UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

THE STATISTICS CONCEPT INVENTORY: DEVELOPMENT AND ANALYSIS OF A COGNITIVE ASSESSMENT INSTRUMENT IN STATISTICS

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

degree of

Doctor of Philosophy

By

KIRK ALLEN
Norman, Oklahoma
2006

Downloaded from SHAREOK. Persistent Handle URL: https://hdl.handle.net/11244/1038

UMI Number: 3212015



UMI Microform 3212015

Copyright 2006 by ProQuest Information and Learning Company.
All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

THE STATISTICS CONCEPT INVENTORY: DEVELOPMENT AND ANALYSIS OF A COGNITIVE ASSESSMENT INSTRUMENT IN STATISTICS

A DISSERTATION APPROVED FOR THE
SCHOOL OF INDUSTRIAL ENGINEERING

BY

Teri Reed Rhoads

Randa L. Shehab

Suleyman Karabuk

Teri J. Murphy

Robert A. Terry

© Copyright by KIRK ALLEN 2006
All Rights Reserved.

Great edifices,
like the great mountains,
are the work of ages.
Often art undergoes a transformation
while they are waiting pending completion
— *pendent opera interrupta* —
they then proceed imperturbably
in conformity with the new order of things.

-- Victor Hugo, *Notre Dame de Paris*

How to read this dissertation

The dissertation is organized into four books. The basic content is as follows:

- Book One. This would have been my Master's thesis. It documents the creation of the SCI focusing up to Spring 2004.
- Book Two. This is a transition. The conclusions drawn from the SCI are expanded and further explorations begin.
- Book Three. The validity of the SCI is re-examined at test-level.
- Book Four. This grand finale summarizes the work and suggests directions for future studies.

Some chapters have an "a" and "b" designation. This signifies either (1) a parallel structure, such as methodology and results; or (2) the chapter is sufficiently long that it could almost be two chapters.

Book One was written in a "traditional" five-chapter format. The literature review was later divided into "Test Theory" and "Concept Inventories," while the "Methodology" and "Results" were combined due to the brevity of the methodology and in keeping with the five-chapter idea.

Book Two, Book Three, and Book Four are intended as more-or-less independent works which could be publishable articles (although generally shortened). As such, each of these chapters has introduction, results, references, etc. In turn, there may be overlap between chapters, especially in the background and associated references.

A general model for test creation is presented in Chapter I (Figure 1) and re-visited in the concluding Chapter XII. The organization of the dissertation is depicted in complementary diagrams in Chapters V (Figure 1) and XII (Figures 1 and 9).

CONTENTS

Book One

	Front Matter	**1**
I	**Commencement**	**8**
II	**Test Theory**	**13**
III	**Concept Inventories**	**38**
IV	**a. Methodology**	**77**
	b. Results	**86**
V	**Preliminary Conclusions**	**182**
	References	**184**

Book Two

	Front Matter	**191**
VI	**Assessing and Improving Test Reliability: An Engineer's Perspective**	**197**
VII	**Development and Equivalency of the Online SCI**	**216**
VIII	**Self Efficacy of Statistical Reasoning Skills**	
	a. Literature Review	**271**
	b. Confidence Analysis of the SCI	**295**

Book Three

	Front Matter	**331**
IX	**Statistics as a multi-dimensional construct**	**336**
X	**Content Validity of the Statistics Concept Inventory**	**386**
XI	**Concept Inventory Cookbook**	**433**

Book Four

	Front Matter	**452**
XII	**The Statistics Concept Inventory: A Tool for Measuring Cognitive Achievement in Introductory Statistics**	**455**
	Full References	**476**

The researchers wish to acknowledge the support provided by a grant from the National Science Foundation, Division of Undergraduate Education, Assessment of Student Achievement program (DUE-0206977). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Abstract

The Statistics Concept Inventory (SCI) is a multiple choice test designed to assess students' conceptual understanding of topics typically encountered in an introductory statistics course. This dissertation documents the development of the SCI from Fall 2002 up to Spring 2006. The first phase of the project essentially sought to answer the question: "Can you write a test to assess topics typically encountered in introductory statistics?" Book One presents the results utilized in answering this question in the affirmative. The bulk of the results present the development and evolution of the items, primarily relying on objective metrics to gauge effectiveness but also incorporating student feedback. The second phase boils down to: "Now that you have the test, what else can you do with it?" This includes an exploration of Cronbach's alpha, the most commonly-used measure of test reliability in the literature. An online version of the SCI was designed, and its equivalency to the paper version is assessed. Adding an extra wrinkle to the online SCI, subjects rated their answer confidence. These results show a general positive trend between confidence and correct responses. However, some items buck this trend, revealing potential sources of misunderstandings, with comparisons offered to the extant statistics and probability educational research. The third phase is a re-assessment of the SCI: "Are you sure?" A factor analytic study favored a uni-dimensional structure for the SCI, although maintaining the likelihood of a deeper structure if more items can be written to tap similar topics. A shortened version of the instrument is proposed, demonstrated to be able to maintain a reliability nearly identical to that of the full instrument. Incorporating student feedback and a faculty topics survey, improvements to the items and recommendations for further research are proposed. The

state of the concept inventory movement is assessed, to offer a comparison to the work presented on the SCI. Finally, the dissertation concludes with a summary of the four years' progress, acknowledging that work is never complete but that the results thus far place the SCI in a strong position to grow for years to come.

Book One

Table of Contents

List of Tables 5
List of Figures 7

I Commencement 8

1. Problem Statement 9
    1.1 Model 10
2. Test Theory 11
3. Concept Inventories 12
4. Methodology and Results 12
5. Preliminary Conclusions 12

II Test Theory 13

1. Validity 13
    1.1 Content Validity 13
    1.2 Concurrent Validity 13
    1.3 Predictive Validity 13
    1.4 Construct Validity 14
2. Reliability 15
    2.1 Kuder-Richardson 16
    2.2 Cronbach 18
    2.3 Guttman 23
    2.4 Critiques of Reliability 26
    2.5 Confidence Intervals and Hypothesis Tests on Alpha 28
3. Discrimination 29
    3.1 Ferguson's Delta 29
    3.2 Discriminatory Index 29
    3.3 Point-Biserial Correlation 30
4. Item Analysis 32
    4.1 Optimal number of distracters 33
5. Conclusion 37

III Concept Inventories 38

1. Introduction 38
2. Force Concept Inventory 39
    2.1 Early Work 39
    2.2 Force Concept Inventory 43
    2.3 Interpreting the FCI 44
    2.4 Uses of the FCI 46
    2.5 Mechanics Baseline Test (MBT) 47
3. Engineering Concept Inventories 49

3.1 Materials Concept Inventory (MCI) 49
3.2 Statics Concept Inventory 52
3.3 Thermal and Transport Science Concept Inventory 55
3.4 Wave Concepts Inventory (WCI) 56
3.5 Heat Transfer Concept Inventory (HTCI)....57
3.6 Fluid Mechanics Concept Inventory (FMCI) 58
3.7 Dynamics Concept Inventory (DCI)....59
3.8 Circuits Concept Inventory (CCI)....61
3.9 Computer Engineering Concept Inventory (CPECI) 61
3.10 Electromagnetics Concept Inventory (EMCI) 62
3.11 Electronics Concept Inventory (ECI)....62
3.12 Signals and Systems Concept Inventory (SSCI)....63
3.13 Strength of Materials Concept Inventory (SOMCI) 64
3.14 Thermodynamics Concept Inventory....65
3.15 Chemistry Concept Inventory (CCI)....66
4. Other Concept Inventories 67
4.1 Concept Inventory of Natural Selection (CINS)....67
4.2 Chemical equilibrium....68
4.3 Conceptual Survey of Electricity and Magnetism 69
4.4 Test of Understanding Graphs in Kinematics....70
4.5 Force and Motion Conceptual Evaluation 72
4.6 Resistive Electric Circuit Concepts Test....72
4.7 Geoscience Concept Inventory (GCI)....74
5. Conclusion 76

IV Methodology and Results 77

A. Methodology 77
1. Scores....77
2. Validity 78
2.1 Content Validity....78
2.2 Concurrent Validity 81
2.3 Predictive Validity 81
2.4 Construct Validity....82
3. Reliability....83
3.1 Reliability of the SCI 83
3.2 Confidence Intervals and Hypothesis tests on Alpha....83
3.3 Guttman coefficients....83
4. Discrimination....83
5. Item Analysis 84

B. Results 86
1. Scores....86
2. Validity 91
2.1 Content Validity....91
2.2 Concurrent Validity 96

2.3 Predictive Validity 100
2.4 Construct Validity 102
3. Reliability 109
3.1 Coefficient Alpha 109
3.2 Confidence Intervals and Hypothesis tests on Alpha 112
3.3 Guttman coefficients 114
4. Discrimination 115
4.1 Ferguson's Delta 115
4.2 Discriminatory Index 116
4.3 Point-biserial correlation 117
5. Item Analysis 118
5.1 Description of questions and changes 119
5.2 Item analysis conclusions 181

V Preliminary Conclusions 182

References 184

List of Tables

Chapter II

Table 1: Cronbach's Split-Half Results 19

Chapter III

Table 1: List of Concept Inventories 37
Table 2: Summary of Results of Physics Diagnostic Test 39
Table 3: Course Grade compared to Diagnostic score for one course 40
Table 4: Correlation Coefficients for data in Figure 1 47
Table 5: Answer and Reason Conditional Probabilities on TISC 68

Chapter IV

Table 1: Abbreviations Used to Identify Classes 77
Table 2: Classes by Semester 78
Table 3: Summary Statistics for SCI, Summer 2003 86
Table 4: Summary Statistics for SCI, Fall 2003 86
Table 5: Summary Statistics for SCI, Spring 2004 86
Table 6: Summary Statistics for SCI, Summer 2004 87
Table 7: Summary Statistics for SCI, Spring 2004 87
Table 8: Summary Statistics for SCI, Spring 2005 87
Table 9: Summary Statistics for SCI, Summer 2005 87
Table 10: Grouping by Who Took Pre, Post, or Both, Fall 2003 88
Table 11: Percent Correct (and Gains) for Sub-Topics, Fall 2003 Post-Test 90
Table 12: Percent Correct (and Gains) for Sub-Topics, Spring 2004 Post-Test 90
Table 13: Percent Correct (and Gains) for Sub-Topics, Fall 2004 Post-Test 90
Table 14: Percent Correct (and Gains) for Sub-Topics, Spring 2005 Post-Test 90
Table 15: Item Categorization by Faculty Survey Topics 92
Table 16: Important Topics Missing from the SCI 93
Table 17: Number of Items in Faculty Survey General Areas 94
Table 18: Number of Items in AP Statistics Areas 94
Table 19: Correlation of SCI Scores with Overall Course Grade(%) 96
Table 20: Correlation of SCI Sub-Scores with Overall Course Grade(%) 97
Table 21: Correlation of SCI Sub-Scores with Overall Course Grade(%) 100
Table 22: Results for 3-Factor FIML Model 103
Table 23: Results for 4-Factor FIML Model 104
Table 24: Item Groupings for three Factor Analytic Solutions, 11 Factors 106
Table 25: Item Groupings for three Factor Analytic Solutions, 5 Factors 107
Table 26: Possible Constructs based on 11-Factor Models 107
Table 27: Coefficient Alpha, Pre-Test and Post-Test 109
Table 28: Coefficient Alpha for Sub-Topics, Post-Test 110
Table 29: Summary of Hypothesis Tests on Alpha 112
Table 30: Confidence Intervals and Hypothesis Tests for Alpha, Post-Tests only 113

Table 31: Guttman Reliability Estimates, Fall 2003 Post-Test....114
Table 32: Discriminatory Power, Post-Tests....115
Table 33: Item Discriminatory Index, Number of Questions in Each Range....116
Table 34: Comparison between Discriminatory Index and Point-biserial Correlation....117
Table 35: Detailed Breakdown of Discriminatory Index and Point-biserial Correlation....117
Table 36: Correlation between Discriminatory Index and Point-biserial Correlation....118
Table 37: Item Analysis Statistics for Median question....119
Table 38: Knowledge Gain on Median question....120
Table 39: Item Analysis Statistics for Change of Units question....122
Table 40: Item Analysis Statistics for Rolling Dice question....124
Table 41: Item Analysis Statistics for Memory-Less Property question....126
Table 42: Item Analysis Statistics for Height question....128
Table 43: Item Analysis Statistics for Bias question....129
Table 44: Item Analysis Statistics for Normal Distribution question....131
Table 45: Item Analysis Statistics for Confidence Interval question....133
Table 46: Item Analysis Statistics for Percentile question....136
Table 47: Item Analysis Statistics for Confidence Interval question....137
Table 48: Item Analysis Statistics for Hospital question....138
Table 49: Item Analysis Statistics for Conditional Probability question....141
Table 50: Item Analysis Statistics for p-value question....142
Table 51: Item Analysis Statistics for t-distribution question....144
Table 52: Item Analysis Statistics for Chance of Rain problem....145
Table 53: Item Analysis Statistics for False Positives question....147
Table 54: Item Analysis Statistics for Normal Distribution question....148
Table 55: Item Analysis Statistics for Population question....150
Table 56: Item Analysis Statistics for Coin Flipping question....153
Table 57: Item Analysis Statistics for Coin Sequence question....155
Table 58: Item Analysis Statistics for GPA question....157
Table 59: Item Analysis Statistics for p-value question....158
Table 60: Item Analysis Statistics for p-value question....159
Table 61: Item Analysis Statistics for Parent Distribution question....161
Table 62: Item Analysis Statistics for Graphical Variability question....165
Table 63: Item Analysis Statistics for Central Tendency question....167
Table 64: Item Analysis Statistics for ANOVA question....168
Table 65: Item Analysis Statistics for Multiple Regression question....169
Table 66: Item Analysis Statistics for Hypothesis Definition question....171
Table 67: Item Analysis Statistics for Regression question....172
Table 68: Item Analysis Statistics for Regression question....174
Table 69: Item Analysis Statistics for Standard Deviation question....175
Table 70: Item Analysis Statistics for Standard Deviation question....176
Table 71: Item Analysis Statistics for Waiting Time problem....177
Table 72: Item Analysis Statistics for Variability question....178
Table 73: Item Analysis Statistics for Outlier question....178
Table 74: Item Analysis Statistics for Chance of Rain question....179
Table 75: Item Analysis Statistics for Interpreting p-value question....180
Table 76: Knowledge Gain on Interpreting p-value question....180

List of Figures

Chapter I

Figure 1: Test creation process 11

Chapter III

Figure 1: FCI Scores vs. MBT Scores (Hestenes and Wells, 1992).....47

Chapter IV

Figure 1: Scree Plot for Principal Component Analysis, from SPSS ™ 105

Chapter V

Figure 1: Creation process presented in Book One 183

CHAPTER I

Commencement

The concept inventory movement was spurred by the development and successful implementation of the Force Concept Inventory (FCI) (Hestenes, 1985). The FCI was developed as a pre-post test to identify student misconceptions about Newtonian force when entering a physics course and check for gains upon completing the course. After many rounds of testing, it was discovered that students gain the most conceptual knowledge in interactive engagement courses, as opposed to traditional lectures (Hake, 1998).

The success of the FCI prompted educators to develop instruments in other fields. In light of recent Accreditation Board for Engineering and Technology (ABET) standards which focus on outcomes rather than simply fulfilling seat time requirements, many engineering fields have begun to development concept inventories, such as Thermodynamics, Statics, and Heat Transfer (Evans, *et al.,* 2003). The development of a Statistics Concept Inventory (SCI) is the topic of this dissertation.

Statistics play a large part in the changing face of engineering education. For instance, according to the ABET criteria for 2006-2007, "Engineering programs must demonstrate that their students attain … an ability to design and conduct experiments, as well as analyze and interpret data" (Engineering Accreditation Commission, 2006). In addition, 15 of the 24 engineering programs that ABET accredits list statistics in the criteria, with 10 of these including probability. Most of these contain the exact phrase "probability and statistics." Statistical knowledge can also be inferred in other programs through terms such as "analyze," "model," and "stochastic."

Statistics play a large part in the changing face of engineering education. For instance, according to the new ABET Criterion 3, "Engineering programs must demonstrate that their graduates have (a) an ability to apply knowledge of mathematics, science, and engineering, and (b) an ability to design and conduct experiments, as well as analyze and interpret data" (Engineering Accreditation Commission, 2003). In addition, 16 of the 24 engineering programs which ABET accredits list statistics in their accreditation criteria, and 11 of 24 mention probability. Most of these contain the exact phrase "probability and statistics" (Engineering Accreditation Commission, 2003).

1. Problem Statement

The goal of the project is to develop an instrument which measures students' conceptual knowledge of statistics while meeting accepted standards for test validity, reliability, and discriminatory power. This first book of the dissertation documents the development process of the Statistics Concept Inventory (SCI) on these grounds. The validity, reliability, and discriminatory power of the SCI are documented for the test as a whole, and detailed analyses are given for how each question was modified (or not) to improve the instrument.

Whereas most classroom tests measure computational ability, the intent of the SCI is to measure students' conceptual knowledge. In this sense, it can be viewed as a supplement to traditional student-teacher interaction. The SCI is best-suited for use as a pre-post test so that gains can be tracked from the beginning to end of instruction. However, it could be useful strictly as a post-test if the instructor preferred, still allowing comparison to scores from other instructors, courses and/or institutions.

The ultimate goal is to develop an instrument which is recognized on a national level as a useful tool for monitoring student learning on an individual or classroom basis to comparing scores across universities. This will require not just a sound instrument but also the appropriate dissemination methods, including journal publications, conference presentations, and even informal communication between faculty. The instrument has been tested extensively in introductory statistics courses in the College of Engineering and Department of Mathematics at the University of Oklahoma. Two other four-year universities have used the instrument in introductory engineering statistics courses, and one two-year college has participated with its introductory course. Data for these courses are presented, and their results are compared and contrasted with the perspective that an introductory statistics course in an engineering department is the target audience.

1.1 Model

Figure 1 offers a general model for test construction, which is followed throughout the dissertation. Book One first presents the background on test theory (Chapter II) and concept inventories (Chapter III) which led to item development. Once items are developed and administered, they can be analyzed both on the test and item level, with interaction between the methods and some overlap (e.g., content validity). Book One essentially encompasses four iterations through this loop. Later work is more difficulty to classify; the dashed arrow from test level analyses to test theory acknowledges that the creation of a valid and reliable SCI allowed greater insight into test theory. Some amount of dissemination occurred through the conference papers utilizing this first round of data, although the dissemination node of Figure 1 refers to a public disclosure of the SCI, such as through the Journal of Engineering Education. The

sampling of courses thus far is broad, but wider dissemination is needed to draw pedagogical implications leading to an enhanced understanding of statistics education.

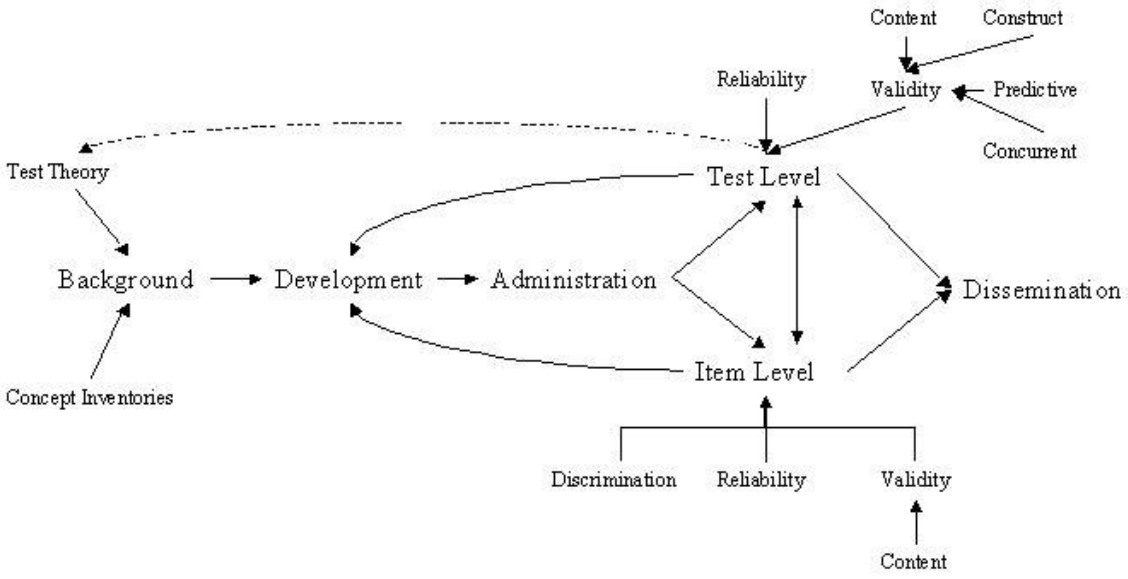


Figure 1: Test creation process

This figure will be re-visited in the concluding chapter (XII) of the dissertation. Diagrams of the dissertation layout can be found in Chapters V and XII; these diagrams illustrate how the chapters fit together, as a guide to the reader who is unfamiliar with the project.

2. Test Theory

The SCI development occurred in tandem with research into prevailing standards of validity and reliability. The techniques are comparable and in many cases more advanced than those used on other concept inventories. However, these techniques are at times still rather naïve and simplistic.

In brief, reliability is a pre-requisite (a necessary condition) for validity. A reliable test is one in which the measurement error is small. Validity, then, refers to the extent that the test measures what it is intended to measure (in this case, conceptual knowledge

of introductory statistics). There are many types of validity referenced in psychometric textbooks. For a concept inventory, content validity is typically viewed as the most essential: the test adequately covers all content areas which it proposes to measure. Additionally, a test is expected to be discriminating, both at the item level and as an aggregate measure: higher-ability students should be more inclined to answer each item correctly, which in turns produces a wide range of scores for the test.

3. Concept Inventories

The development processes and uses of other concept inventories are documented in Chapter III. This serves as a reference for the type of analysis that is considered publishable within the field of educational research.

4. Methodology and Results

The methods utilized in constructing and editing the SCI are described in the first part of Chapter IV. Results are first presented across the test as a whole. Each item's evolution and the reasons for doing so are detailed, along with the item analysis statistics utilized in making these decisions.

5. Preliminary Conclusions

This short chapter is a reminder that Book One is but the beginning of the Statistics Concept Inventory. The demarcation between Book One and the later work serves as a reminder that the methods of the researchers evolved along with the instrument itself.

CHAPTER II

Test Theory

This chapter describes the methods used to construct and analyze the Statistics Concept Inventory. Validity seeks to answer the question: "Does the test measure what it claims to measure?" In constructing the SCI, the first considerations were given to the topics to be assessed, and therefore validity is discussed first herein. As a necessary condition for validity, reliability may be considered more fundamental, but the test had to be constructed and piloted before reliability could be assessed. As such, reliability theory is discussed second. Many sections of this chapter are quite short and rather definitional. The application of these methods to other concept inventories is found in Chapter III. This parsing facilitates an understanding of the level of analysis utilized on concept inventories, whereas splicing Chapters II and III hinders the exposition.

1. Validity

Validity refers to the extent that an instrument measures what it claims to measure. Validation is the process of accumulating evidence supporting this claim. Validation is an on-going process: the instrument must be constantly evaluated as its uses and needs evolve (Nunnally, 1978). There are many types of validity, such as face validity, concurrent validity, predictive validity, incremental validity, and construct validity (Kline, 1986). However, they are not mutually exclusive. The following sections describe the types of validity that are relevant to this project – content, concurrent, predictive, and construct.

1.1 Content Validity

Content validity refers to the extent to which items are (1) representative of the knowledge base being tested and (2) constructed in a "sensible" manner (Nunnally, 1978). Achieving content validity requires finding an adequate sampling of possible topics and incorporating value judgments as to which topics to include in the instrument.

1.2 Concurrent Validity

Concurrent validity is "assessed by correlating the test with other tests" (Klein, 1986). This implies a decision as to what constitutes the "other test." Of course, if another test already exists, it raises the question of whether the test being constructed is even necessary. Therefore, the term "other test" should be loosely interpreted. On concept inventories, a logical selection is the course grade or final exam score, with the caveat that a concept inventory does not focus on computation.

1.3 Predictive Validity

Predictive validity refers to a test's ability to accurately predict future performance. This is often discussed in the context of training, such as whether a training program can be considered valid at increasing job performance (Thorndike, 1982). A decision must be made as to what constitutes success in a future endeavor. In an academic setting, future performance may be considered the grade earned for a course or graduation.

1.4 Construct Validity

A test is "constructed" to measure some latent ability of its subjects (Thorndike, 1982). This is often applied to personality scales where the desired measure may be a quality not directly observable, such as aggression or courage. This concept is often

extended to achievement tests to define sub-tests within a larger instrument. The hope is to find specific abilities within a larger domain of knowledge.

The technique of factor analysis is commonly used to find and analyze these sub-topics. It is expected that similar items should be highly correlated. Factor analysis analyzes the correlation matrix between items to find a smaller number of groups or dimensions through which the variables (item scores) can be expressed. It is generally desired to find a solution which places each item on only one factor with near-zero loadings on the other factors. This is called the simple structure (Kline, 1993). The loading is the correlation of an item with the factor. Each item will load on all factors.

There are many ways to find the simple structure. It is widely accepted to consider all factors with eigenvalues greater than one. Some researchers consider this arbitrary and use subjective judgment such as a Scree test. This test looks for large differences between consecutive eigenvalues. Another decision involves rotation of the factors, which affects the loadings but not the overall fit of the solution. Varimax and Direct Oblimin rotations are recommended for their likelihood to approach the simple structure. (Kline, 1993)

2. Reliability

A reliable instrument is one in which measurement error is small, which can also be stated as the extent that the instrument is repeatable (Nunnally, 1978). There are several types of reliability: test-retest measures answer stability on repeated administrations; alternative forms requires subjects to take two similar tests on the same subject; internal consistency is based on inter-item correlations and describes the extent to which the test measures a single attribute (e.g., statistical knowledge).

Internal consistency is the most common measure because it requires only one test administration, reducing costs and eliminating the issue of students gaining knowledge between test administrations. Internal consistency is typically measured using Cronbach's alpha (1951), which is a generalized form of Kuder-Richardson Formula 20 (1937). Typically, a test is considered reliable if alpha is above 0.80 (Nunnally, 1978). Other sources consider a value of 0.60 to 0.80 to be acceptable for classroom tests (Oosterhof, 1996). A historical description of the development of the reliability coefficient follows.

2.1 Kuder-Richardson

One of the first attempts to quantify test reliability was set forth by Kuder and Richardson in their 1937 article "The Theory of the Estimation of Test Reliability." They comment that a reliability coefficient based on test-retest will often result in a reliability that is too high due to material remembered on the second administration. Further, increased time between administrations is impractical because subjects may gain knowledge.

The authors next focus on the split-half coefficient, where the test is split in two parts and a correlation is calculated between those two parts. For a test of length k, there are $\frac{k!}{2\left(\frac{k}{2}!\right)^2}$ ways to split a test in two (combination formula divided by 2). For a test with 10 items, there are 126 combinations. Each split-half will result in a different reliability. To overcome this problem, the authors "present certain deductions from test theory which lead to unique values of the reliability coefficient" (p. 152).

The first result of the paper is given by equation (1). This formula assumes that the matrix of inter-item correlations has a rank of one.

$$r_{tt} = \frac{\sigma_t^2 - \Sigma pq + \Sigma r_{ii}\, pq}{\sigma_t^2} \quad (1)$$

where: r_{tt} is the reliability of the test
σ_t^2 is the total score variance for the test
p is the proportion of students who answer each item correctly
q is the proportion of students who answer each item incorrectly
r_{ii} is reliability of item i
$\sum$ represents the sum over all items on a test

Unfortunately, the authors point out that this equation is not directly usable because r_{ii} is not "operationally determinable" (p. 154). As an approximation, the authors recommend using the average correlation of item i with the other *k – 1* items.

To put the equation in a more usable form, several assumptions are made. First, all inter-item correlations are equal to $\overline{r_{ii}}$, the average inter-item correlation for all items. This yields the following expression (obtained by substituting $\overline{r_{ii}}$ into equation 1):

$$r_{tt} = \frac{\overline{r_{ii}}(\Sigma\sqrt{pq})^2}{\sigma_t^2} \quad (2)$$

Next, the following partition of total test variance is given for a test with *k* items:

$$\sigma_t^2 = \sigma_a^2 + \sigma_b^2 + \ldots + \sigma_n^2 + 2(r_{ab}\sigma_a\sigma_b + r_{ac}\sigma_a\sigma_c + \ldots + r_{(k-1)k}\sigma_{k-1}\sigma_k) \quad (3)$$

Holding the $\overline{r_{ii}}$ assumption true as above and substituting $\sqrt{pq}$ as the standard deviation of each item, the following equation for test reliability is obtained:

$$r_{tt} = \frac{\sigma_t^2 - \Sigma pq}{(\Sigma\sqrt{pq})^2 - \Sigma pq} * \frac{(\Sigma\sqrt{pq})^2}{\sigma_t^2} \quad (4)$$

This equation has the advantage that it uses quantities which are directly calculated from the test data (i.e., no inter-item correlations). To further simplify matters, equal item variance can be assumed ($\overline{pq}$, i.e., items have equal difficulty: *p* is constant

over all items). The term $\sum p_i q_i$ can be substituted in place of $k\overline{pq}$ to give a more general result. The general result is actually equation 20, but what is commonly cited as KR-20 has the more precise sum. Equation 20 (KR-20) is given below:

$$r_{tt} = \frac{k}{k-1}\frac{\sigma_t^2 - k\overline{pq}}{\sigma_t^2} \tag{5}$$

2.2 Cronbach

Following Kuder and Richardson, the study of reliability was advanced by Lee J. Cronbach in several articles. The first of these, "On Estimates of Test Reliability" (1943), dealt with criticisms of the split-half method and made recommendations to overcome these problems.

The first criticism was the same as pointed out by Kuder and Richardson – namely, that there are many possible ways to split a test in half. Cronbach comments that "the split-half method gives erroneous estimates whenever the assumption that the halves are of equal difficulty, variability, and reliability is not met" (p. 486). This means that a random split or even something simple such as odd-even could result in a low split-half reliability. Cronbach later presents guidelines for selecting an appropriate split.

Several criticisms of the KR-20 are discussed. The primary concern is lack of the extent that the KR-20 serves as the lower bound to reliability. As Cronbach shows in a later article, KR-20 is the mean of all split-half coefficients. This has the disadvantage of including poor splits which do not meet the criteria of equal difficulty and variability. In this way, KR-20 is a conservative estimate of reliability. Evidence of this conservatism can be found in a negative KR-20. As Cronbach puts it, "This is of course meaningless, since complete heterogeneity would yield a coefficient of 0.00" (p. 487).

To overcome this problem, Cronbach proposes the "Parallel-Split Method." The method is analogous to parallel-forms, in which subjects take two tests where items are as identical as possible in terms of form, content, difficulty, and range of difficulty (p. 489). This can be achieved in a single administration by making the two halves as similar as possible. The technique for obtaining a parallel split is described below (p. 490):

> "To obtain a parallel split, the investigator requires an item-analysis. This is made, using a representative, but small, sample of papers not used in the actual correlation. Using this analysis, pairs of items are selected which test the same behaviors or knowledge, and which are of roughly equal difficulty. 'Testing the same behavior' means not only similarity of content, but similarity of response behavior."

To ensure the split is accurate, several further measures can be taken. First, each half is split again to yield fourths. Each subject's score is obtained for each of the fourths, and the correlations between the fourths are calculated. If the computation "$r_{12}r_{34} - r_{13}r_{24}$ is within sampling expectancy of zero, the halves are comparable" (p. 490). Further, the following equalities should hold:

Mean first half = Mean second half

Variance first half = Variance second half

Following these theoretical proposals regarding split-halves, Cronbach (1946) analyzed the "Test of Silent Reading Vocabulary" to determine the variability of the split-half reliability for different splits. He used four pre-determined splits (odd-even, first-last, easy-hard, low variance-high variance), thirty random splits, and fourteen parallel splits (as nearly parallel as possible, as defined above).

The KR-20 for the test is 0.820. The reliability for each split ("a" and "b") is calculated using the Guttman (1945) formula, given below:

$$r = 2\left(1 - \frac{\sigma_a^{\,2} + \sigma_b^{\,2}}{\sigma_t^{\,2}}\right) \tag{6}$$

A good split (i.e., the halves are comparable) is one that has a ratio of standard deviations (σ_a/σ_b) between 0.90 and 1.10. The results are presented below:

Table 1: Cronbach's Split-Half Results (1946)

Split	**σ_a/σ_b**	**n**	**Statistic**	**Reliability**
Odd-even	1.03	1	--	0.796
First-last	1.01	1	--	0.862
Easy-hard	0.95	1	--	0.806
Low V-high V	1.23	1	--	0.809
Random	0.90 to 1.10	17	Median Q1 Q3	0.829 0.808 0.844
Random	< 0.90 or > 1.10	13	Median Q1 Q3	0.815 0.803 0.828
Parallel	0.90 to 1.10	11	Median Q1 Q3	0.832 0.810 0.850

The results of this study indicate that the parallel split method finds the maximum reliability coefficient in fewer attempts than other methods. Single splits should not be taken because they may yield lucky (first-last, 0.862) or unlucky (odd-even, 0.796) splits relative to the overall reliability. When multiple splits are taken (random or parallel), those which are similar yield the highest reliability. In practice, the parallel method may save time because it yields a higher proportion of comparable splits compared to the random method (11 of 14, compared to 17 of 30).

Cronbach's most-cited work regarding reliability is "Coefficient Alpha and the Internal Structure of Tests" (1951). This article is in many ways a summary of the work

already presented. Cronbach dubs the KR-20 "coefficient alpha" to serve as a short-hand notation. He also gives a more generalized form, presented below:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\Sigma V_i}{V_{test}}\right) \tag{7}$$

where: α is Cronbach's coefficient alpha
k is the number of questions on the test
V_i is the variance of each question
V_{test} is the total score (not percentage) variance of the entire test

For dichotomously scored (binary: 0 or 1) items, V_i reduces to $p_i q_i$ and KR-20 is obtained. This relationship is derived using the basic definition of population variance:

$$\sigma^2 = V_i = \frac{\Sigma(x_i - \mu)^2}{n} \tag{8}$$

where: x_i are the individual observations (0 or 1)
μ is the population mean (p_i for each question)
n is the total number of observations (students)

For binary data, the sum portion of the variance equation can be broken down into the 0 and 1 scores:

For 0 scores on a question: $\Sigma(x_i - \mu)^2 = (0 - p_i)^2 q_i n = p_i^2 q_i n$

The term $(0 - p_i)^2$ represents the fact that 0 is the value of each observation (x_i) and that the overall mean for each question is p_i. The term $q_i n$ accounts for summing all incorrect scores for that question (the proportion incorrect multiplied by the total number).

For the correct students, the same logic holds in calculating *Vi*, but each x_i is 1 and the total number of correct students is $p_i n$. The term $(1 - p_i)$ is the proportion incorrect, referred to as q_i.

For 1 scores on same question: $\Sigma(x_i - \mu)^2 = (1 - p_i)^2 p_i n = q_i^2 p_i n$

Combining the 0 and 1 portions and dividing by n yields the total variance for an individual question (equation 9, V_i).

$$V_i = \frac{p_i^2 q_i n + q_i^2 p_i n}{n} \tag{9}$$

The next step is to divide out the n's and re-arrange the numerator:

$$V_i = p_i q_i (p_i + q_i) \tag{10}$$

The term $p_i + q_i$ is the sum of the proportion correct plus the proportion incorrect, which totals 1. Therefore, the final result for each question's variance is:

$$V_i = p_i q_i \tag{11}$$

A common statement about alpha is that it can be inflated by increasing the length of the test. This can be explained by considering the definition of variance, shown below.

$$V = \frac{\Sigma(x-\mu)^2}{n} \tag{12}$$

where: x is each individual score
μ is the mean for the measure of interest
n is the sample size

For example, there is a hypothetical class with 10 students taking a test with 10 questions. The overall scores are 2, 2, 2, 3, 4, 6, 7, 8, 8, 8. This yields a mean of 5.0 and a variance of 6.4. For simplicity, it is assumed that each question was answered correctly by 5 students (i.e., each question has a variance of 0.25). Using Equation 2.1, this yields an alpha of 0.677:

$$\frac{10}{9}\left(1 - \frac{10 * 0.25}{6.4}\right) = 0.677 \tag{13}$$

For comparison, this hypothetical class now takes a 20-question test and all the scores are doubled (i.e., 4, 4, 4, 6, …, 16). This is equivalent to doubling the test length but covering the same material. The mean is now 10.0 and the variance has increased fourfold, because of the squared term in equation 12, to 25.6. However, the sum of item

variances will merely double. Alpha is now 0.847 – an increase of 0.170 into an "acceptable" range simply by doubling the test length.

$$\frac{20}{19}\left(1-\frac{20*0.25}{25.6}\right)=0.847 \tag{15}$$

2.3 Guttman

Guttman (1945) approaches the concept of reliability according to "what seems to have been Spearman's original purpose" (p. 256). One of his equations matches the KR-20 formula and Cronbach's alpha. However, Guttman believes his derivations are superior because he does not require as many assumptions, such as that inter-item correlations are constant or that the rank of the matrix of inter-item correlations is one. Guttman's work follows Kuder-Richardson by approximately eight years and precedes Cronbach's first major publication on this subject by two years.

Fundamental concepts are laid out which differ from conventional derivations in several respects. First, Guttman identifies three distinct sources of variation: trials, persons, and items. He considers unreliability to be "variation over trials" (p. 257, point 1). Second, the total test variance is the sum of error variance and variance of expected scores. It follows that reliability is "the complement of the ratio of error variance to total variance" (p. 257, point 2). Guttman is interested in information that can be obtained from a single trial. This yields a lower bound to the true reliability, but it is advantageous to avoid difficulties associated with multiple test administrations (trials). Finally, he states that the one basic assumption is that "errors of observation are independent between items and between persons over the *universe of trials*" (p. 257, point 5, his italics). No assumptions are made about the relationships between the items (point 6).

Guttman makes three basic assumptions throughout the paper. They are essentially necessary only to be rigorously correct from a mathematical standpoint and are usually attained in practice. The assumptions are described below:

- Assumption (A) states that the following moments exist:

$$\underset{i}{E}\,\underset{k}{E}\,x_{ijk}^{\,p} \tag{16}$$

where: E refers to the expected value
i are the subjects (students)
j are the items (1, 2, …, n)
k are the trials
p = 1, 2, 3, 4

  In practice, this assumption is invariably fulfilled because all moments exist for a finite distribution, and tests do not permit infinite scores.

- Assumption (B) states that the population of individuals and the universe of trials are indefinitely large. This assumption is not explicitly used but is necessary to make Assumption (C) hold.

- Assumption (C) states two things: (C_1) "the observed value of an individual on an item is experimentally independent of his values on any other items"; and (C_2) "the observed value of an individual on an item is experimentally independent of the observed values of any other individual on that or any other item." This basically means that (C_1) items should be arranged in a way to prevent carry over from one item to the next; and (C_2) subjects do not copy from one another.

With these assumptions in mind, Guttman derives six measures of reliability. They are considered lower bounds because a single test administration is not sufficient to provide evidence against the hypothesis that the true test reliability is one (H_0: $\rho_t^2 = 1$). It follows that the estimated lower bound (L) is less than or equal to the true reliability

which is less than or equal to one ($L \leq \rho_t^2 \leq 1$). The simplest estimate of reliability (L_1) is the complement of the ratio of error variance to total variance (shown below, respecting Guttman's notation for variance).

$$L_1 = 1 - \frac{\sum_{j=1}^{k} s_j^{\,2}}{s_t^{\,2}} \tag{17}$$

where: s_j^2 are the item variances (V_i for Cronbach)
s_t^2 is the total test variance (V_{test} for Cronbach)
k is the number of items

A better lower bound (i.e., higher) is L_2, which accounts for covariance between items. The formula is:

$$L_2 = L_1 + \frac{\sqrt{\frac{k}{k-1} C_2}}{s_t^{\,2}} \tag{18}$$

where: C_2 is the sum of the squares of the covariances over all items

Rather than calculating the covariance matrix (presumably difficult in 1945), Guttman suggests weakening L_2 as an estimate to the reliability. This result is the same as Cronbach's alpha:

$$L_3 = \frac{k}{k-1} L_1 = \frac{k}{k-1} \left(1 - \frac{\sum_{j=1}^{k} s_j^{\,2}}{s_t^{\,2}} \right) \tag{19}$$

L_3 will be only slightly less than L_2 if the covariances are homogenous and positive. However, L_2 will be a better estimate if there are negative covariances.

The fourth estimate (L_4) is the split-half reliability analyzed by Cronbach in detail (equation 6). Guttman remarks that any split can serve as a lower bound, but splits which correlate more highly with each other will yield larger L_4 values. Guttman comments that L_3 and L_4 are the most useful in practice due to the ease of calculation.

The fifth estimate (L_5) is again based on covariances. It will be greater than L_2 when one item has large covariances compared to the covariances of other items. Otherwise, L_5 is less than or equal to L_2. The formula is:

$$L_5 = L_1 + \frac{2\sqrt{\overline{C_2}}}{s_t^2} \tag{20}$$

where: $\overline{C_2}$ is calculated by first finding the $k - 1$ sum of the squares of the covariances for all k items (C_{2j}). The largest of these sums is $\overline{C_2}$.

The final estimate of reliability (L_6) is based on a linear multiple regression for each item with the other $k - 1$ items. The variances of the errors (e_j^2) replaces the item variance of L_1. L_6 will tend to be larger than L_2 when the items have low zero-order correlations but high multiple correlations. The formula is:

$$L_6 = 1 - \frac{\sum_{j=1}^{k} e_j^2}{s_t^2} \tag{21}$$

2.4 Critiques of Reliability

Cortina (1993) provides a solid background for various interpretations of alpha present in the literature. They are outlined below (p. 98):

- (a) "Alpha is the mean of all split-half reliabilities." It depends on what is meant by "split-half reliabilities." If the Spearman-Brown prophecy formula is used, then this statement is false. However, if the Flanagan (1937) and Rulon (1939) equation for split-half reliability is used, this statement is true. It is this equation that Cronbach used when making claim (a).
- (b) "Alpha is the lower bound of reliability of a test." This is discussed with (d).
- (c) "Alpha is a measure of first-factor saturation." This suggests "alpha is a measure of the extent to which there is a general factor present in a set of items

and, therefore, the extent to which the items are interrelated" (p. 99). However, this is contradictory to Cronbach's original article and has been proven false by later research.

- (d) "Alpha is equal to the reliability in conditions of essential τ-equivalence." Essential τ-equivalence means that two forms (or halves) have true scores that are linearly related but not necessarily equal. This relates to (b) in that alpha approaches the true reliability as a test becomes τ-equivalent (p. 101). Novick and Lewis (1967) define τ-equivalence as the condition where all individuals have the same true score on two alternate forms of an instrument. Symbolically, this is represented by the following:

$$\tau_{ga} = \tau_{g'a} \tag{22}$$

where: τ is the true score
g and *g'* are equivalent forms of an instrument
a is the subject

- This definition is a rigorous way of saying that two forms (or halves) of a test measure the same thing.
- (e) "Alpha is a more general version of the Kuder-Richardson coefficient of equivalence." This can be shown by deriving that $\sum Vi = \sum p_i q_i$ for dichotomously-scored items.

Cortina examined alpha using simulated data. He varied average item intercorrelation, number of items, and number of orthogonal dimensions (table 2, p. 102). The findings indicate that alpha increases as the number of items or the average item intercorrelation increases. Alpha decreases as the number of orthogonal dimensions increases.

Streiner (2003) points out several misconceptions about alpha and summarizes earlier work. His states that alpha can actually be too high, evidence of "unnecessary duplication

of content across items" and "redundancy" rather than "homogeneity" (p. 102). This is similar to Cortina's finding that alpha increases as the inter-item correlations increase.

An off-shoot of the Streiner statement could be related to reliability of sub-tests. If a test is partitioned into sub-tests by grouping similar items, then it should be expected that each sub-test will have higher inter-item correlations compared to all items on the test. Higher inter-item correlations taken alone will raise the reliability. However, by splitting the test into pieces, the total score variance will decrease more quickly than the sum of the item variances. Therefore, the only way to determine if the sub-tests are more reliable than the entire instrument is to define the sub-tests and perform the calculations. It is possible to be higher or lower depending on the nature of the items and the test.

2.5 Confidence Intervals and Hypothesis Tests on Alpha

Confidence intervals on alpha can be calculated with the equations 23a and 23b (Thompson and Fan 2003):

$$CI_{upper} = 1 - [(1 - \hat{\alpha}) \cdot F_{\gamma/2,\, df1,\, df2}] \quad (23a)$$

$$CI_{lower} = 1 - [(1 - \hat{\alpha}) \cdot F_{1-\gamma/2,\, df1,\, df2}] \quad (23b)$$

where: $\hat{\alpha}$ is the observed alpha from the test
F refers to the F-distribution
$\gamma/2$ is the percentile of the F-distribution, calculated as $\frac{1 + Conf.Level}{2}$
df1 is subjects minus 1, (n-1)
df2 is items minus 1 times subjects minus 1, (n-1)(k-1)

The observed alpha can also be tested against a theoretical population value (α_0). The hypothesis test (H_0: $\alpha = \alpha_0$) is constructed in a similar manner to the confidence intervals. The test statistic is the following:

$$\frac{1 - \alpha_0}{1 - \hat{\alpha}} = F_{df1,\, df2} \quad (24)$$

3. Discrimination

Discrimination refers to a test's ability to produce a wide range of scores. It is desirable because tests are designed to look for differences between subjects. The discriminatory power depends on the shape of the score distribution. For example, if scores are normally distributed, it is easiest to differentiate between scores at the tails because there are few extreme scores; the middle scores are hard to differentiate because they are clustered. (Ausubel, 1968)

3.1 Ferguson's Delta

Discriminatory power can be measured by Ferguson's delta, which ranges from 0 (all scores the same) to 1 (each person has a unique score). A test is considered discriminating if delta is above 0.90 (Kline, 1986). The formula for delta is given below:

$$\delta = \frac{(k+1)(k^2 - \Sigma f_i^{\,2})}{kn^2} \tag{25}$$

where: k is the number of items
n is the number of students
f_i is the frequency of each score
(e.g., if 5 people scored a 20, and 10 people scored 21, these f_i are 5 and 10)

3.2 Discriminatory Index

It is most important for each question to be discriminating. Item discrimination is measured by the discriminatory index, which compares the top-scoring students to the low-scoring students. For example, if 75% of the top students and 30% of the bottom students get a question correct, the item has a discriminatory index of 0.45. To determine the top and bottom students, the optimal split is considered to be 27% at each end (Kelley, 1939). The value 27% is derived by Kelley as the point which maximizes the difference between the means of the upper and lower groups, divided by the standard deviation of this difference.

An item is considered poor if the discrimination index is below 0.20, while above 0.40 is considered high (Ebel 1954). Other sources consider above 0.30 to be discriminating, with above 0.40 labeled "very discriminating" (Hopkins, *et al.,* 1990). Above 0.20 may also be taken as a "good rule of thumb" for acceptability (Brown, 1983). Very poorly discriminating items can have a negative discriminatory index, which means that low students answered the question correctly more often than high students.

3.3 Point Biserial Correlation

The point biserial correlation (r_{pbis}) is another measure of item discrimination. It is the correlation between scores on a question (i.e., 0 = incorrect; 1 = correct) and the overall test score. The formula can be derived from the basic Pearson correlation coefficient by assuming that the *x* variable is dichotomous. The formula is somewhat of a relic from the pre-computer days. Any statistical software will perform the correlation, such as Microsoft Excel's ™ *correl* command. The formula is given below.

$$r_{pbis} = \frac{\bar{y}_c - \bar{y}_i}{S_t} \cdot \sqrt{p_i q_i} \tag{26}$$

where: $\bar{y}_c$ is the average test score (points, not percent) of people who got an item correct
$\bar{y}_i$ is the average test score of people who got an item incorrect
S_t is the standard deviation of scores (again, points not percent)
p_i is the percentage of students who got an item correct
q_i is the percentage of students who got an item incorrect

The derivation of the point biserial correlation is conducted in a manner somewhat similar to the derivation of the expression $V_i = p_i q_i$. The first step is to recall the formula for the Pearson correlation coefficient (equation 27).

$$r = \frac{S_x}{S_y} b \tag{27}$$

where: S_x is the standard deviation of the x variable (item scores)
S_y is the standard deviation of the y variable (total test scores)
b is the slope of the regression line (sometimes denoted a or m)

The quantity S_y is the same as S_t. To arrive at the point-biserial formula, it is necessary to specialize S_x and b to the case where x is a dichotomous variable. S_x is simply the square root of the item variance, which has previously been shown to equal $p_i q_i$.

The general formula for the slope of the regression line, b, is the following:

$$b = \frac{\sum \left(x - \bar{x}\right)\left(y - \bar{y}\right)}{\sum \left(x - \bar{x}\right)^2} \tag{28}$$

The x portion of the numerator is the following:

For correct students: $\sum \left(x - \bar{x}\right) = \sum (1 - p)$
For incorrect students: $\sum \left(x - \bar{x}\right) = \sum (0 - p)$

For a class with n students, np are correct on each item with an average total score of $\overline{y_c}$.

There are $n(1-p)$ incorrect students with an average total score of $\overline{y_i}$.

Each of the x terms is also multiplied by the y portion of the sum. Again, it is partitioned into correct and incorrect students.

For correct students: $\sum \left(x - \bar{x}\right)\left(y - \bar{y}\right) = (1 - p)\sum \left(y - \bar{y}\right) = (1 - p)\left(\bar{y}_c - \bar{y}\right)np =$
$= np(1 - p)\left(\bar{y}_c - \bar{y}\right)$

For incorrect students: $\sum \left(x - \bar{x}\right)\left(y - \bar{y}\right) = (0 - p)\sum \left(y - \bar{y}\right) =$
$(0 - p)\left(\bar{y}_i - y\right)n(1 - p) = -np(1 - p)\left(\bar{y}_i - \bar{y}\right)$

Combining the expressions for correct and incorrect students, yields:

$$np(1-p)\left(\bar{y}_c - \bar{y}\right) - np(1-p)\left(\bar{y}_i - \bar{y}\right)$$
$$= np(1-p)\left[\left(\bar{y}_c - \bar{y}\right) - \left(\bar{y}_i - \bar{y}\right)\right] = np(1-p)\left(\bar{y}_c - \bar{y}_i\right)$$

Now, attention is turned to the denominator. Again, it is partitioned into correct and incorrect students:

For correct students: $\sum\left(x - \bar{x}\right)^2 = np(1-p)^2$

For incorrect students: $\sum\left(x - \bar{x}\right)^2 = n(1-p)(0-p)^2 = n(1-p)p^2$

Combining the two, the following expression is the denominator of *b*:

$$np(1-p)^2 + n(1-p)p^2 = np(1-p)\left[(1-p) + p\right] = np(1-p)$$

The numerator and denominator have the term *np(1-p)* in common, which is divided out to yield:

$$b = \frac{np(1-p)\left(\bar{y}_c - \bar{y}_i\right)}{np(1-p)} = \bar{y}_c - \bar{y}_i \tag{29}$$

And finally, the pieces are put back together to arrive at the point biserial correlation formula:

$$r = \frac{S_x}{S_y} b = \frac{\sqrt{p_i q_i}}{S_t}\left(\bar{y}_c - \bar{y}_i\right) = \frac{\bar{y}_c - \bar{y}_i}{S_t} \cdot \sqrt{p_i q_i} = r_{pbis} \tag{30}$$

4. Item Analysis

The background on item analysis has been described previously. The concepts of validity, reliability, and discrimination which apply to the test as a whole are applicable to the items. The methods utilized on the SCI are discussed in the Chapter IV. One additional consideration which has received attention in the literature is the optimal number of choices for multiple choice items; this topic is explored below.

4.1 Optimal number of distracters

Theory

Ebel (1969) derived an estimation of KR-20 as a function of the number of choices per item. The following assumptions are most crucial to this estimate: the mean of the test is the midpoint between chance-level and the maximum; the standard deviation is one-sixth of the difference between the maximum and chance-level; KR-21 is a reasonable estimate for KR-20. The reliability then becomes the following:

$$r = \frac{k}{k-1}\left[1 - \frac{9(a+1)}{k(a-1)}\right] \tag{31}$$

where: r is the reliability
k is the number of items
a is the number of choices per item (i.e., "a" as in alternative)
9 is the number nine

The standard deviation assumption is most appropriate for long tests (e.g. 100 items). A standard deviation of $\frac{1}{5}$ the difference between the maximum and chance-level is recommended for tests with 21 to 60 items. Re-deriving the reliability, the value *9* in the numerator becomes $\frac{25}{4}$, while the other portions of the formula do not change. With these assumptions, the reliability increases monotonically as the number of alternatives increases. The largest increase in reliability occurs going from 2 to 3 options, with minimal increases as a increases further. These results are merely intended to serve as guidelines, as the actual reliability will depend on factors such as item quality, homogeneity, and subject variability.

Grier (1975) examined test reliability as a function of the number of choices per item, assuming the total number of choices on a test was constant ($c=na$, where n is the

number of items, *a* is the number of alternatives per item, and the total number of options, *c*, is constant). This condition assumes that examinees will spend most of their time reading and considering the alternatives rather than reading the stem. Using Ebel's approximation to KR-21, Grier showed that three choices per item (*a*=3) provided maximum reliability for all values of *c*. Two choices (*a*=2) yielded only slightly lower reliability, with the difference most pronounced at small values of *c*; however, Ebel's approximation may not hold as *n* decreases below 18. At large values of *c*, the reliability showed little variability across *a* (when *c*=600, reliability is 0.910, 0.915, 0.890, and 0.880 for *a* of 2, 3, 4, and 5, respectively).

Practice

Costin (1970) investigated differences between three- and four-choice items in terms of discrimination, reliability, and difficulty. Four tests in introductory psychology, each with 50 or 60 items, were constructed from well-established item pools; the items were designed to measure empirical generalizations in the areas of perception, learning, motivation, and intelligence. All items originally contained four choices, and half of the items had one distracter randomly removed to create three-choice items. Each of the four tests was analyzed as two separate tests of 25 or 30 items. The tests were administered to 207 students in four sections over two semesters. The three-choice versions were superior in terms of discrimination (mean discrimination index +0.02 to +0.08) and reliability (KR-20 +0.02 to +0.10), although the differences were small. The three-choice items were only slightly easier (+1.7% to +3.6% correct), a much smaller difference than would be expected by chance guessing. In a larger follow-up study (*n*=1566), Costin (1972) found four-choice items slightly preferable in terms of discrimination (mean point-

biserial correlation +0.01) and reliability (KR-20 +0.03) on a 50-item test. The three-choice items were again slightly easier (+0.8% correct). Taken together, Costin's work suggests there are no practical differences between three- and four-choice items.

Ramos and Stern (1973) studied the difference between four- and five-choice multiple choices items on second- and third-year college-level French (n=1340) and Spanish (n=1083) reading examinations. For each language exam, the subjects were assigned to one of two groups: either four- or five-choice items. As a control measure, all students also completed a common, four-choice section of the respective test. For both language tests, the five-choice items had marginally higher reliability (α +0.04 and +0.05 on French and Spanish, respectively) and discrimination (mean point-biserial correlation +0.03 and +0.02). However, these small differences become even less meaningful considering the outstanding values even at the low end (lowest α 0.85, lowest mean r_{pbis} 0.53). On the common section of each test, the five-choice group had higher reliability (+0.01 and +0.03), further diminishing the differences.

Rogers and Harley (1999) examined differences between tests with three and four multiple choice options, using test-wiseness guidelines to remove one option from each item. The instrument was a high-stakes mathematics test given to 12^{th} grade students in Alberta, Canada. On one 40-item section of the test, two alternate forms were used: one contained items with four options, with half of the items identified as test-wiseness susceptible (Form4); the second form was altered to three options per item by revising potential test-wise items or else by removing the option which had been least chosen in the previous year (Form3).

Form4 was completed by 75 students and Form3 by 80 students. Form3 yielded significantly higher mean scores ($p < 0.05$). However, no other test metrics varied significantly between the two forms: Cronbach's alpha 0.70 on Form4, 0.75 on Form3; mean point-biserial correlation 0.315, 0.344; time to complete test (min) 106.67, 107.07. At the item level, 26 of the 31 items were more difficult on the four-item test, with seven differences greater than one (pooled) standard deviation above the mean difference. The point-biserial correlations are less clear-cut. Nine items were higher on the three-option version by at least one standard deviation, while this is true of six items for the four-option version. The remaining 16 items were within one standard deviation of the mean difference between three- and four-option point-biserial correlations.

Implications

The papers reviewed offer no insight into the development of items, as they focus on either mathematical formulations or empirical studies with well-developed tests. Taken together, the theory implies that three options is optimal, as three provides the maximum incremental increase in reliability for a fixed number of items (Ebel, 1969) and maximum reliability for a fixed number of total choices (Grier, 1975). The empirical studies are inconclusive: three is better than four (Costin, 1970); four is better than three (Costin, 1972; Rogers and Harley, 1999); five is better than four (Ramos and Stern, 1973). In the end, three-choice items are preferable for the simple reason that it is easier to develop two meaningful distracters than three or four. However, identifying which distracters are, in fact, the best will require developing more than two to begin with and then eliminating the worst distracters. The quality of the distracters is likely to play a

larger role in determining the item characteristics than any theoretical considerations about the optimal number of choices.

5. Conclusion

This chapter documented the test theory topics which were consulted in constructing the SCI. The following chapter illustrates that these topics are sufficient for inclusion in the concept inventory literature. Later chapters feature enhanced analyses when appropriate, especially Chapter IX, which analyzes the dimensionality of the SCI.

CHAPTER III

Concept Inventories

1. Introduction

This chapter describes the design and use of concept inventories. Basic scores are presented if they are available, along with comments on the reliability, validity, and discrimination. Table 1 lists the concept inventories which are cited in this document, along with the authors and title abbreviations.

Table 1: List of Concept Inventories (CI) and similar instruments

Instrument	Abbreviation	Authors (year)
Physics diagnostic instrument	none	Halloun and Hestenes (1985)
Force CI	FCI	Hestenes, *et al.* (1992)
Mechanics Baseline Test	MBT	Hestenes and Wells (1992)
Materials CI	MCI	Krause, *et al.* (2003 and 2004b)
CI of Natural Selection	CINS	Anderson, *et al.* (2002)
Chemical equilibrium (Test to Identify Students' Conceptualizations)	TISC	Voska and Heikkinen (2000)
Heat Transfer CI	HTCI	Jacobi, *et al.* (2003)
Fluid Mechanics CI	FMCI	Martin, *et al.* (2003 and 2004)
Statics CI	none	Steif (2003 and 2004) Steif and Dantzler (2005) Steif and Hansen (2006)
Thermal and Transport Science CI	TTSCI	Olds, *et al.* (2004)
Dynamics CI	DCI	Gary, *et al.* (2003 and 2005)
Wave CI	WCI	Roedel, *et al.* (1998) Rhoads and Roedel (1999)
Circuits CI	CCI	Helgeland and Rancour
Computer Engineering CI	CPECI	Michel, *et al.*
Electromagnetics CI	EMCI	Notaros
Electronics CI	ECI	Simoni, *et al.* (2004)
Signals and Systems CI	SSCI	Wage, *et al.* (2002 and 2005)
Strength of Materials CI	SOMCI	Richardson, *et al.* (2003); Morgan and Richardson
Thermodynamics CI	none	Midkiff, *et al.* (2001)
Chemistry CI	CCI	Pavelich, *et al.* (2004) Krause, *et al.* (2004a)
Conceptual Survey of Electricity and Magnetism	CESM	Maloney, *et al.* (2001)
Test of Understanding Graphs in Kinematics	TUG-K	Beichner (1994)
Force and Motion Conceptual Evaluation	FMCE	Thornton and Sokoloff (1998)
Determining and Interpreting Resitive Electric Circuit Concepts Test	DIRECT	Engelhardt and Beichner (2004)
Geoscience Concept Inventory	GCI	Libarkin and Anderson (2005)

2. Force Concept Inventory

2.1 Early work

The stimulus of the concept inventory was the work in the early 1980's to develop a diagnostic test for introductory physics courses (Halloun and Hestenes, 1985). This eventually led to the Force Concept Inventory, but it was simply called a physics diagnostic instrument at this early stage. The authors utilized pedagogical research which found that (1) common sense beliefs about mechanics usually conflict with Newtonian mechanics and (2) conventional instruction does little to change these beliefs.

The goal of the physics diagnostic instrument is to assess students' qualitative conceptions of motion and its causes. This is accomplished through carefully-written questions which identify students' misconceptions. Ideally, the instrument is used as a pre-test and post-test to measure changes in student conceptions as a result of instruction.

Over a three-year period, versions of the instrument were administered to over 1000 students in introductory, college-level physics. Early versions were open-ended, and common misconceptions were used to develop distracters for multiple choice answers.

Detailed results are presented from three sources: (1) four sections, each with a different instructor, of University Physics at Arizona State University (ASU), a calculus-based course composed primarily of engineering majors; (2) three sections, with two different instructors, of College Physics at ASU, a trigonometry-based course; and (3) one honors and one regular high school physics course, taught by the same instructor. The results are summarized in Table 2.

Table 2: Summary of Results of Physics Diagnostic Test (Halloun and Hestenes, 1985)

Instructor	**N of Students**	**Pre-Test**	**Post-Test**	**Gain**
University Physics				
A	97	51%	65%	13%
B	192	51%	64%	13%
C	70	50%	64%	13%
D	119	53%	64%	11%
College Physics				
E	82	37%	53%	15%
E	196	37%	n/a	n/a
F	127	40%	n/a	n/a
High School Physics				
G (honors)	24	30%	52%	22%
G (regular)	25	30%	44%	14%

The low test scores lead the authors to conclude that "students are prone to misinterpreting almost everything they see and hear in physics class" (p. 1045). It is interesting that 55% of students in College Physics took physics in high school, but those students scored only 2 points (6%) higher than students in the same course who had not taken high school physics. The conclusion is that pre-test and post-test scores from High School Physics and College Physics are similar to each other and also to the pre-test scores from University Physics.

Another experiment involved a small group of students who took the test at mid-semester in addition to pre- and post-tests. The mean score for these students is reported as 22.79 on the mid-test and 23.58 on the post-test. The pre-test score can be inferred to be around 18.5 from tables in the article. This corresponds to a gain of around four questions from pre-test to mid-test but a gain of less than one from mid-test to post-test, suggesting that students gain the most conceptual knowledge early in a course.

The possibility of pre-post test/re-test effects was ruled out by comparing a group of students who took only the post-test to students in the same class who took both pre-

and post-tests. The mean and standard deviation for both groups were nearly identical (no numbers are given).

Reliability was established in two ways. First, an informal test/re-test reliability was assessed via interviews with students who had taken the test. Almost without exception, students gave the same answers in the interviews as they had given on the test. The authors conclude: "the students' answers reflected stable beliefs rather than tentative, random, or flippant responses" (p. 1044). The reliability was formally established by calculating Cronbach's coefficient α to be 0.86 on the pre-test and 0.89 on the post-test. This appears to combine all groups instead of looking at the reliability only for specific courses, sub-tests, or instructors. The authors claim that similar scores for multiple choice and written versions of the instrument give "comparable results" and therefore can be said to measure the "same thing." No evidence is given to support this claim.

Several steps were taken to ensure the content validity of the instrument. First, suggestions were gathered from physics professors and graduate students, and these suggestions were included in the instrument. Second, eleven graduate students took the exam and agreed to the correct answer to each question. Third, interviews were conducted with 22 introductory physics students to ensure that they understood the questions and alternatives. Fourth, the answers of 31 students who received an A in introductory physics were analyzed to ensure that there were no misunderstandings due to question formulation. All four steps suggest that the instrument is valid.

A claim for the concurrent validity of this instrument is implied by showing that students with higher course grades tend to score higher on instrument. The results are

presented in Table 3 for one course. The authors comment that other courses yield similar correlation coefficients for instrument score versus course grade ($r = 0.56$, $p = 0.0001$).

Table 3: Course Grade compared to Diagnostic score for one course

Grade	Number	Pre-Test	Post-Test
A	31	63%	75%
B	61	55%	67%
C	66	47%	62%
D	25	43%	56%
E	9	38%	46%

Predictive validity was assessed in a stepwise regression which included the diagnostic pre-test, a mathematics pre-test, and prior physics and mathematics coursework. The diagnostic was shown to have the highest correlation with final course grade for both University (r^2=0.30) and College Physics (r^2=0.32) courses, with the mathematics pre-test slightly less predictive (r^2=0.26 and 0.22, respectively). The combined effect of coursework was much smaller (r^2=0.15 and 0.16, respectively). Using a simple linear regression with only the diagnostic pre-test, 53% of students were correctly classified according to their final course grade.

The final step is to make inferences about teaching. All the courses are taught in a lecture-recitation format, with typically three or four hours of lecture and one hour of recitation per week. The instructors show a wide range of style and quality but fall into this same basic format. One of the professors (B, Table 3) twice received awards for teaching, whereas another (D) was teaching the course for the first time and closely followed the textbook. The ultimate conclusion is that “basic knowledge gain under conventional instruction is essentially independent of the professor” (p. 1048).

2.2 Force Concept Inventory

The work on this as-yet unnamed physics diagnostic test laid the foundation for the Force Concept Inventory (FCI) (Hestenes, *et al.,* 1992), which has proven instrumental in improving physics education. The structure of the FCI is more meticulously detailed by stating the topics covered and the types of misconceptions the instrument aims to detect. Most of the results and conclusions are compatible to those in the original article by Halloun and Hestenes (1985).

The FCI's topic coverage is broken down in two manners. First, the FCI is divided based on correct Newtonian concepts into six categories: Kinematics, First Law, Second Law, Third Law, Superposition Principle, and Kinds of Force. Each category has approximately five questions, further classified into sub-topics. Only four of the 29 questions fall into multiple categories. Second, the FCI is divided into six new categories based on student misconceptions: Kinematics, Impetus, Active Force, Action/Reaction Pairs, Concatenation of Influences, and Other Influences on Motion. Each incorrect answer is then placed in a sub-topic.

The FCI's validity is established rather informally by comparing its results to those on the diagnostic. For a physics course at Arizona State, the average scores on the FCI were 52 on the pre-test and 63 on the post-test, compared to 51 and 64 for the diagnostic. Further, they have post-test scores from seven different professors which are "nearly identical" (p. 10). About half of the FCI items are the same as those on the diagnostic, as the authors felt those questions could not be improved.

Interviews were conducted to probe students' thought processes. Students tended to have firm reasons for their choices and seldom wavered between options. Among a

group of physics graduate students, who were also interviewed, three students who possessed severe misconceptions struggled in a graduate mechanics course, with two barely passing and the other failing. During interviews, the authors learned that several students exhibited difficulty reading the text. Five of these students did not speak English as a first language, but another five were native English speakers. The biggest obstacle tended to be overlooking "little words" (their quotes) such as prepositions.

2.3 Interpreting the FCI

Huffman and Heller (1995) present results of a principal-component analysis to question the interpretation of the FCI. For a group of high school students, they found only two significant factors, one of which contained all four questions on Newton's Third Law and the second of which contained three of the FCI's twelve questions on Kinds of Forces. For a group of university students, only one significant factor was found. It contained four of the twelve Kinds of Force questions and one First Law question. The authors conclude that "the questions on the FCI are only loosely related to each other and do not necessarily measure a single force concept or the six conceptual dimensions of the force concept as originally proposed by the authors" (p. 140). They also believe that responses are highly dependent on a question's context due to the lack of a fundamental understanding of Newtonian concepts.

In response to this criticism, Hestenes and Halloun (1995) write that the FCI score should be considered a measure of student disparity between Newtonian and non-Newtonian thinking. They feel that factor analysis on a population of non-Newtonian thinkers is irrelevant to the validity of the FCI. They suggest performing factor analysis

only on students who scored in the 60% to 85% range because these will most closely resemble Newtonian thinkers and help eliminate noise caused by false positives.

Heller and Huffman (1995) reply with two possible interpretations of their findings: 1) students have coherent knowledge of Newtonian concepts, but the FCI does not measure this; or 2) the FCI measures only bits and pieces of students' knowledge which do not form a coherent concept. As suggested, results are presented for a factor analysis of only high-performing students. Unfortunately, the sample sizes were too small to yield many significant correlations, and the factor analysis yielded even less information.

A potential pitfall of the FCI is mentioned in the comparison of results at Arizona State University (ASU) and the University of Minnesota (UM) in a later study (Heller and Huffman 1995). A basis for the validity of the original diagnostic instrument, and by extension the FCI, was that the post-test was highly correlated with the overall course grade ($r = 0.56$, $p = 0.0001$) at ASU. However, at UM, the correlation was too low to accurately predict student success in an introductory physics course ($r = 0.27$). In light of this, the authors urge caution in using the instrument at other institutions until it has been validated for that student population.

Rebello and Zollman (2004) questioned the validity of distracters on the FCI, using four items on the FCI. In a pilot study, 25 students in second-semester physics and 238 students in first-semester physics (both algebra-based and at the beginning of the semester) answered open-ended versions of FCI items. The design was counter-balanced so that each student answered two open-ended items along with two original FCI items. It was found that FCI distracters do not necessarily capture all possible responses and not in

the correct proportions. In a follow-up study, the four items were administered to 234 students at the beginning of an algebra-based introductory physics course. Three versions of distracters were used: 1) original FCI; 2) modified based on responses to open-ended pilot study; 3) the union of (1) and (2). From this experiment, several interesting conclusions are reached. First, the FCI appears to accurately measure percent correct compared to the open-ended versions. However, the FCI distracters may not capture all possible responses and may alter response tendencies as well. Further, the revised distracters proved more effective than the originals in some instances. Although the number of items examined is small, the study raises an important question about the FCI's ability to diagnose misconceptions.

2.4 Uses of the FCI

Hake (1998) uses FCI scores to provide a comprehensive comparison between interactive engagement (IE) and traditional teaching. The results are striking: 14 traditionally-taught courses had average normalized gains of 0.22, while 41 IE courses had normalized gains ranging from 0.34 to 0.69; what's more, the *lowest* IE course was above the *highest* traditional course. These results include high school, college, and university courses, with a total of 5832 students.

The FCI has been used to assess the effectiveness of the Peer Instruction (PI) teaching method (Crouch and Mazur, 2001, for a review), which advocates conceptual mastery by engaging all students in small-group discussions. FCI normalized gains of 0.49 to 0.74 attest to the effectiveness of PI, compared to normalized gains of 0.25 and 0.40 for two courses taught using traditional methods.

2.5 Mechanics Baseline Test (MBT)

Two of the three authors of the FCI developed a complementary instrument entitled the Mechanics Baseline Test (MBT) (Hestenes and Wells, 1992). It is conceptual in nature, but approximately one-third of the problems require simple calculations (e.g., tension in a rope). The MBT is recommended for use as a post-test for introductory physics courses but could be used as a placement test for advanced courses.

Data suggest that good performance on the FCI is a necessary condition for good performance on the MBT. A score of 60% on the FCI is considered as "a kind of conceptual threshold for problem-solving competence on physics" (p. 161). This is evident by examining a graph ("Fig. 1" in the article) which shows FCI Post-Test scores vs. MBT Post-Test scores for 26 courses. There are 16 courses with average scores below 55% on the FCI. All but one of these courses scored below 35% on the MBT; it should be noted that these are high school courses. Ten courses scored near or above 60% on the FCI; these courses show a strong linear relationship between FCI and MBT scores (estimated $r = 0.67$). The lowest MBT scores among this group are nearly equal to the highest MBT scores among the lower group. Only one course performed better on the MBT than the FCI; this course is referred to as an outlier below because it does not fit the pattern of other classes which performed better on the FCI than the MBT. The graph from the article is re-produced in Figure 1.

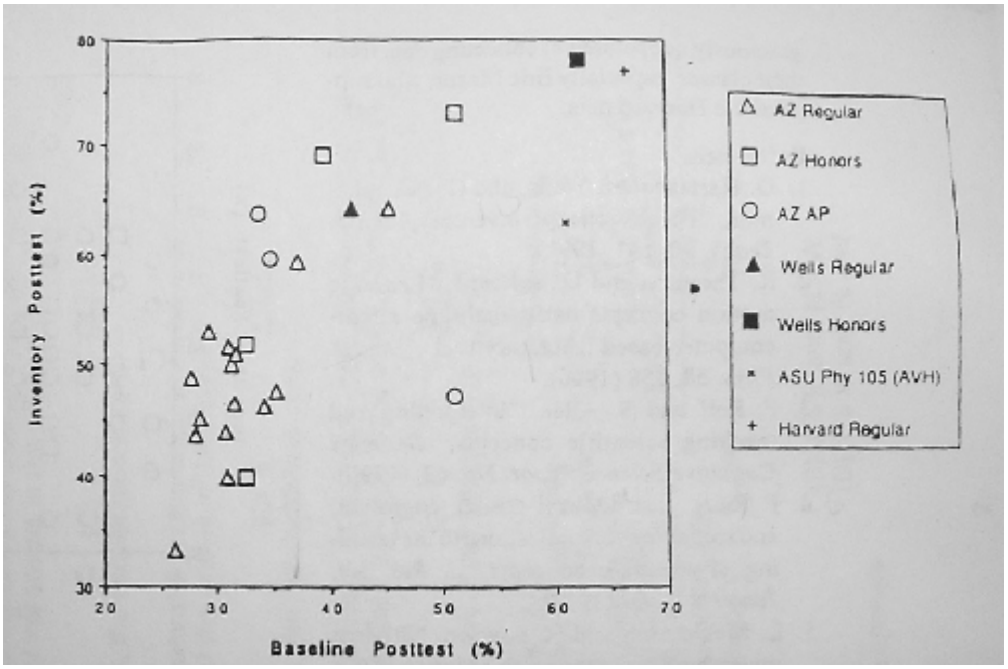


Figure 1: FCI Scores vs. MBT Scores (Hestenes and Wells, 1992)

No correlation coefficients are given in the article. By estimating the values on the graph, correlation coefficients have been calculated. Logarithmic and square root transformations are included because the data appear to follow these patterns when viewed as a whole. The data are reported with and without the outlier in Table 4.

Table 4: Correlation Coefficients for data in Figure 1

Relationship	**Correlation including outlier**	**Correlation with outlier removed**
FCI vs. MBT	$r = 0.74$	$r = 0.80$
FCI vs. ln(MBT)	$r = 0.76$	$r = 0.83$
FCI vs. sqrt(MBT)	$r = 0.75$	$r = 0.81$
FCI vs. MBT, low group only	$r = 0.44$	n/a
FCI vs. MBT, high group only	$r = 0.67$	n/a

There is essentially no discussion of the validity or reliability of the MBT. The validity and reliability can be loosely implied by the comparison to the FCI and the earlier physics diagnostic instrument, which share many questions.

Kim and Pak (2000) used the MBT as a mid-point of a study comparing students' self-reported copiousness in solving textbook problems in preparation for university entrance exams with their results on conceptual quizzes. Solving a great number of textbook problems proved to be no aid in performing well on the MBT: in fact, a slight negative correlation was found (r = -0.15), suggesting that students who solved additional problems (up to 2900) did so to compensate for lack of understanding. The MBT suggests a slight degree of predictive validity (r = 0.31) for performance on conceptual quizzes, while the number of problems solved has essentially no value (r = -0.04) as a predictor of conceptual understanding.

3. Engineering Concept Inventories

Following the successful implementation of the FCI, concept inventories have been developed in many fields of engineering. Most of these are affiliated with the Foundation Coalition, a National Science Foundation-funded coalition headed by D.L. Evans of Arizona State University. The concept inventories fit the Foundation Coalition's goal to improve curricula and learning environments. It is informative to compare the structure and development of these concept inventories to the FCI and the Statistics Concept Inventory. Overviews (e.g., format and scores) of several of these instruments are presented in this section. Details are presented as much as they are available, but it will be noted that several of these instruments are either stalled or abandoned along the educational research highway.

3.1 Materials Concept Inventory (MCI)

The Materials Concept Inventory (MCI) states that the "overall goal is to analytically link relationships of scientific fundamentals to macroscopic materials

behavior" (Krause, *et al.,* 2003). The primary topic areas are atomic structure and bonding, band structure, crystal geometry, defects, microstructure, and phase diagrams. The questions are applied to metals, ceramics, polymers, and semiconductors.

The MCI contains 30 questions, with ten based on previous knowledge of chemistry and geometry, and 20 based on content from a materials course. Concept inventories in thermodynamics, chemistry, and mechanics were consulted for additional information and content. Student misconceptions were gathered from chemical education journals. The initial distracters were from a faculty survey of students' conceptual difficulties. To find more "authentic" distracters, the researchers conducted weekly student interviews and gave short-answer, open-ended "intuition quizzes" during materials courses.

The MCI was administered to several classes of 16 to 90 students at Arizona State (ASU) and Texas A&M (TAMU) in Summer and Fall 2002. Most classes had a limited gain of 15% to 20%; however, one class that used active learning had a gain of 38%. The low end gains are comparable to those on the early testing of the physics diagnostic instrument for traditional lecture courses.

Four questions are discussed in detail in their conference paper. The first question (background geometry) is interesting because it had large differences between institutions on the pre-test (61% ASU, 79% TAMU). However, the gap was narrowed on the post-test (81%, 88%); the difference on the pre-test is likely due to students at TAMU taking a CAD class as freshmen. The post-test is interesting because a large portion of the ASU students drew a wire cube to help visualize the question.

The second question (background chemistry) relates to phase diagrams and solubility. The distracters were based on intuition quizzes. Students have a "prior misconception," concluded from relatively low percent correct on the pre-test (39% ASU, 50% TAMU). This misconception is not completely overcome and becomes a "persistent prior misconception" because there is not complete mastery on the post-test (67%, 66%).

The third question covers a topic learned in a materials course (conductivity). It is interesting because of the different gains between institutions (ASU: 20% pre, 75% post; TAMU: 35% pre, 33% post). The difference is likely due to ASU focusing more on electrical properties, whereas TAMU focuses on mechanical properties

The fourth question also covers a materials-course topic (strength). Intuition quizzes gave very good distracters, which caused pre-test scores to be well below guessing (8% ASU, 7% TAMU, with 5 choices). The authors believe that students developed "spontaneous misconceptions" based on prior experience and course knowledge, and these persisted on the post-test (19% ASU, 15% TAMU).

The MCI authors have also documented the focus group efforts used to revise the instrument (Krause, *et al.,* 2004b). Focus groups held for the initial study were not found to be as informative as desired. For the new focus groups, students were given only selected MCI questions, which helped to steer the discussions. Group size was increased from two or three to six to ten, which made students more comfortable and willing to speak. The students were first given their 10 to 12 selected questions to answer individually. They then met in their group to discuss why they had chosen certain answers but were not told by the moderator which answer was correct until discussions

had concluded. This format proved informative for developing and validating distracters and also helped students gain a deeper understanding of the material.

Jordan, *et al.* (2005) is not in the normal mold of concept inventory papers because it was not written by the MCI authors and does not include data from the MCI authors' schools. It might, therefore, be considered a less biased account of the use and interpretation of the results. One important difference is that the Jordan group teaches a two semester-hour course, whereas the MCI was designed for a three semester-hour course. The initial results were disappointing, with 32% correct on pre-test and 39% on post-test. However, when adjustment was made for topics not covered, the scores increased to 37% and 49%, respectively, slightly lower than the 15% to 20% originally reported by the MCI authors. The concurrent validity of the instrument is assessed by computing average MCI scores for students grouped by letter grade. The MCI scores formed three groups – High A and Low A students averaged 63% and 60%, respectively; High B, Low B, C, and D students averaged between 30% and 42% correct; F students averaged 17% correct. Based on the groups, a correlation (r^2) of 0.9403 is reported, but this grouping is not a valid way to perform the correlation analysis. To overcome the topic coverage issue, it is proposed to write a new concept inventory for a two-hour course. However, it is not clear if this means writing an entirely new MCI or merely allowing an *a la carte* use of the existing MCI.

3.2 Statics Concept Inventory

The Statics Concept Inventory is designed to "detect errors associated with incorrect concepts, not with other skills (e.g., mathematical) necessary for Statics" (Steif, 2004). Questions which contain numbers require only trivial calculations, and incorrect

answers are based on incorrect assumptions rather than incorrect calculations. The results presented are from a pre-test for a Statics course, but most students had some exposure to the topics during an introductory mechanical engineering course. The average score was 10.6 of 27 (39.3%).

The questions on the Statics Concept Inventory were developed primarily through the experience of the author and two statics professors at different universities. Distracters were also based on students' written responses to questions which require multiple statics concepts. Based on these findings, the instrument is divided into five sub-topics, one of which contains four further specific situations.

Five items are presented, but little detail is given beyond the thought process for the distracters. These five items contain diagrams with forces acting on various objects. Several of the misconceptions are: forces are missing from the diagrams; extra forces are included; incorrect understanding of a couple; incorrect understanding of normal forces; inability to balance forces in equilibrium.

The instrument has 7 of 27 items with discriminatory indices below 0.20 on a pre-test, with 12 of 27 above 0.40. Three of the low discriminating items relate to friction and another three relate to static equivalence. The friction questions had very low percent correct, which limits the discriminatory index. The static equivalence questions were not quite as low, but other research has shown that misconceptions in this area persist even after a statics course.

The Statics Concept Inventory is reported to have an alpha of 0.712 as a pre-test; no post-test results are reported. The author considers the instrument reliable for "an initial version" but would like to attain a value above 0.80.

In a second study, Steif and Dantzler (2005) perform more psychometric analyses of the Statics Concept Inventory using data from 245 students at five universities. Total scores indicated no significant differences due to gender or ethnicity in a 2×2 ANOVA. The instrument was highly reliable with α of 0.89. At one university (n = 105), a strong correlation was found between course grade (coded A=1, B=2, C=3) and inventory score (Spearman's ρ = -0.547, $p < 0.001$; ρ is negative because higher grades are coded with lower numbers). At this same university, mean scores nearly doubled from pre-test (39.2%, n = 127) to post-test (75.3%, n = 105).

Items performed very well on the discriminatory index, with values ranging from 0.26 to 0.84, with only the former below 0.30. Difficulties were also in a preferred range, from 31% to 85%. A confirmatory factor analysis (CFA) model was fit to the item scores with each item assumed to measure one of eight hypothesized constructs. The fit indices suggest the model is "acceptable" (e.g., goodness-of-fit 0.90), although there is some room for improvement.

Further results (Steif and Hansen, 2006) are tempered, although still encouraging. For a survey of 1331 students at seven universities, the reliability was 0.82 and four items had a discrimination index below 0.30. For five courses, the correlation (Pearson's r) between inventory score and course final exam score ranged between 0.24 and 0.62. Two additional courses provided data from two mid-term examinations in addition to the final. The inventory correlated most-highly with the first exam (r = 0.65 and 0.46), while the correlations were mixed across second-exam/inventory and final-exam/inventory. Further, six classes provided data which showed that correlations across test scores *within each class* were of similar magnitude to the various inventory-exam correlations (read:

there is no pattern of course exams being better or worse predictors of future exams than the inventory as a predictor of exam scores). These data were collected from a online test [http://engineering-education.com/CATS/intro.htm], and more recent results are available therein.

3.3 Thermal and Transport Science Concept Inventory (TTSCI)

The Thermal and Transport Science Concept Inventory (TTSCI) alpha version has 15 questions (Olds, *et al.,* 2004). Five items ask for a reason for choosing the answer to a previous question. The TTSCI was completed by 66 of 93 students in two senior-level courses (one chemical engineering, one mechanical engineering). All students had taken at least one course in thermodynamics, heat transfer, and fluids. The test was given as a take-home extra credit assignment. The scores on the items range from 28.8% to 71.2% correct, and the overall average is 52.2%. Based on overall scores, there were no significant difference by gender, but chemical engineering students significantly out-performed mechanical engineering students. This may be because the primary author of the test is a chemical engineer and inadvertently biased the items towards the chemical engineering curriculum. The authors plan to test this hypothesis at other universities.

The process began with asking faculty to rate 28 topics for the degree to which undergraduates understand the topic and how important it is for them to understand. The list was pared down to 10 topics which respondents felt were important but not adequately learned by students. The topics include 2nd Law of Thermodynamics, steady-state vs. equilibrium processes, and energy-related topics (heat, temperature, enthalpy, internal energy). From these 10 concepts, sample questions were devised and tested in "think aloud" sessions with six undergraduate students. This helped ensure students were

properly interpreting the questions and gave insight into the students' thought processes which helped develop distracters.

To revise the instrument for further use, the authors plan to revise choices which were chosen by very few students by conducting another faculty survey. After a round of beta testing, the plan is to analyze the reliability, construct validity, and predictive validity of the TTSCI to give more concrete evidence of its usefulness.

3.4 Wave Concepts Inventory (WCI)

The Wave Concepts Inventory (WCI) was designed to examine differences between a traditional course in electromagnetics and an integrated course (Roedel, *et al.,* 1998). Both courses contained over 50% seniors, with the remainder primarily juniors and a few graduate students. The results show that the integrated course had a significantly higher knowledge gain (+3.4 questions from pre-test to post-test) compared to the traditional course (+0.9). The traditional course's gain was not significant compared to zero. Further analysis shows that every student in the integrated course improved from pre-test to post-test, but only around half of the traditional class improved.

The WCI is different from other concept inventories in that it allows more than one correct answer to some questions. One of the choices should be obvious to students with a basic understanding, but the other correct answer is indicative of deeper understanding. There are 20 items with a total of 34 possible correct answers.

Later work framed the WCI in terms of Bloom's taxonomy, integrating focus group discussions to strengthen conclusions (Rhoads and Roedel, 1999). Seven students participated in the focus group who had taken the WCI the previous semester. Students

gained their Knowledge skills (Bloom terminology) from introductory physics courses, whereas higher-level skills such as Analysis and Synthesis were gained in the most recent course. The focus group comments also helped to guide revision by uncovering misunderstanding about the instructions (i.e., multiple correct answers) and identifying confusing wording.

3.5 Heat Transfer Concept Inventory (HTCI)

The Heat Transfer Concept Inventory (HTCI) (Jacobi, *et al.,* 2003) has a greater level of student involvement than most other concept inventories. The abstract of the conference paper states that the process was initiated with "student identification of the conceptual problems rather than with faculty perceptions of student misunderstandings." However, the body of the paper seems contradictory by stating that faculty were asked for input about concepts they felt were important for subject mastery.

Students were hired to participate in the development during Spring and Summer 2003. They were selected to ensure diversity and to be representative of the makeup of the classes. Activities consisted of assignments to identify and explain concepts, participating in video-taped discussions, and development of concept questions.

The first assignment was to make two lists based on the course syllabus: (1) 10 topics you are *most* confident of that you think are important and (2) 10 topics you are *least* confident of that you think are important. This statement is mildly ambiguous, but it seems to ask the students for 10 important and unimportant topics that they are confident or not confident, respectively, about knowing. The lists indicate that students are good at calculations but are uncertain of basic concepts (e.g., "the difference between convection, conduction, and radiation").

For the second assignment, students selected two topics which they felt confident about and wrote three to six statements to demonstrate their understanding of the concept. Even though they could use a textbook, the "statements showed various degrees of confusion."

After completing the written assignments, the students participated in video-taped group discussions. The common observations from these sessions are summarized below.

- Students are not really sure about anything; they say "I think" or "I'm not sure but" to preface many of their statements.
- They cannot make connections between different things they know.
- There did not appear to be misconceptions, but rather a poor grasp of the concepts.
- Students are comfortable with a lack of understanding as long as they can solve the problems.
- Students lack a technical vocabulary.
- Students do not understand why topics are covered; this relates to not making connections between what they know.

Based on these findings, the authors plan to proceed with the development of questions and then administering the HTCI.

3.6 Fluid Mechanics Concept Inventory (FMCI)

The Fluid Mechanics Concept Inventory (FMCI) (Martin, *et al.,* 2003) is being developed by the same group as the HTCI. Most of the methods are the same, such as listing 10 topics and video-taping student discussions. The process is further along than the HTCI, with questions already written. Five questions are discussed in the paper, but

no scores are given. The instrument contains multiple questions on most concepts, usually varying in difficulty. Graphical representations are important for the FMCI because it contains topics such as velocity profile of a fluid in a pipe. Some questions contain units of measure, which could aid students who are unable to solve the problem based solely on their conceptual understanding.

Four of the six items presented have point biserial values about 0.30, although the highest of these is 0.39 (Martin, *et al.,* 2004). The r_{pbis} is also calculated for all incorrect answers. Presumably, this is done by assuming each incorrect answer is correct (i.e., scoring it as 1) and performing the correlation. A negative r_{pbis} on an incorrect answer is taken to mean that the distracter is appropriate because people with low scores would tend to choose it, while people with high scores would tend not to.

3.7 Dynamics Concept Inventory (DCI)

The Dynamics Concept Inventory (DCI) is designed based on the need to quantitatively assess innovated teaching practices in mechanics education (Gary, *et al.,* 2003). The first step in the development was for each team member to devise a list of the topics which are most important to dynamics. Simultaneously, a survey was conducted of faculty members for the same purpose. The 25 faculty members ranged from two-year colleges to research universities. The faculty members were asked to identify topics which they felt students had conceptual difficulty learning, as opposed to difficulty with problem-solving. The focus was on rigid body dynamics because the authors felt the FCI effectively assessed particle dynamics. This initial survey yielded 24 topics which were passed on to step two. The participants were then asked to give the percentage of students they believed adequately learn the topic (re-scaled on a 0 to 10 scale) and the importance

of the topic (also on a 0 to 10 scale). The list was pared to 13 topics by eliminating those with average importance scores below 8. Several topics with similarities were combined and an extra topic was included due to the authors' lists of the important topics, to yield a final list of 11 topics.

Each member of the DCI team wrote questions for two of the topics identified in the survey described above, including distracters (Gary, *et al.,* 2005). The questions were critiqued by other team members until a consensus was reached. Student participation was then incorporated. The first focus groups responded to open-ended versions of the questions to help further identify and refine distracters. In later focus groups, the students were divided into multiple-choice and open-ended groups.

In 2003-04, the DCI was administered at two universities. The scores range from 31% to 35% on pre-tests and 56% to 63% on post-tests. Cronbach's α is generally acceptable, reported as 0.730 and 0.837 on post-tests and 0.640 and 0.719 on pre-tests. Post-test scores from Fall 2003 and Spring 2004 at one university are analyzed for possible pedagogical differences. The Spring 2004 course performed significantly ($p < 0.05$) better on 10 of the 11 topics. Students in Spring 2004 were given short weekly conceptual quizzes and results were discussed to analyze misconceptions. The authors believe this accounts for the higher post-test scores in Spring 2004, although it cannot be stated conclusively because no pre-test was given in Fall 2003 (i.e., did the two groups start at the same points?).

Four items are presented on which students attain large gains from pre- to post-test, at the two different universities. On each of these, answers are distributed almost entirely between the correct answer and only one of the distracters (i.e., three of the five

choices are seldom selected), on both pre- and post-tests. Two other items are presented which indicate little gain or even loss of understanding. Discrimination indices can be loosely estimated from a table containing student scores grouped by quartile. One of the questions with poor performance has low discrimination for both universities, and two different items have low discrimination at each university. The other items appear to have discrimination indices above 0.30.

3.8 Circuits Concept Inventory (CCI)

The developers (Helgeland and Rancour) met with three graduate students to brainstorm question ideas, with topics determined from the Circuits I and II course catalog descriptions. The initial version given to students had multiple correct answers for many of the questions and varying number of total choices. This made data analysis too difficult, and a revised version was written with exactly four choices (only one correct) per item.

The above process took place during 2001 and 2002. The CCI as of early 2003 contained 43 questions in 12 topic areas, ranging from 1 to 10 questions per topic). Future versions are expected to include 9 additional topics. There are no publications available and no more current information.

3.9 Computer Engineering Concept Inventory (CPECI)

The Computer Engineering Concept Inventory (CPECI) is in its early development (Michel, *et al.*). The core concept areas are digital components, computer architecture and organization, and programming fundamentals. Thirty questions have been developed as an alpha version, and thirty-seven more are under review for inclusion as well. The goal is to have 20 to 30 questions in each of the core concept areas.

3.10 Electromagnetics Concept Inventory (EMCI)

The Electromagnetics Concept Inventory (EMCI) is intended for junior-level courses in Electrical Engineering departments but might be applicable in other areas (Notaros). The EMCI contains three tests: Fields, Waves, and Fields & Waves. The tests contain 23, 23, and 25 questions, respectively. The Fields test is intended for a first-semester electromagnetics course in a two-course sequence, with the Waves test targeted at the second semester. The combined Fields & Waves test is designed to cover all topics taught in undergraduate electromagnetics. The test selection and sequencing is up to the instructor depending on the curriculum. The Fields and Waves test each contain 16 topics, with the number of questions ranging from 1 to 4 per topic, with some questions covering multiple topics. Version 1.0 of the instrument was developed during 2000 and 2001. Student interviews were to be conducted after administering the instrument to Electromagnetics I and II courses at The University of Massachusetts Dartmouth, but no more up-to-date information is available.

3.11 Electronics Concept Inventory (ECI)

The Electronics Concept Inventory (Simoni, *et al.,* 2004) is in its development phase, with the intended audience a first-semester (of two) course on topics such as diodes and resistors. The paper details four heuristics that are used to develop items: 1) items should cover a single concept; 2) computation should be eliminated; 3) incorrect answers should represent students' misconceptions; and 4) non-standard terms and definitions should not be present. The instrument has been used in five courses at the authors' institution (Rose-Hulman Institute of Technology) and two courses at other schools. A focus group was conducted with 13 students from Rose-Hulman to provide

feedback on item wording. Limited feedback has been gathered from external faculty. The authors plan to ask external faculty to assess current item validity, identify ambiguities or confusing terminology, generate one new item, and suggest one item for deletion. The beta draft of the ECI contains 31 items in five topic areas (four to nine items per topic). Background items in Basic Circuit Analysis are included for two reasons: 1) to assess if misconceptions in advanced topics are due to basic misconceptions; and 2) to boost student morale while taking the test.

3.12 Signals and Systems Concept Inventory (SSCI)

The Signals and Systems Concept Inventory (Wage, *et al.,* 2005) has two versions: continuous time (CT) and discrete time (DT). Each instrument contains 25 items designed to assess students' conceptual understanding of topics covered in undergraduate linear systems and signals courses, arranged in five and six topic areas, respectively. Some items fall into multiple areas because they require synthesis of multiple topics. This makes interpretation difficult, but the authors feel these types of items are crucial to an understanding of Signals and Systems.

Early work (Wage, *et al.,* 2002) assessed the validity of the instrument in terms of its performance across gender and race. No significant differences were found in any of the combinations presented, which speaks well for the fairness of the instruments.

Over a three-year period, the SSCI has been given to over 900 students at seven universities (Wage, *et al.,* 2005). The alpha version of the CT-SSCI was too long and difficult, with most students struggling to finish in one hour and a mean score of 29.5%. Revisions were made, eliminating the least-chosen distracter from each item and deleting

some items to arrive at 25 items per instrument. The primary results of the paper are based on "Version 2.0" of the SSCI.

Twenty courses are included in the major portion of the study. Fifteen of these are traditional lecture courses, while five are defined as interactive-engagement (IE), in the spirit of Hake's FCI work. The traditional lecture courses achieved gains of 0.20±0.07, while the IE courses attained gains of 0.37±0.06. Scores for the CT and DT versions are similar at both pre- and post-instruction. To assess validity, a number of correlations between SSCI (pre-test, post-test, gain) and course grades (e.g., CT systems & signals, DT systems & signals, calculus, overall GPA) are computed. To highlight several of these, both CT-SSCI and DT-SSCI post-test and gain have significant positive correlations with their respective course letter grades (4.0 scale). The CT course grade also correlates significantly with the DT-SSCI pre-test and post-test. Circuits course grade correlates positively with the CT-SSCI pre-test. GPA correlates positively with both SSCI pre-test scores. The authors feel this type of correlation analysis provides valuable insight into the role of course sequencing and can evaluate whether signals and systems courses are conceptual in nature. No information on reliability or discrimination is available.

3.13 Strength of Materials Concept Inventory (SOMCI)

Development of the Strength of Materials Concept Inventory (SOMCI) appears to be stalled (Richardson, *et al.,* 2003; Morgan and Richardson). A Strength of Materials course usually follows Statics, and the knowledge should be new, which may preclude pre-testing. The initial SOMCI was developed after meeting with David Hestenes, one of the developers of the Force Concept Inventory, who offered tips for the SOMCI. The

initial SOMCI was developed during Spring 2001, then shared with SOM professors. A revised version was administered in Summer 2001. Interviews were conducted to identify poorly written items. After more revisions, the instrument was given to 60 students in an SOM course and 60 students in a follow-up to SOM. Eighteen of 25 items are of acceptable difficulty, defined as between 20% and 80% correct. Fifteen of the 25 items have discrimination indices above 0.20, with 11 of these above 0.40. Scores on the instrument correlate positively with course grade but not at a statistically significant level ($r = 0.343$, $p = 0.069$). Three items are presented, but there is no discussion.

The authors offer a blueprint for continuing development. The first step is to develop a list of concept definitions, including colleagues in the process. The second step is to draft questions. Working independently, the team generated 55 questions, but many of these were discarded or revised because they focused on more than one concept. The multiple choice answers are to be sought from students taking the test open-ended. The third and final step is to build a working version of the inventory using data from the first two steps. The authors plan to give it to as many students at as many universities as possible. No further work has been presented since the 2003 conference paper.

3.14 Thermodynamics Concept Inventory

Topic coverage of the Thermodynamics Concept Inventory is weighted towards what a student is expected to know from chemistry and physics upon entering a thermodynamics course rather than what is learned in the course (Midkiff, *et al.,* 2001). The beta version of the instrument contains 30 items in six topic areas, ranging from 1.5 to 11 items per area. Distracters were developed by recalling common student misconceptions encountered while teaching thermodynamics. Although no data are

given, the authors list several insights gained from initial testing. They would like to shorten the instrument, possibly make it more pictorial, and relate items to more real-world experience.

3.15 Chemistry Concept Inventory (CCI)

The Chemistry Concept Inventory (CCI) was motivated by the overlap of chemistry concepts with other engineering fields, such as Materials (Pavelich, *et al.,* 2004). The authors brain-stormed and arrived at two tests, each with three topics and each topic having one to four sub-topics. The two CCI tests (Chem I and Chem II) contain 30 and 31 items, respectively. The post-test scores are 49.1% and 59.8% correct, with reliability (α) of 0.7883 and 0.7855. Revisions were made based on discrimination, reliability, difficulty, and expert opinion to arrive at two 20-item tests. These revised instruments were administered at a university (Chem I) and a community college (Chem II). The post-test scores were 53% and 55%, respectively, corresponding to gains of 26% and 19% from the pre-tests. Chem I was reliable, with an alpha of 0.7135, but Chem II was not reliable, checking in at an alpha of 0.4188. Seven items from Chem I were discussed in focus groups, leading to further refinement of distracters and wording. Only pre-test scores are available from the subsequent version of the instrument, but these are very close to those on the previous iteration.

Another publication on the CCI (Krause, *et al.,* 2004a) contains essentially the same information described above. Additional information about student interviews indicates the importance of including little words (e.g., “each”) for clarity.

4. Other Concept Inventories

4.1 Concept Inventory of Natural Selection (CINS)

Engineering is not the only discipline to pursue concept inventories. The Concept Inventory of Natural Selection (CINS) is intended for use in biology courses (Anderson, *et al.,* 2002). The instrument has three passages approximately five sentences in length related to different animals. Each passage is followed by six to eight questions about the situation. There are 20 questions total.

Distracters were generated from responses to open-ended items by non-biology majors, since biology majors presumably already understand evolution (Anderson, *et al.,* 2002). The authors also examined relevant literature, which tended to include and extend ideas gathered from the open-ended items. The responses were designed to "distinguish between different basic assumptions about the nature of the universe" (p. 957).

In the initial round of testing, students took the test and then participated in one-on-one interviews that asked more in-depth questions. The results indicate that the score on the CINS is generally correlated with the score from the interview, which matches findings from the FCI. For the second round of testing, objective analysis was incorporated. This included scoring the readability of the items by removing approximately every seventh word from the stem and seeing if students could fill in the blanks. This exercise illustrated that the item stems were appropriate for the target audience. The test was expanded to include more concepts and some items were improved based on these results.

A principal component analysis was conducted on the item phi correlation coefficients. A varimax rotation was used, and solutions with two to eight components

were examined. The solution with seven components was found to be optimal because it had (a) a large proportion of the variance explained (53%); (b) all items loaded at least 0.40 on at least one component; (c) only one item loaded at least 0.40 on multiple components; and (d) nine of the ten evolution concepts grouped along the same respective component.

The CNIS is reported to have an alpha value of 0.58 for one class and 0.64 for another. The authors consider this acceptable for a classroom test. Six of the 20 questions had undesirable values for the point biserial value (< 0.30).

4.2 Chemical equilibrium (Test to Identify Students' Conceptualizations – TISC)

An instrument is being developed to identify concepts used when solving chemical equilibrium problems where Le Chatelier's principle is applied (Voska and Heikkinen, 2000). The test is referred to as the Test to Identify Students' Conceptualizations (TISC). The test has 10 multiple-choice questions. Each question also requires the student to put a reason for his selection. A question's letter choice and reason are scored separately.

The development process of the TISC is similar to other concept inventories and includes literature reviews and content reviews by professors. An additional step involved students taking the test and meeting a few days later to go over their answers and reasons with an instructor. Interviews showed that the most common change was when students who had incorrect reasons on the test gave a new incorrect reason (30%); the least common change was when students who gave a correct reason on the test changed to an incorrect reason (4%). The TISC correctly identified both the answer and reason given by an interviewee 47% of the time. Based on this, the authors claim "that TISC possesses a

moderate level of construct validity" (p. 165). However, this is more similar to what other test developers refer to as content validity; construct validity is typically established through factor analytic techniques. An alpha of 0.79 is reported, which is excellent for a 10-item instrument and nearly at the widely-accepted value of 0.80.

There tends to be a strong relationship between reasons and answers. The conditional probabilities are shown below. The most telling relationships reveal that students who possess correct reasons nearly always respond correctly to the multiple choice option.

Table 5: Answer and Reason Conditional Probabilities on TISC

Combination	**Conditional Probability**
Correct Answer if Correct Reason	P(A|R) = 0.99
Correct Answer if Wrong Reason	P(A|R') = 0.32
Correct Reason if Correct Answer	P(R|A) = 0.59
Correct Reason if Wrong Answer	P(R|A') = 0.01

4.3 Conceptual Survey of Electricity and Magnetism (CESM)

The Conceptual Survey of Electricity and Magnetism (CESM) is designed to assess students' qualitative understanding of electricity and magnetism in introductory physics, both algebra- and calculus-based, as a pre-test and post-test (Maloney, *et al.,* 2001). Because electricity and magnetism are broad fields, separate instruments were initially written for electricity (CSE) and magnetism (CSM). A group of experienced physics professors developed the topic list and draft items at a workshop. Open-ended responses were used to refine distracters, and the best items from the two instruments were combined into the CSEM. After three more rounds of testing, the instrument contained 32 items in 11 topic areas.

Test items are analyzed according to difficulty and discrimination. Difficulty ranges from 0.10 to "a little over 0.8" (fraction of correct responses), which the authors

consider a reasonable range. However, they would prefer more easy items, as only seven of the 32 were answered correctly by over 60% of students. Discrimination indices ranged from 0.10 to 0.55, with only four items below 0.20.

The content validity was assessed by asking 42 physics professors to rate the reasonability and appropriateness of the items. On a scale of 1 to 5, with 5 being the highest, all items averaged above 4 in terms of reasonability and appropriateness for either algebra-based or calculus-based physics courses; most items scored above 4.7.

Reliability of the CSEM is assessed with Cronbach's alpha. Values of "around 0.75" are reported, which the authors consider good. Principal components analysis was conducted as well. Eleven factors were identified with eigenvalues greater than 1, but the largest of these accounted for only 16% of the variance. The authors do not consider this result meaningful and advise against adding additional items which may flesh out structure.

Pre-test scores on the CSEM are 25%, 31%, and 41% for algebra-based, calculus-based, and Honors calculus-based courses. Post-test scores are 44%, 47%, and 69% for these same groups and 70% for a "Majors/Grad" group. The CSE and CSM were given separately to some algebra- and calculus-based courses as well. The post-test scores are close to the respective CSEM scores. The CSE pre-test scores are close as well, but the CSM pre-test scores are lower.

4.4 Test of Understanding Graphs in Kinematics (TUG-K)

The Test of Understanding Graphs in Kinematics (TUG-K) was developed to assess students' ability to interpret graphs (Beichner, 1994). Beichner, citing other research, feels that interpreting graphs is an important gateway to higher understanding.

Eight objectives were formulated from studying physics textbooks and test banks; one objective was dropped after a pilot study because it was too easy. Three items were written for the remaining objectives, a total of 21 items. Open-ended items were used to help develop distracters. The draft version was sent to 15 high school and college physics instructors to assess the content validity.

To estimate test/re-test reliability, four groups (three high school, one university) took the TUG-K after being exposed to kinematic concepts. One week later, after participating in laboratory exercise, students took a slightly-modified TUG-K and a correlation coefficient (r) of 0.79 was found between the alternate forms of the instrument. A statistically significant increase ($p<0.01$) in test scores was also achieved as a result of the laboratory activities. These results guided further revision, and a new version was administered to 524 college and high school students.

A KR-20 of 0.83 was attained, which is quite good for a test of this length. The point-biserial correlations averaged 0.74, compared to a desired value of ≥ 0.20. Ferguson's delta, measuring overall discrimination, was 0.98, again more than adequate (≥ 0.90). The average discrimination index was 0.36, which is quite a bit lower than the average point-biserial correlation but still above the acceptable 0.30.

The overall scores are disappointing to the author, averaging only 40%. Considering that the instructors were volunteers, it is further possible that "only good teachers would 'risk' giving an outsider an opportunity to closely examine what their students were learning" (p. 753). There is some sloppy statistical analysis of the results when broken down by high school vs. college, gender, and calculus- vs. algebra-based. Each of these comparisons is made singly rather than in a three-way ANOVA. This

makes the findings (males > females, calculus > algebra, high school not different from college) difficult to accept.

4.5 Force and Motion Conceptual Evaluation (FMCE)

The Force and Motion Conceptual Evaluation (FMCE) "was designed to probe conceptual understanding of Newtonian mechanics" (Thornton and Sokoloff, 1998). This clearly sounds similar to the FCI and browsing the items suggests the same, but only a passing reference is made to the FCI. The article focuses on the pedagogical effects of two laboratory curricula compared to traditional coursework. The pre-test results are universally poor, with typically less than 20% of students giving correct Newtonian responses before instruction, across several semesters at two universities in both calculus- and non-calculus-based physics courses. After completing one of two laboratory curricula, the percent correct increased to over 70% in most cases and some over 90% when coupled with an additional interactive lecture series. Data from traditional teaching indicate gains around 20%, compared with the 50%+ gains from the interactive and laboratory methods. Additional testing at the end of the semester (other testing was conducted immediately after labs/lectures) suggests that students further assimilate knowledge even with no further instruction, showing small to moderate gains. No formal psychometric analysis of the FMCE is presented. A small case for content validity is asserted by mentioning that student answers "correlate well (above 90%)" with short-answer reasons.

4.6 Determining and Interpreting Resistive Electric Circuit Concepts Test (DIRECT)

The Determining and Interpreting Resistive Electric Circuit Concepts Test (DIRECT) was developed to assess students' conceptual understanding of direct current

(DC) circuits (Engelhardt and Beichner, 2004). The instrument was developed to increase the breadth of physics test coverage beyond existing instruments such as the FCI and TUG-K. Eleven topic areas for DIRECT were identified by reviewing textbooks and physics education literature, then consulting instructors to ensure no critical topics were overlooked. Items were then written to cover these areas and administered in open-ended format to generate authentic distracters.

Version 1.0 of DIRECT, with 29 items, was administered to 454 high school students and 681 university students across the country. The overall mean was 48% (52% university; 41% high school), with scores ranging from 14% to 97%. The KR-20 value is 0.71 and considered acceptable for group administration (i.e., ≥0.70). The average point-biserial correlation is 0.33, ranging from 0.07 to 0.51. The average discrimination index is only 0.26, below a desired value of 0.30, with values ranging from 0.00 to 0.43. The authors believe, "The low average discrimination values may indicate that the test is indeed uncovering students' misconceptions" (p. 102-3). The validity of this claim is tenuous at best. The percent correct ranged from 15% to 89%. This information was used to revise the instrument to Version 1.1, which proved to be more difficult (average 41% correct) but did not have appreciably different psychometrics (omitted due to similarity). A factor analysis was conducted using the "Little Jiffy method," identifying eight factors for 1.0 and eleven for 1.1; no specific results are reported, and the accuracy of the groupings is not discussed.

Student interviews deserve added attention. They were conducted in three parts: 1) identification of symbols in the test; 2) definitions of terms in the test; and 3) answering the items, providing reasons and stating their confidence. The interviewer had

access to each student's original answer, asking the student to recall his original reasoning if his answer changed from the original. The responses indicate that nearly all students understood the symbols, with only a light bulb symbol causing confusion. The symbols for voltage, current, and resistance (V, I, R) were often confused. Specific misconceptions were found to match global expectations of student misconceptions, but these varied widely from student to student.

Comparisons between and within groups were performed incorrectly, as they were with the TUG-K. With that in mind, the following differences are reported as statistically significant: overall mean, university > high school; mean, males > mean females; misconceptions university, males < females; interview confidence, males > females; textbook type, microscopic description of phenomena > traditional approach; instructional approach, hands-on > traditional (both algebra- and calculus-based). There were no significant differences based on the math basis for the course in either high school or university courses.

4.7 Geoscience Concept Inventory (GCI)

Libarkin and Anderson (2005) document the development and analysis of the Geoscience Concept Inventory (GCI), designed for entry-level college geoscience courses. The dissemination and subsequent analysis is novel compared to other concept inventories. Twenty-nine items were written; eleven formed a core, while two alternate versions each contained nine of the remaining items. Each student therefore answered 20 questions. Data were collected from 43 courses at 32 institutions in 22 states. The pre-test was administered to 2500 students; the post-test to 1295 students, with 930 students completing both.

Rather than reporting item statistics (i.e., discriminatory index, difficulty), a Rasch model was fit and scores were placed on an adjusted scale of 0 to 100. The relationship between raw scores and scaled scores were nearly identical across the two versions. The 930 matched students showed a small effect size of 0.17 (pre-test 43±11, post-test 47±12). While the gain is statistically significant, further analysis questions its practical significance. The gain is primarily attributable to low-achieving students (pre-test < 40%; average gain +9), as opposed to intermediate scorers (pre-test 40% to 60%; gain +2) and high scorers (> 60%; no gain). The authors insightfully note that low pre-testers are likely to suffer from bad luck, with the gains then a *regression to the mean* phenomenon. Using regression, the authors determine that the low achievers do in fact score 4% higher than would be expected as a result of bad luck, "indicating that improvement is real" (p. 397).

When results are analyzed by course, only 8 of 30 exhibited significant gains, with the overall gain of just one question. Details are presented from three courses, although it is unclear how these were selected.

- A small course (n = 11 pre; n = 9 post; n = 8 matched) showed a slight decrease in scaled scores (47±13; 43±13). This instructor reported some use of "alternative methods" but is implied to rely on traditional lecture and laboratory.
- A mid-sized course (42; 38; 28) was more successful, with 57% of students showing a gain, although only one or two questions per student. The instructor utilized in-class discussion in addition to lecturing.

- A large course (190; 183; 135) showed the largest gain, but again it was small (+4) and had a lower starting value (38). The instructor reported 100% lecture.

Finally, a small section is offered on the "entrenchment of ideas" (referred to elsewhere as "persistent misconceptions"). These concepts are unchanged by instruction, but the overall low gains would suggest more items belong in this section than the five presented. For example, prior to instruction, 78% of students believe that Earth's age can be determined by "fossils, rock layers, or carbon" as opposed to uranium or lead content of rocks; this misconception is held by 72% of students after instruction.

5. Conclusion

This chapter serves as a listing of concept inventories in development and provides an idea of the level of analysis performed. The following chapter describes the methods and the first two years' results for the SCI. The documentation of other concept inventories should be considered comparatively, although this formal comparison is not presented until the end of the dissertation.

CHAPTER IV

Methodology and Results

A. Methodology

1. Scores

Summary statistics are reported for all students who took either the pre-test or post-test, rather than only students who took both. Typically, the students who only took either pre-test or post-test have similar scores to those who took both. The number of students who took both is also usually much larger, dampening any effect of those who only took either pre-test or post-test.

There are several abbreviations used to define the different classes who have participated in this research project. The classes may also be followed by a number to indicate that there are multiple sections (e.g., Math #2) or multiple universities (e.g., External #2). The following table gives this information for reference.

Table 1: Abbreviations Used to Identify Classes

Abbreviation	Explanation
Engr	Introductory statistics course in OU's College of Engineering
Math	Introductory statistics course in OU's Department of Mathematics
REU	Research Experience for Undergraduates, two separate summer research groups in OU's School of Industrial Engineering
DOE	Design of Experiments, an upper-division Industrial Engineering course at OU
Ext.	External universities outside of OU

The following table also lists all the classes who have taken the SCI and the semesters in which they participated. The instructors may or may not be the same for classes who are listed under multiple semesters. A number in parenthesis indicates

multiple sections. The introductory statistics classes (labeled "intro") are calculus-based except Communications and External #3.

Table 2: Classes by Semester

Course	Level	Fa 02	Su 03	Fa 03	Sp 04
Communications	Intro	√			
Engr	Intro	√	√	√	√
Math	Intro	√ (2)	√ (2)	√ (2)	√ (2)
DOE	Advanced	√		√	
REU	Varies		√		
External #1	Intro		√	√	
External #2	Intro			√ (2)	
External #3	Intro, 2-yr			√	

2. Validity

The researchers focused on content, concurrent, and construct validity because they are broad and are described in most psychometric textbooks. Predictive validity is also mentioned because it relates to students' pre-conceptions. The SCI's validity is measured in terms of its target audience, introductory statistics courses in engineering departments. Statistics courses from the Mathematics department, a two-year college, and an advanced Design of Experiments course are included, but their data are not given as much weight when evaluating the instrument.

2.1 Content Validity

As a starting point for item construction, the researchers searched textbooks and statistics journals and used personal experience to identify important concepts. The Advanced Placement (AP) Statistics course outline (College Board, 2003) was used as a guide to ensure breadth of coverage. Next, a survey was used to verify the appropriateness of topics and to fill gaps in coverage. The survey was sent to all faculty members in the College of Engineering at the University of Oklahoma during the Spring 2001 semester. The respondents were asked to rank the importance of statistics topics for

their curricular needs. The scale ranged from 1 (not at all important) to 4 (very important), along with the option of "No opinion" if the topic was unfamiliar. Respondents were instructed to list additional items if they felt something was missing. The responses indicate that no major topic was omitted from the original list. Simultaneously, a literature search was conducted to identify misconceptions in statistics. Both journal articles (Garfield and Ahlgren, 1988; Kahneman and Tversky, 1972; Konold, 1995; Konold, *et al.,* 1993; Pollatsek, *et al.,* 1981) and textbooks (Montgomery and Runger, 1994; Moore, 1997) were utilized.

The Advanced Placement (AP) Statistics Course Description lists four general topic areas, listed below. Each of these areas is broken down further into approximately 20 sub-topics. These are generally similar to the faculty survey sub-topics and are therefore not analyzed.

I. Exploring Data: Observing patterns and departures from patterns

II. Planning a Study: Deciding what and how to measure

III. Anticipating Patterns: Producing models using probability theory and simulation

IV. Statistical Inference: Confirming models

Each item is placed into its appropriate faculty survey sub-topic and AP area, as nearly as possible. The faculty survey list is divided into nine topic areas, and each of these contains between four and ten sub-topics. It is not always possible to fit a question perfectly into a sub-topic, but all questions have an appropriate topic area. For example, there is no faculty survey sub-topic regarding the interpretation of p-value, but this does

fit into the general area Confidence Intervals and Hypothesis Testing. The categorization of items is presented to demonstrate the overall content validity of the SCI.

The content validity of individual items is also essential. After the first round of test administration, items were analyzed in several ways. First, answer distributions were examined to find choices which were consistently not chosen. These options were either thrown out or revised. Second, focus groups were conducted to gain insight from students as to why they chose certain answers and identify other choices to serve as distracters. Items were revised based on these results. Several new items were constructed in a similar manner as described above.

This revised SCI was administered in Summer 2003. A similar revision process was conducted as described above, and new items were constructed where necessary. In addition, specific effort was made to identify poorly written questions. It is necessary to identify poorly written questions because the SCI is not intended to identify good test-takers. Each question was evaluated on the basis of seven criteria identified by Gibb (1964) that may lead students with good test-taking skills to figure out the answer. The criteria are listed below:

1. Phrase-Repeat: Correct answer contains a key sound, word, or phrase that is contained in the question's stem.
2. Absurd Relationship: Distracters are unrelated to the stem.
3. Categorical Exclusive: Distracters contain words such as "all" or "every."
4. Precise: Correct answer is more precise, clear, or qualified than the distracters.
5. Length: Correct answer is longer than the distracters.

6. Grammar: Distracters do not match the verb tense of the stem, or there is not a match between articles (“a”, “an”, “the”).
7. Give-Away: Correct answer is given away by another item in the test.

(It is also sometimes said that students should “always choose C,” but this criteria implies that one should never use C as the correct answer, which is unreasonable.)

These problems were revealed both through focus groups and the researchers’ analysis. Information on the application of these criteria is provided in section 5 of Results.

2.2 Concurrent Validity

Concurrent validity is assessed by correlating a test with some “other test.” The “other test” is the overall course percentage grade, which is correlated with the SCI Pre-Test, SCI Post-Test, SCI Gain (Post minus Pre), and SCI Normalized Gain (Gain as a percentage of the maximum possible Gain). Due to various administrative difficulties, this is not possible for every course. The p-value listed is from a 2-tailed test. If a 1-tail test is desired, the reported p-value can be divided by two.

Based on the results of full-information Maximum-Likelihood Factor Analysis, the instrument is also divided into four sub-tests (see Construct Validity section, 2.4). The concurrent validity of the sub-tests is assessed by correlating the overall course grade with the score on each of the sub-tests.

2.3 Predictive Validity

The correlation between SCI Pre-Test and overall course grade is taken as a measure of predictive validity. The correlation between pre-test sub-scores and overall course grade is also included to examine the predictive validity of the sub-tests.

2.4 Construct Validity

The SCI as a whole can be said to measure the construct "statistics knowledge" or, more precisely, "conceptual understanding of statistics." As statistics is composed of many sub-topics, it is the goal of this analysis to determine those more precise constructs.

The SCI is divided into the sub-topics probability, descriptive statistics, and inferential statistics. Items which used graphical displays are also grouped separately because they have a strong common question format even though they are not highly-related topically. Items which are considered advanced (e.g., design of experiments or regression) are also grouped together because of the expectation that most introductory students will be guessing.

Using full-information Maximum-Likelihood (FIML) Factor Analysis, various models were fit, which are combinations of the sub-topics described above. The software package TestFact was used for the analysis. The data are from the combined post-test data for all classes from Fall 2003. It is recommended to have around 400 subjects for this type of analysis. The total number available is 332. The model is fit using the tetrachoric correlation matrix, as opposed to the traditional Pearson's r. The tetrachoric correlation accounts for the fact that items which vary in difficulty do not have a maximum possible correlation of 1.

A traditional factor analysis is also performed using the Varimax and Direct Oblimin rotations, as recommended by Klein (1993). The sample size of 332 is sufficient, compared to recommendations of 100 or 200 depending on the source. The subject to item ratio should be 10 to 1 by the most conservative suggestion, and this is met when the

advanced items are omitted. The goal of this analysis is to see if similar items load on the same factors.

3. Reliability

3.1 Reliability of the SCI

Coefficient alpha was calculated for each class who took the SCI. Alpha is calculated for both pre- and post-tests, and all students who took either test are included. It is more crucial that the post-test exhibit reliability (*cf.* impact of guessing, Chapter VI), but comparison between pre- and post-tests can be insightful.

3.2 Confidence Intervals and Hypothesis Tests on Alpha

The observed alpha values are tested against theoretical values of 0.60 and 0.80 because these have been cited as minimally-acceptable values in the literature (Oosterhof, 1996, and Nunnally, 1978, respectively). The alternate hypothesis is one-sided, and the directionality depends on whether the observed alpha is above or below 0.60 and 0.80. For example, an alpha of 0.65 will be tested against the alternate hypotheses $\alpha > 0.60$ and $\alpha < 0.80$.

3.3 Guttman Coefficients

The Guttman reliability estimates are calculated for the Fall 2003 SCI Post-Test. The results are based on all students who took the post-test. MatLab code was written to perform the calculations.

4. Discrimination

Ferguson's delta was calculated for each class separately to determine the overall discriminatory power of the instrument, rather than combining all the data from each semester into one large dataset.

For items, the discriminatory index is based on the bottom and top 25% of students with all ties included, rather than the recommended 27% because it is simpler to compute the quartiles. The inclusion of ties usually means that more than 27% are included. The discriminatory indices are considered for the instrument as a whole (e.g., how many items are "poor") and for individual items when making revision decisions. The criteria for poor, moderate, and high discriminatory indices are those recommended by Ebel (1954).

The point-biserial values were calculated for the Spring 2004 data to serve as a comparison between the discriminatory index and point-biserial value.

5. Item Analysis

Each item's reliability is measured by alpha-if-deleted. A question which contributes favorably to reliability will have an alpha-if-deleted below the overall alpha because deleting that one item would lower the overall alpha. Except for small classes, an item will rarely vary from the overall alpha by more than ±0.02. The alpha-if-deleted is interpreted both by the sign of the difference from the overall alpha and by ranking the questions according to alpha-if-deleted. The value of alpha-if-deleted is presented in tables for each question, but the values discussed in the text are the difference from the overall alpha. The reader can make the distinction because the difference is always preceded by a + or – sign which tells whether the question tends to raise or lower the overall alpha.

Discrimination is measured by the discriminatory index. The point-biserial value is not explicitly analyzed due to its similarity, except the aforementioned Spring 2004 comparison.

Content is validated to ensure that the item is not answered via testing-taking tricks. This knowledge is gathered from focus groups primarily but also based on the research group's knowledge of how they would approach the question. Construct validity of items is analyzed through the factor analytic methods and presented in that section.

For the pilot study (Fall 2002), statistics were calculated for the entire population of six statistics courses (Engr, two Math, Communications, Regression, Design of Experiments) to give a general idea of how the question behaved. For later work, the data were divided by class to examine differences across courses. The combined data are also reported, and they are used in situations where the classes do not have a pattern.

B. Results

1. Scores

The average score for the Fall 2002 courses is 36.9% (standard deviation 12.1%); this test was conducted near the end of the semester but is not strictly a post-test. When the Communications class is removed, the average is 39.9% (standard deviation 12.4%). Communications is the only course from Fall 2002 which is not calculus-based and also had poor testing conditions. The scores for later semesters are presented below and discussed thereafter.

Table 3: Summary Statistics for SCI, Summer 2003

Course	Level	N Pre	N Post	Mean Pre	Mean Post	Gain	S.D. Pre	S.D. Post
Engr	Intro	25	24	35.5%	48.7%	+15.2%	13.4%	16.9%
Math #1	Intro	17	14	35.7%	52.4%	+16.7%	12.8%	18.7%
Math #2	Intro	28	n/a	37.2%	n/a	n/a	13.5%	n/a
REU	Varies	27	n/a	53.0%	n/a	n/a	11.7%	n/a
Ext. #1	Intro	n/a	38	n/a	47.8%	n/a	n/a	11.7%

Table 4: Summary Statistics for SCI, Fall 2003

Course	Level	N Pre	N Post	Mean Pre	Mean Post	Gain	S.D. Pre	S.D. Post
Engr	Intro	70	53	42.2%	45.1%	+1.9%	13.1%	15.7%
Math #1	Intro	29	n/a	43.4%	n/a	n/a	13.6%	n/a
Math #2	Intro	19	19	49.7%	49.1%	-0.6%	12.8%	13.9%
Ext. #1	Intro	42	43	43.1%	49.5%	+6.4%	12.6%	15.0%
Ext. #2a	Intro	60	54	48.1%	52.4%	+5.3%	11.0%	11.8%
Ext. #2b	Intro	58	48	46.1%	49.9%	+3.8%	12.0%	10.8%
Ext. #3	2-yr	42	37	29.0%	31.6%	+2.6%	8.1%	10.1%
DOE	Adv.	34	26	33.7%	36.4%	+2.7%	11.4%	10.4%

Table 5: Summary Statistics for SCI, Spring 2004

Course	Level	N Pre	N Post	Mean Pre	Mean Post	Gain	S.D. Pre	S.D. Post
Engr	Intro	67	31	40.3%	48.0%	+7.7%	13.1%	12.3%
Math #1	Intro	38	31	45.6%	53.1%	+7.5%	12.7%	15.3%
Math #2	Intro	35	30	43.1%	47.9%	+5.8%	13.3%	13.7%

Table 6: Summary Statistics for SCI, Summer 2004

Course	Level	N Pre	N Post	Mean Pre	Mean Post	Gain	S.D. Pre	S.D. Post
REU	Varies	28	n/a	50.0%	n/a	n/a	12.0%	n/a

Table 7: Summary Statistics for SCI, Spring 2005

Course	Level	N Pre	N Post	Mean Pre	Mean Post	Gain	S.D. Pre	S.D. Post
Engr	Intro	62	8	40.5%	50.7%	+10.2%	11.3%	9.1%
Math #1	Intro	31	28	48.6%	49.0%	+0.4%	13.4%	13.4%
Math #2	Intro	31	31	48.2%	49.9%	+1.7%	13.7%	14.0%
Ext. #1	Intro	n/a	41	n/a	51.4%	n/a	n/a	12.0%

Table 8: Summary Statistics for SCI, Spring 2005

Course	Level	N Pre	N Post	Mean Pre	Mean Post	Gain	S.D. Pre	S.D. Post
Engr	Intro	24	17	40.7%	44.9%	+4.2%	12.2%	14.6%
Math #1	Intro	22	25	46.8%	44.0%	-2.8%	11.6%	14.4%
Math #2	Intro	21	15	48.5%	45.6%	-2.9%	14.8%	13.0%
Ext. #1	Intro	n/a	50	n/a	49.8%	n/a	n/a	13.8%
Quality	Adv.	35	n/a	44.9%	n/a	n/a	13.4%	n/a
Psych	Intro	106	106	38.6%	43.9%	+5.3%	9.8%	10.9%

Table 9: Summary Statistics for SCI, Summer 2005

Course	Level	N Pre	N Post	Mean Pre	Mean Post	Gain	S.D. Pre	S.D. Post
Engr	Intro	n/a	24	n/a	41.1%	n/a	n/a	13.7%
Math #1	Intro	n/a	24	n/a	47.0%	n/a	n/a	11.4%
Math #2	Intro	n/a	12	n/a	52.4%	n/a	n/a	12.1%
Psych	Intro	n/a	14	n/a	36.1%	n/a	n/a	9.3%

There is a noticeable increase in the average scores from the Fall 2002 pilot test to Summer 2003. For Fall 2002, the instrument was administered around the middle of November rather than during the final week of classes, as has been done for later semesters. This can explain part of the difference. Most of the questions have also been improved, including removing several choices which were not indicative of statistical reasoning but were highly-chosen nonetheless.

For Summer 2003 onward, the post-test scores are very similar across all classes with a mean around 50%. The pre-test scores from Summer 2003 are lower compared to Fall 2003. This results in essentially zero gains from pre-test to post-test for Fall 2003, whereas Summer 2003 had at least modest gains. The Spring 2004 pre-test scores are between the Summer and Fall 2003 results, but once again the post-test scores are right around 50%. The gains are 5% to 7% for Spring 2004, slightly larger than Fall 2003 but smaller than Summer 2003.

The post-test scores are similar to those on the FCI for high school and non-major physics courses, taught in the standard lecture format. The gains are smaller than on the FCI's early testing, which were usually around 10% to 15%. On the SCI, only the Summer 2003 Engr and Math courses had gains this large. All others were single digits, usually around 4%.

To demonstrate that the approach of including all students rather just those who took both pre-test and post-test is valid, the following table breaks the results into three groups for the Fall 2003 data: 1) those who took only the pre-test, 2) those who took only the post-test, and 3) those who took both.

Table 10: Grouping by Who Took Pre, Post, or Both, Fall 2003

Course	**Pre Only**	**Both, Pre Scores**	**Both, Post Scores**	**Post Only**
Engr	41.8% (n=23)	42.4% (n=47)	43.6% (n=47)	47.5% (n=6)
Math #2	52.4% (n=5)	48.7% (n=14)	51.5% (n=14)	42.4% (n=5)
External #2a	51.6% (n=9)	47.5% (n=51)	52.4% (n=51)	57.8% (n=3)
External #2b	42.9% (n=10)	46.8% (n=48)	49.9% (n=48)	(n=0)
External #3	25.4% (n=10)	30.4% (n=32)	32.9% (n=32)	23.5% (n=5)
DOE	32.9% (n=3)	33.1% (n=16)	39.0% (n=16)	32.4% (n=10)

Note: External #1 is not available because the pre-test was given anonymously.

There is no pattern of one group consistently out-performing another. For example, it is not apparent that low students drop the course and are therefore unavailable

to take the post-test. This fact, along with the smaller sample sizes of the pre-only and post-only groups, suggests that the approach is valid.

There are several interpretations of the SCI scores. One possibility is that the SCI is too hard for introductory students. However, the limited results from advanced courses are comparable, but the lower scores from a two-year college suggest that engineering majors possess more statistical knowledge. Another possibility is that students are not learning concepts in their coursework. This was a finding of the FCI, but more research is needed to draw this conclusion, such as in-class observations, interviews, and comparison with courses which use inter-active teaching. There is also a consideration that students do not take the SCI seriously since it typically is administered by someone not affiliated with the class and not for a grade. However, one course which counted the SCI as a small portion of the final exam scored almost identically to other classes (Fall 2003 External #1, mean 49.5%). Until more extensive research is available, no firm conclusions can be made.

The test was divided into four sub-topics – probability, descriptive statistics, inferential statistics, and graphical displays. The selection of these groups and the item categorization are discussed in the Construct Validity section (2.4). The sub-scores are shown in the tables below, along with the gain from pre-test to post-test (in parenthesis).

Table 11: Percent Correct (and Gains) for Sub-Topics, Fall 2003 Post-Test

Course	Probability	Descriptive	Inferential	Graphical
Engr	42.8% (+1.5%)	61.1% (-1.8%)	38.8% (+5.1%)	26.4% (+10.0%)
Math #2	51.1% (no change)	66.3% (+2.6%)	33.8% (+6.8%)	47.4% (+13.2%)
External #1	50.9% (+7.4%)	65.4% (+2.5%)	35.6% (+5.3%)	37.2% (+15.4%)
External #2a	52.0% (+6.2%)	75.9% (+1.7%)	42.9% (+8.9%)	30.6% (-2.7%)
External #2b	52.3% (+7.6%)	73.8% (+0.5%)	38.1% (+5.1%)	27.1% (+7.3%)
External #3	28.9% (-1.6%)	46.8% (+13.7%)	29.3% (-1.7%)	2.7% (-15.0%)
DOE	31.2% (no change)	53.8% (+1.4%)	27.5% (+5.4%)	28.8% (+8.2%)

Note: Math #1 is absent because that course did not take the post-test.

Table 12: Percent Correct (and Gains) for Sub-Topics, Spring 2004 Post-Test

Course	Probability	Descriptive	Inferential	Graphical
Engr	45.2% (+3.5%)	69.1% (+13.6%)	36.4% (+9.5%)	19.4% (+6.7%)
Math #1	49.2% (-0.5%)	62.9% (+5.0%)	25.9% (+15.5%)	22.4% (+8.3%)
Math #2	40.0% (+3.3%)	58.6% (+3.1%)	32.2% (+2.5%)	27.1% (+9.5%)

Table 13: Percent Correct (and Gains) for Sub-Topics, Fall 2004 Post-Test

Course	Probability	Descriptive	Inferential	Graphical
Engr	51.4% (+7.9%)	75.0% (+22.1%)	45.5% (+12.1%)	23.2% (-0.6%)
Math #1	48.0% (-1.1%)	63.2% (-2.1%)	40.3% (+9.2%)	43.9% (+12.3%)
Math #2	45.5% (+4.2%)	65.5% (+6.9%)	41.3% (+3.9%)	46.5% (+17.0%)

Note: External #1 did not take the pre-test; Engr is based on only 8 students who took the online post-test, whereas the Math courses had approximately 30 students each on pre- and post-test. There is no chart for Summer 2004 because that was only REU.

Table 14: Percent Correct (and Gains) for Sub-Topics, Spring 2005 Post-Test

Course	Probability	Descriptive	Inferential	Graphical
Engr	36.5% (-1.0%)	50.5% (+2.2%)	42.7% (+2.9%)	37.1% (+2.0%)
Math #1	36.0% (-12.0%)	54.5% (-4.6%)	41.5% (+3.9%)	38.3% (-3.9%)
Math #2	43.3% (-2.2%)	55.2% (-7.2%)	42.4% (+1.1%)	38.1% (-2.7%)

The descriptive sub-test has the highest scores for all classes on both pre-test and post-test. The probability sub-test is next-highest for most classes, followed by

inferential. The graphical sub-test is not very meaningful at this stage because it only has two questions. The graphical scores vary widely from class to class.

The four sub-tests have small gains, just as the overall instrument does. The only sub-test which attains double-digit gains consistently is the graphical section, but again this is not very meaningful due to the lack of questions.

The descriptive sub-test tends to have the lowest gains, most likely because it includes many questions which could be considered pre-knowledge. The students know this information upon entering and tend to maintain that level. The exception is the Fall 2003 External #3 (two-year college), which has a large gain in the descriptive sub-test but negative gains on the other three. Approximately 80% of that class indicated no prior statistics experience on the demographics survey, whereas most other classes are around 30% with no prior statistics experience. The Spring 2004 Engr course also has a large gain, but this is due more to a low pre-test. The results from all four sub-tests for this course are strikingly similar to the Fall 2003 Engr course.

The probability and inferential sub-tests have gains usually between +4% and +8%. One is not consistently higher than the other. Most instructors touch on probability briefly at the beginning of the course. Inferential statistics is normally taught during the last half of a course, but it is also considered conceptually difficult.

2. Validity

2.1 Content Validity

Table 15 (next page) shows the item categorization for SCI questions based on the faculty survey topics. The mean score of all topics is 2.62, and the median is 2.63. Table 16 (following Table 15) shows which rate above the median but contain no items.

Table 15: Item Categorization by Faculty Survey Topics

General Area	Specific Topic	Average	Fa 02	Su 03	Fa 03	Sp 04
Data Summary & Presentation	Measure of variability	3.68	√ (2)	√ (5)	√ (5)	√ (7)
Data Summary & Presentation	Importance of data summary	3.65	√ (2)	√ (2)	√ (2)	√ (2)
Cont. Rand. Vars. & Prob. Dist.	Normal dist.	3.48	√ (2)	√ (2)	√ (2)	√ (2)
Data Summary & Presentation	Methods of displaying data	3.43	√ (4)	√ (2)	√ (2)	√ (2)
Cont. Rand. Vars. & Prob. Dist.	Continuous uniform dist.	3.32	√	√	√	√
Probability	Interpretation of prob.	3.26	+	+	+	+
Discrete Prob. Distributions	Poisson dist.	3.14	√	√	√	√
Data Summary & Presentation	Frequency dist and histograms	3.09	+	+	+	+
Random Variables	Expected values	3.09	√ (2)	√	√	
Probability	Independence	3.00	√	√	√	√ (2)
Parameter Estimation	The central limit theorem	3.00	√	√	√	√
Parameter Estimation	Random sampling	2.95	√ (3)	√ (3)	√ (3)	√ (3)
Probability	Sample space and events	2.95	+	+	+	+
Cont. Rand. Vars. & Prob. Dist.	Standardized normal	2.87	+	+	+	+
Discrete Prob. Distributions	Binomial dist.	2.86	√	√	√	√
Probability	Conditional prob.	2.85	√	√	√	√
Probability	Mult. And total prob. Rules	2.81	+	+	+	+
Probability	Axiomatic rules	2.80	+	+	+	+
Probability	Counting concepts	2.77	√	√	√	√
Discrete Prob. Distributions	Discrete uniform dist.	2.76	+	+	+	+
Parameter Estimation	Sampling dist.	2.75	+	+	+	+
Conf. Intervals & Hypo. Testing	Inference on the mean of a pop.	2.74	√ (8)	√ (6)	√ (6)	√ (6)
Probability	Addition Rules	2.72				√
Linear Regression	Assessing the adequacy of reg.	2.71		√	√	√
Linear Regression	Hypothesis tests in reg.	2.67		√	√	√
Probability	Bayes' theorem	2.63	√	√	√	√
Data Summary & Presentation	Percentiles and quartiles	2.59	√	√	√	√
Cont. Rand. Vars. & Prob. Dist.	Exponential dist.	2.50			√	
Multi-factor designs	2 factor factorial design	2.24		√	√	√

Key: √ means there is one item in that category
√ (#) means there are multiple items, with # being how many
+ means an item (or items) is categorized elsewhere but has elements of that topic

Table 16: Important Topics Missing from the SCI

General Area	Specific Topic	Average	Comment
Other	Other	3.75	No topic was mentioned more than once
Linear Regression	Simple linear regression	3.52	Regression not usually taught in Intro
Joint prob. Distributions	Covariance and corr.	3.10	Correlation item added for Summer 04
Linear Regression	Properties of the least squares	3.10	Regression not usually taught in Intro
Data Summary & Presentation	Time sequence plot	3.00	Could be added as part of a graphical item
Linear Regression	Correlation	2.95	Correlation item added for Summer 04
Linear Regression	Use of the reg. for prediction	2.86	Regression not usually taught in Intro
Parameter Estimation	Properties of estimators	2.84	An under-lying theory of hypo. Testing but hard to explicitly ask
Linear Regression	Confidence intervals for the reg.	2.81	Regression not usually taught in Intro
Random Variables	Linear combinations	2.80	Taught briefly but usually computational
Conf. Intervals & Hypo. Testing	Testing for a goodness of fit	2.78	Advanced topic
Random Variables	Functions of random var.	2.76	Taught briefly but usually computational
Joint prob. Distributions	Two discrete random vars.	2.75	Plan to add joint prob. Item
Conf. Intervals & Hypo. Testing	Sample size determination	2.68	Advanced topic
Conf. Intervals & Hypo. Testing	Inference on the var. of a norm	2.63	Potential good topic to replace other hypo. Tests

The faculty survey can also be analyzed by looking only at the general area, rather than the specific topics. Table 17 shows the general areas, sorted by average score on importance, and the number of items in each category for each semester. Items are only included in their primary category.

Table 17: Number of Items in Faculty Survey General Areas

General Area	Importance	Fa 02	Su 03	Fa 03	Sp 04
Data Summary & Presentation	2.90	6	8	8	10
Probability	2.88	4	4	4	5
Linear Regression	2.86	0	2	2	2
Random Variables	2.76	2	1	1	0
Joint prob. Distributions	2.72	0	0	0	0
Parameter Estimation	2.71	4	4	4	4
Discrete Prob. Distributions	2.67	2	2	2	2
Cont. Rand. Vars. & Prob. Dist.	2.67	6	4	5	4
Conf. Intervals & Hypo. Testing	2.49	8	6	6	6
Time Series, etc.	2.42	0	0	0	0
Single factor experiments	2.30	0	0	0	0
Multi-factor designs	2.09	0	2	2	2

Table 18 shows the number of items which fall into each of the AP Statistics (College Board, 2003) general topic areas.

Table 18: Number of Items in AP Statistics Areas

AP Area	Fa 02	Su 03	Fa 03	Sp 04
I: Exploring Data	6	8	8	10
II: Planning a Study	3	3	3	3
III: Anticipating Patterns	13	14	15	15
IV: Statistical Inference	10	8	8	7

Based on the faculty survey, the SCI makes a strong case for content validity. There have been only three items which ranked below the median importance, and two of these have been deleted. Most of the important topics which are missing relate to regression, which is not usually taught in the introductory engineering statistics course at OU. If surveys of other universities show that regression is commonly taught, items should be devised to cover the most important areas of regression. Other topics, such as joint probability and correlation, are expected to be added to the SCI for future versions. Still others are likely too advanced for an introductory class and therefore do not conform to the target audience. The only item topic which does not directly conform to the faculty survey is the *t*-distribution, which is used in hypothesis tests of small samples.

Considering the general areas of the faculty survey (Table 17), there seem to be too many items in the categories Continuous Random Variables & Probability Distributions and Confidence Intervals & Hypothesis Testing. Most of the items in the former relate to probability distributions, and Probability is the second-most important area. All of the items in the latter relate to inferences about a mean, and this sub-topic scores 2.74, which is above the median. Further, browsing an introductory statistics textbook will reveal that inferences on the mean are a primary topic covered. Therefore, the excess of items in these two areas is not as negative as appears at first glance.

The Linear Regression area again shows as lacking, as previously noted. The Random Variables area has two sub-topics, functions of a random variable and linear combinations, which might be good candidates for new items. The Joint Probability area has already been cited for its deficiencies, and new items are being devised. Time Series could be incorporated as part of a graphical question, while the Single Factor Experiments and Multi-Factor Experiments are advanced.

Based on the AP categorization (Table 18), Area II (Planning a Study) appears lacking, while Area III (Anticipating Patterns) may have too many items. Area II has the fewest number of sub-topics. In addition, most of these concepts are implied by several of the hypothesis test items. Therefore, Area II is not as lacking as it appears. Area III contains all of the probability items, around 10 on the various versions of the SCI. There is discussion about reducing the number of probability items because statistics and probability are somewhat separate, although probability is usually covered in an introductory statistics course. If the number of probability items is reduced in favor of Area II, then the SCI can have nearly equal coverage across the four AP areas.

2.2 Concurrent Validity

The correlation of SCI scores with overall course grade is shown below for Summer 2003, based on students who took both pre-test and post-test.

Table 19a: Correlation of SCI Scores with Overall Course Grade(%), Summer 2003

Course	SCI Pre	SCI Post	SCI Gain	SCI Norm.Gain
Math #1 (n=12)	r = -0.392 (p = 0.207)	r = -0.023 (p = 0.944)	r = 0.318 (p = 0.314)	r = 0.288 (p = 0.365)
Engr (n=22)	r = 0.360 (p = 0.109)	r = 0.593** (p = 0.005)	r = 0.511* (p = 0.018)	r = 0.604** (p = 0.004)

For all Tables 19, 20, and 21: (2-tailed p-values in parenthesis)
** significant at 0.01, * significant at 0.05, † significant at 0.10 (0.05 if one-tailed)

For Fall 2003, grade data are available for seven courses. Five are introductory statistics courses, with four in engineering departments and the same Math course as above with a different professor. Three courses are at four-year universities outside OU, and one course is at a two-year college. Data from an advanced Design of Experiment course (DOE) are included as well. The data are presented in Table 19b. Spring 2004 data are in Table 19c.

Table 19b: Correlation of SCI Scores with Overall Course Grade(%), Fall 2003

Course	SCI Pre	SCI Post	SCI Gain	SCI Norm.Gain
Engr (n=47)	r = 0.360* (p = 0.012)	r = 0.406** (p = 0.005)	r = 0.114 (p = 0.444)	r = 0.139 (p = 0.352)
Math #2 (n=14)	r = -0.066 (p = 0.823)	r = -0.054 (p = 0.854)	r = 0.011 (p = 0.970)	r = 0.200 (p = 0.492)
External #1 (n=43)	n/a	r = 0.343* (p = 0.024)	n/a	n/a
External #2a (n=51)	r = 0.224 (p = 0.113)	r = 0.296* (p = 0.035)	r = 0.094 (p = 0.514)	r = 0.052 (p = 0.716)
External #2b (n=48)	r = 0.400** (p = 0.005)	r = 0.425** (p = 0.003)	r = -0.034 (p = 0.818)	r = -0.041 (p = 0.780)
External #3 (n=31)	r = 0.206 (p = 0.266)	r = 0.438* (p = 0.014)	r = 0.305 (p = 0.095)	r = 0.348† (p = 0.055)
DOE (n=16)	r = 0.148 (p = 0.585)	r = 0.085 (p = 0.754)	r = -0.098 (p = 0.718)	r = -0.157 (p = 0.561)

Note: Math #1 for Fall is not listed because the Post-Test was not given due to a scheduling conflict. Gains are not given for External #1 because the Pre-Test was given anonymously.

Table 19c: Correlation of SCI Scores with Overall Course Grade(%), Spring 2004

Course	SCI Pre	SCI Post	SCI Gain	SCI Norm.Gain
Engr (n=29)	r = 0.060 (p = 0.758)	r = 0.133 (p = 0.493)	r = 0.080 (p = 0.679)	r = 0.108 (p = 0.578)
Math #1 (n=30)	r = 0.323† (p = 0.081)	r = 0.502** (p = 0.005)	r = 0.316† (p = 0.089)	r = 0.353† (p = 0.056)
Math #2 (n=26)	r = 0.219 (p = 0.282)	r = 0.384† (p = 0.053)	r = 0.303 (p = 0.133)	r = 0.336† (p = 0.094)

The concurrent validity of the sub-tests is measured by correlating the post-test scores with the overall course grade. The correlation of sub-test pre-scores with overall course grade is presented later as a measure of predictive validity.

Table 20a: Correlation of SCI Sub-Scores with Overall Course Grade(%), Fall 2003 Post-Tests

Course	Probability	Descriptive	Inferential	Graphical
Engr (n=47)	r = 0.256† (p = 0.082)	r = 0.302* (p = 0.039)	r = 0.437** (p = 0.002)	r = 0.183 (p = 0.218)
Math #2 (n=14)	r = -0.091 (p = 0.757)	r = -0.298 (p = 0.301)	r = 0.256 (p = 0.377)	r = 0.009 (p = 0.976)
External #1 (n=43)	r = 0.321* (p = 0.036)	r = 0.323* (p = 0.035)	r = 0.158 (p = 0.313)	r = 0.003 (p = 0.986)
External #2a (n=51)	r = 0.235† (p = 0.097)	r = 0.138 (p = 0.336)	r = 0.329* (p = 0.018)	r = -0.024 (p = 0.868)
External #2b (n=48)	r = 0.371** (p = 0.009)	r = 0.079 (p = 0.592)	r = 0.374** (p = 0.009)	r = 0.202 (p = 0.169)
External #3 (n=31)	r = 0.020 (p = 0.917)	r = 0.463** (p = 0.009)	r = 0.223 (p = 0.227)	r = 0.129 (p = 0.490)
DOE (n=16)	r = -0.142 (p = 0.599)	r = -0.026 (p = 0.925)	r = 0.497* (p = 0.050)	r = 0.253 (p = 0.345)

Table 20b: Correlation of SCI Sub-Scores with Overall Course Grade(%), Spring 2004 Post-Tests

Course	Probability	Descriptive	Inferential	Graphical
Engr (n=29)	r = -0.010 (p = 0.961)	r = 0.058 (p = 0.766)	r = 0.279 (p = 0.143)	r = 0.150 (p = 0.438)
Math #1 (n=30)	r = 0.326† (p = 0.079)	r = 0.270 (p = 0.150)	r = 0.637** (p < 0.001)	r = 0.080 (p = 0.674)
Math #2 (n =26)	r = 0.584** (p = 0.002)	r = -0.031 (p = 0.882)	r = 0.485* (p = 0.012)	r = 0.290 (p = 0.150)

The Summer 2003 results (Table 19a) indicate that the SCI attains concurrent validity for the Engr class but not for the Math class. In fact, the Math course has

negative correlations on the pre-test and post-test. When gains are considered, the Math course yields moderate correlations, but they are not significant. While this could be random, it is encouraging that the correlations are positive. It should also be noted that this class has a small sample size. The Engr course yields a moderate (but not significant) correlation on the Pre-Test and significant positive correlations on the three other measures. The correlation is strongest with Normalized Gain.

The failure of the Math course to attain concurrent validity is disappointing, though perhaps not surprising. The SCI was constructed to serve as an assessment instrument for Engineering Statistics courses. In general, mathematics courses are taught from a more theoretical perspective, while engineering courses are typically more applied. The use of a different textbook and different topic coverage may also contribute. If the results were reversed (low correlations for Engr), this would be a concern. Future work will aim to improve the SCI so that it is valid across all statistics courses (e.g., Mathematics, Engineering, Psychology).

From Fall 2003 (Table 19b), the SCI consistently has significant positive correlations on the Post-Test except the Math and DOE courses, which have near-zero correlations on all measures. However, unlike Summer 2003, the Gain and Normalized Gain do not provide significant correlations except for the normalized gain on External #3, which is significant if one considers the 1-tail *p*-value. In fact, most of the other courses' correlations are near zero (between -0.20 and +0.20). These results imply that the SCI Post-Test measures the same basic material as the introductory courses cover, which is the most important evidence for concurrent validity.

From Spring 2004 (Table 19c), the Engr course fails to attain concurrent validity. The course was taught by a different professor than previously, and he also used a different textbook. This result serves as a caution about extrapolating the validity of one professor to another, even though they are in the same department. More data are needed from the Spring 2004 professor to determine the merit of these statements. The Spring 2004 Math courses display concurrent validity. In fact, the correlation of course grade with SCI Post-Test for Math #1 is the highest correlation to date ($r = 0.502$). The other three correlations with course grade are significant at 0.05 by a one-tailed test. The Math #2 course grades correlate significantly with the Post-Test and Normalized Gain.

From Fall 2003 (Table 20a), the probability and inferential sub-tests have the strongest evidence for concurrent validity. All four introductory engineering statistics courses have significant positive correlations by a 1-tailed test (two by a 2-tailed test). Three of the four yield significant positive correlations on the inferential sub-test, including two at $p < 0.01$. The DOE course is also significant on the inferential sub-test. The descriptive sub-test has significant positive correlations for the Engr, External #1, and External #3 courses. The graphical sub-test has no significant correlations.

By course, Engr has the strongest concurrent validity with three of the four sub-tests correlating significantly. The other introductory engineering statistics courses (External #1, #2a, #2b) correlate significantly on two of the four sub-tests, while External #3 and DOE only on one. The Math course has no significant correlations, with the probability and descriptive sub-tests even correlating negatively with the grade.

From Spring 2004 (Table 20b), the Engr course fails to attain concurrent validity for any sub-tests, just as it failed for the overall instrument. Both Math courses obtain

significant positive correlations for the probability and inferential sub-tests. Each course has one correlation above 0.50, which is higher than any correlation for Fall 2003. This Math professor had not previously administered a post-test with grades available. The prior Math courses with no significant correlations were taught by different professors.

2.3 Predictive Validity

The correlation of SCI Pre-Test with Overall Course Grade is considered a measure of predictive validity. This is found in the first column of Tables 19a and 19b. The correlation of pre-test sub-scores with the overall final course grade is calculated as a measure of predictive validity of the sub-tests (Tables 21a and 21b).

Table 21a: Correlation of SCI Sub-Scores with Overall Course Grade(%), Fall 2003 Pre-Tests

Course	**Probability**	**Descriptive**	**Inferential**	**Graphical**
Engr (n=47)	r = 0.245† (p = 0.097)	r = 0.304* (p = 0.038)	r = 0.069 (p = 0.643)	r = -0.047 (p = 0.755)
Math #2 (n=14)	r = 0.208 (p = 0.476)	r = -0.149 (p = 0.611)	r = -0.275 (p = 0.341)	r = 0.229 (p = 0.430)
External #2a (n=51)	r = 0.085 (p = 0.551)	r = 0.152 (p = 0.286)	r = 0.266† (p = 0.059)	r = 0.223 (p = 0.115)
External #2b (n=48)	r = 0.343* (p = 0.018)	r = 0.418** (p = 0.003)	r = 0.107 (p = 0.474)	r = 0.479** (p = 0.001)
External #3 (n=31)	r = 0.026 (p = 0.888)	r = 0.460** (p = 0.009)	r = 0.033 (p = 0.859)	r = -0.303† (p = 0.097)
DOE (n=16)	r = -0.095 (p = 0.725)	r = 0.045 (p = 0.868)	r = 0.400 (p = 0.124)	r = 0.033 (p = 0.904)

Table 21b: Correlation of SCI Sub-Scores with Overall Course Grade(%), Spring 2004 Pre-Tests

Course	**Probability**	**Descriptive**	**Inferential**	**Graphical**
Engr (n=29)	r = 0.092 (p = 0.636)	r = 0.053 (p = 0.785)	r = 0.010 (p = 0.957)	r = -0.026 (p = 0.893)
Math #1 (n=30)	r = 0.145 (p = 0.443)	r = 0.339† (p = 0.067)	r = 0.213 (p = 0.258)	r = 0.070 (p = 0.715)
Math #2 (n =26)	r = 0.232 (p = 0.255)	r = 0.290 (r = 0.150)	r = 0.036 (p = 0.860)	r = 0.158 (p = 0.441)

For Summer 2003 (Table 19a), the SCI Pre-Test lacks predictive validity with respect to final course grade. This implies that pre-knowledge is not playing a significant role in how much a student learns in the course. The conclusion is similar to that reached in the pilot study (Stone, *et al.,* 2003), which stated that statistics experience does not play a vital role in the SCI score for students in an introductory statistics course. However, data from Fall 2003 (Table 19b) yield significant correlations of SCI Pre-Test with overall course grade for two courses. The magnitudes of the correlations are similar to the Summer 2003 Engr course, but the larger sample sizes make the Fall 2003 correlations significant.

The magnitude of predictive validity for the sub-tests is comparable to that of the overall test in that there are few significant correlations. External #2b has the strongest predictive validity, with three of the four sub-tests correlating significantly. Engr and External #3 have two based on the 1-tailed p-values. External #2a has one, while Math and DOE have none. The descriptive sub-test has three courses with significant correlations (two at 0.01). Based on 1-tailed p-values, the inferential and probability sub-tests have one significant correlation, while the graphical sub-test has only one. The conclusion is that the SCI generally lacks predictive validity, with the possible exception of the descriptive sub-test, which is largely pre-knowledge. The total instrument and the sub-tests also lack predictive validity for the Spring 2004 Engr and Math #2 courses. There is a moderate argument for predictive validity with Math #1 because the correlation is significant at 0.05 by a one-tailed test ($r = 0.326$, $p = 0.079$). Similarly, the descriptive sub-test has a correlation of $r = 0.339$ ($p = 0.067$).

2.4 Construct Validity

The construct being assessed by the SCI might be called "conceptual knowledge of statistics," although naming a construct is a precarious occupation. The first model uses full-information Maximum-Likelihood (FIML) Factor Analysis with the groupings probability (group 1), descriptive statistics (group 2), and inferential statistics (group 3). Questions deleted based on content validity are not included in the analysis. The results from TestFact are shown in Table 22 (next page).

Questions 21 and 28 share the common feature of graphical displays. Question 21 has the most negative specific factor (-0.5349) of any item, and question 28 has a specific factor close to zero (0.0972). A new group (4) was created for these two items. Although two items is a small group, more items are planned using graphical displays and the group will be larger at that time. The results of the four-factor model are in Table 23 (following Table 22). This model has a slightly worse fit compared to the three-factor model based on the uniqueness (3-factor 70.1552% unique variance; 4-factor 70.4135%). However, the graphical questions now fit better on their own factor.

Table 22: Results for 3-Factor FIML Model

GENERAL	0	20.3208
ITEM GROUP	1	2.6994
ITEM GROUP	2	3.3060
ITEM GROUP	3	3.5187
UNIQUENESS		**70.1552**

ITEM	GROUP	DIFFICULTY	COMMUNALITY	GENERAL	SPECIFIC
5 SSCI6	1	-1.4389	0.2109	0.3834	-0.2528
6 SSCI7	1	0.2233	0.2150	0.4504	0.1103
8 SSCI10	1	0.0741	0.2729	0.5205	0.0450
11 SSCI13	1	0.5272	0.2631	0.3203	0.4006
17 SSCI20	1	-0.4636	0.3866	0.5500	0.2900
20 SSCI23	1	0.4255	0.4099	0.6397	-0.0263
21 SSCI24	1	0.4121	0.5123	0.6686	0.2555
23 SSCI27	1	0.1252	0.3340	0.3569	-0.4545
25 SSCI29	1	1.1683	0.3321	0.3837	-0.4300
28 SSCI33	1	0.0332	0.1897	0.4325	-0.0519
2 SSCI2	2	-0.7993	0.2980	0.5075	0.2011
3 SSCI3	2	-0.4353	0.1025	0.2458	0.2050
4 SSCI4	2	0.3060	0.2770	0.4897	-0.1930
7 SSCI9	2	-0.8684	0.6320	0.6598	0.4435
9 SSCI11	2	-1.3893	0.2591	0.5065	0.0510
10 SSCI12	2	0.2903	0.0309	0.1348	-0.1128
12 SSCI14	2	0.4440	0.0874	0.2928	-0.0411
14 SSCI16	2	-0.5895	0.5620	0.6505	0.3727
15 SSCI18	2	-0.6066	0.6327	0.6705	0.4279
18 SSCI21	2	0.8812	0.5290	0.4928	-0.5349
24 SSCI28	2	0.3565	0.1739	0.4055	0.0972
26 SSCI31	2	-0.6318	0.2483	0.4905	-0.0882
1 SSCI1	3	0.4531	0.0398	0.1253	0.1552
13 SSCI15	3	1.4384	0.7264	-0.2252	0.8220
16 SSCI19	3	0.3024	0.3537	0.4876	0.3406
19 SSCI22	3	-0.5761	0.0812	0.1188	0.2590
22 SSCI25	3	0.8540	0.1267	0.2285	0.2728
27 SSCI32	3	0.2063	0.0808	0.2361	0.1583
29 SSCI34	3	0.1926	0.2872	0.4991	0.1952

Table 23: Results for 4-Factor FIML Model

GENERAL	0	19.8414	
ITEM GROUP	1	2.7951	
ITEM GROUP	2	2.6438	
ITEM GROUP	3	3.4481	
ITEM GROUP	4	0.8581	
UNIQUENESS		**70.4135**	

	ITEM	GROUP	DIFFICULTY	COMMUNALITY	GENERAL	SPECIFIC
5	SSCI6	1	-1.4365	0.2228	0.3949	-0.2586
6	SSCI7	1	0.2262	0.2248	0.4588	0.1194
8	SSCI10	1	0.0766	0.2745	0.5208	0.0574
11	SSCI13	1	0.5288	0.2600	0.3065	0.4076
17	SSCI20	1	-0.4607	0.3897	0.5513	0.2928
20	SSCI23	1	0.4298	0.4208	0.6486	-0.0117
21	SSCI24	1	0.4156	0.5237	0.6714	0.2701
23	SSCI27	1	0.1271	0.3419	0.3684	-0.4540
25	SSCI29	1	1.1707	0.3608	0.4096	-0.4394
28	SSCI33	1	0.0355	0.1958	0.4402	-0.0449
2	SSCI2	2	-0.7957	0.2881	0.5332	0.0615
3	SSCI3	2	-0.4342	0.0878	0.2453	0.1662
4	SSCI4	2	0.3064	0.2432	0.4808	-0.1095
7	SSCI9	2	-0.8624	0.6195	0.6267	0.4761
9	SSCI11	2	-1.3868	0.2579	0.4977	0.1009
10	SSCI12	2	0.2913	0.0191	0.1312	-0.0431
12	SSCI14	2	0.4452	0.0803	0.2826	0.0216
14	SSCI16	2	-0.5831	0.5137	0.6416	0.3196
15	SSCI18	2	-0.6006	0.7712	0.6256	0.6163
26	SSCI31	2	-0.6292	0.2225	0.4694	0.0465
1	SSCI1	3	0.4539	0.0396	0.1159	0.1616
13	SSCI15	3	1.4374	0.7061	-0.2021	0.8156
16	SSCI19	3	0.3050	0.3499	0.4945	0.3247
19	SSCI22	3	-0.5755	0.0847	0.1163	0.2669
22	SSCI25	3	0.8553	0.1319	0.2329	0.2786
27	SSCI32	3	0.2076	0.0826	0.2469	0.1471
29	SSCI34	3	0.1955	0.2813	0.4986	0.1808
18	SSCI21	4	0.8837	0.2091	0.4161	0.1897
24	SSCI28	4	0.3579	0.3767	0.4048	0.4613

The results from the more traditional factor analysis are now presented. The extraction method is principal components analysis. The first decision is how many factors to include. Based on the criteria of eigenvalues greater than one, there are 11

factors. Using a more subjective Scree plot, five factors appear most logical. The Scree plot is shown below, and it should be noted that the curve levels off after the fifth factor.

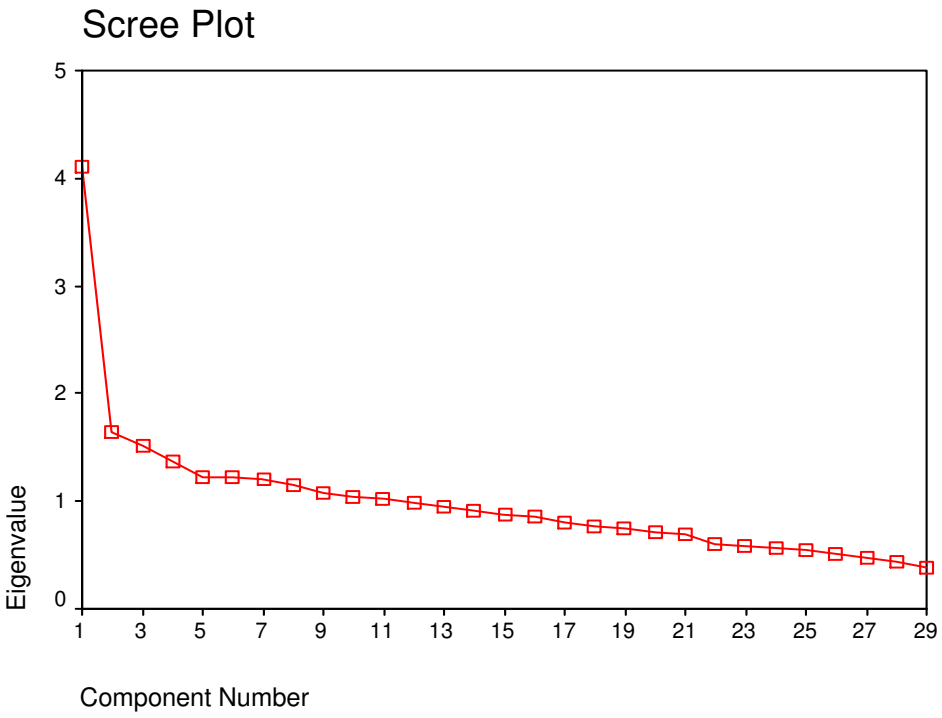


Figure 1: Scree Plot for Principal Component Analysis, from SPSS ™

The five-factor model explains 35.0% of the variance, which is slightly more than the FIML model. The eleven-factor model explains 57.1%. The first factor explains 15.2% for either model, which is smaller than the approximate 20% on the FIML models' general factor.

The rotation method is the other crucial decision. Tables 24 and 25 show which factor each item loads highest on for the unrotated, Varimax, and Direct Oblimin solutions. Both the five-factor and eleven-factor results are included. The items are sorted further by their categorization, as defined in the FIML factor analysis (Prob=Probability, Desc=Descriptive, Infer=Inferential, Graph=Graphical).

Table 24: Item Groupings for three Factor Analytic Solutions, 11 Factors

Factor	Unrotated	Varimax	Direct Oblimin
1	Prob: 7,10,20,23,24,33 Desc: 2,3,9,16,18 Infer: 19,34 Graph: 28	Prob: 33 Desc: 2,3,9,16,18 Graph: 28	Prob: 10 Desc: 3,16,18 Infer: 19,34
2	Prob: 13 Infer: 32 Graph: 21	Prob: 7,23,24 Graph: 21	Prob: 7,23,24 Desc: 4 Graph: 21
3	None	Prob: 10 Infer: 34	Prob: 33 Infer: 32
4	Prob: 6,27 Desc: 31	Prob: 13,20 Desc: 14 Infer: 19	Prob: 6,20 Desc: 31 Graph: 28
5	Infer: 15,25	Prob: 6 Desc: 11,31	Infer: 15
6	Infer: 1	Prob: 27	Infer: 1,25
7	None	Infer: 32	Prob: 13
8	Prob: 29 Desc: 12,14	Infer: 1,25	Prob: 27,29 Desc: 12
9	None	Infer: 15	None
10	Desc: 4 Infer: 22	Prob: 29 Desc: 12	Desc: 2,9 Infer: 22
11	Desc: 11	Desc: 4 Infer: 22	Desc: 11,14

Table 25: Item Groupings for three Factor Analytic Solutions, 5 Factors

Factor	Unrotated	Varimax	Direct Oblimin
1	Prob: 7,10,20,23,24,33 Desc: 2,3,4,9,14,16,18 Infer: 19,34 Graph: 28	Prob: 29,33 Desc: 2,3,9,16,18 Infer: 1,22 Graph: 28	Prob: 10,29 Desc: 3,11,16,18 Infer: 19,34
2	Prob: 13 Desc: 12 Infer: 1,32 Graph: 21	Prob: 7,23,24 Desc: 4,12 Graph: 21	Prob: 7,13,23,24 Desc: 4,12 Infer: 1 Graph: 21
3	Prob: 29	Prob: 10 Infer: 15,34	Prob: 33 Desc: 2,14 Infer: 32
4	Prob: 6,27 Desc: 11,31	Prob: 13,20 Desc: 14 Infer: 19,25,32	Prob: 6,20,27 Desc: 9,31 Graph: 28
5	Infer: 15,22,25	Prob: 6,27 Desc: 11,31	Infer: 15,22,25

The 11-factor model is easier to interpret because it has more factors which contain primarily items from one of the four areas. The five-factor model is too jumbled to easily interpret. From inspection, it is apparent that some items group together on at least two of the rotation methods. Aside from Factor 1 (the "general" factor), the factor numbering is not especially important. Table 26 lists some generalities drawn from the 11-factor models.

Table 26: Possible Constructs based on 11-Factor Models

Grouped Items	Similarities	Construct
Prob: 10,33 Desc: 2,3,9,16,18 Infer: 19,34 Graph: 28	The desc. items include two about st. dev. and two about median	Most likely a general statistics factor since it is Factor 1 in the models
Prob: 7,23,24 Graph: 21	7 & 23 relate to weather, although different aspects	Probability, with perhaps weather as a sub-construct
Infer: 1,25	1 is use of t-distribution, 25 is meaning of p-value	Inferential, although not highly related to each other
Prob: 29 Desc: 12	None, and in fact 29 is very hard and 12 very easy	Not meaningful
Prob: 6 Desc: 31	Both are relatively easy; possibly the same people get both correct	Not meaningful

The SCI has a large portion of unique variance on the FIML model. This is not necessarily bad because the SCI has primarily unique items. Only the content area of hypothesis testing seems to have an abundance of items, but this topic is rather broad in itself and only a few items might be considered redundant. It is apparent from the faculty survey that statistics is a broad topic, and multiple items on a single topic would either make the SCI too long or fail to capture a satisfactory range of knowledge.

Based on the four-factor FIML model (Table 23), the inferential and graphical sub-tests have all of the items loading positively on the specific factor, which suggests they do in fact belong together. The descriptive sub-test is somewhat poorer in that two items load negatively on the specific factor, while three others are only slightly positive (< 0.10). The probability sub-test is the poorest fit, with half the items loading negatively. One solution is to try fitting a different model, such as breaking probability into two sub-tests. Another possibility is to reduce the number of probability items by deleting several of those that load negatively. A similar suggestion was made based on content validity because statistics and probability are sometimes considered separate disciplines.

The factor analytic conclusions are not very different from those of the FIML models. There appears to be somewhat of a general factor with around 10 items depending on the rotation method. Two other possible constructs relate to probability and inferential statistics, but these only included a portion of the items in those sub-topics. Two other constructs have no meaningful interpretation. These account for approximately 10 more items, meaning the remaining one-third of the SCI is unique in that the items do not consistently group on the different rotation methods.

3. Reliability

3.1 Coefficient Alpha

The values of coefficient alpha are presented in the Tables 27. For the Fall 2002 pilot study, the value for the combined data is 0.6115.

Table 27a: Coefficient Alpha, Summer 2003, Pre-Test and Post-Test

Course	Pre-Test Alpha	Post-Test Alpha
Engr	0.6805	0.8100
Math #1	0.6765	0.8587
Math #2	0.6902	--
REU	0.5983	--
External #1	--	0.5781

Table 27b: Coefficient Alpha, Fall 2003, Pre-Test and Post-Test

Course	Pre-Test Alpha	Post-Test Alpha
Engr	0.6863	0.7496
Math #1	0.7122	--
Math #2	0.6715	0.7232
External #1	0.7025	0.7314
External #2a	0.5709	0.6452
External #2b	0.6648	0.5843
External #3	0.1997	0.5424
DOE	0.6215	0.5623

Table 27c: Coefficient Alpha, Spring 2004, Pre-Test and Post-Test

Course	Pre-Test Alpha	Post-Test Alpha
Engr	0.6882	0.6562
Math #1	0.6797	0.7460
Math #2	0.6906	0.7211

Table 27d: Coefficient Alpha, Fall 2004, Pre-Test and Post-Test

Course	Pre-Test Alpha	Post-Test Alpha
Engr	0.5860	0.4886
Math #1	0.7147	0.6994
Math #2	0.7376	0.7344
Ext. #1	--	0.6239

Table 27e: Coefficient Alpha, Spring 2005, Pre-Test and Post-Test

Course	Pre-Test Alpha	Post-Test Alpha
Engr	0.6190	0.7744
Math #1	0.6416	0.7676
Math #2	0.7723	0.7079
Quality	0.7009	--
Psych	0.4284	0.5918

Alpha was also calculated for the four sub-topics presented in the Construct Validity section. These sub-alphas are shown in Tables 28.

Table 28a: Coefficient Alpha for Sub-Topics, Fall 2003 Post-Test

Course	Probability	Descriptive	Inferential	Graphical
Engr	0.6657	0.6163	0.0542	0.2374
Math #2	0.6460	0.2705	0.5873	0.5528
External #1	0.5583	0.5006	0.5631	-0.2011
External #2a	0.6737	0.5631	0.1151	0.2374
External #2b	0.5140	0.0856	0.3790	0.0512
External #3	0.3446	0.4658	0.1416	0.0000
DOE	0.3259	0.5195	0.1326	0.0297

Table 28b: Coefficient Alpha for Sub-Topics, Spring 2004 Post-Test

Course	Probability	Descriptive	Inferential	Graphical
Engr	0.4652	0.4202	0.2685	-0.0690
Math #1	0.0285	0.5136	0.5181	-0.1685
Math #2	0.2766	0.6150	0.3554	-0.4000

For Summer 2003 (Table 27a), the instrument is generally reliable on the Post-Test but is slightly lacking on the Pre-Test. It is interesting that alpha increases from Pre-Test to Post-Test. Students are encouraged to answer all questions, which leads to a large amount of guessing on the Pre-Test and tends to lower alpha. Research on the mathematical behavior of alpha has shown this to be theoretically plausible (Allen, 2004).

The low alpha at External #1 is a concern because it possibly indicates that the instrument is unreliable at other universities. The instrument was written based on the researchers' knowledge of the Engineering Statistics course as it is taught at OU. Data from the Fall 2003 Post-Test (Table 27b) show that External #1 yields an alpha very

similar to the OU courses. However, a new test site (External #2a and #2b) has alphas that are somewhat lower. More data are needed to determine if the test is reliable at universities outside OU.

In the Spring 2004 courses (Table 27c), the Engr course is slightly lower than previous semesters (see comments about different professor), while the Math courses are almost identical to Math courses from Fall 2003. Data from Fall 2004 and Spring 2004 (Tables 27d and 27e) show the reliability to have stabilized around 0.70 for most courses.

Of the four sub-tests, probability tends to have the highest alpha for Fall 2003 (Table 28a). It is above the recommended 0.60 for three of the seven courses, with another two between 0.50 and 0.60. The lowest (DOE) is 0.3259. The descriptive sub-test is not appreciably lower, with one course above 0.60 and another three between 0.50 and 0.60. The lowest, however, is just 0.0856. The inferential sub-test fares poorer, with just two courses above 0.50 and four below 0.20. The graphical sub-test is also very low, but it should be noted that it only has two questions. Considering that the sub-tests have 10 or fewer items, it may be concluded that the probability and descriptive sub-tests are reliable in themselves for most classes.

The data from Spring 2004 are somewhat different. The descriptive sub-test is the best on average, but it now has the slight advantage of having 12 items compared to still 10 for probability and seven for inferential. Inferential rates second on average, with probability third and the two-question graphical area very poor. Caution is necessary in drawing conclusions about these data for several reasons. First, both Math courses were taught by the same professor, a different professor from Fall 2003. This effectively means there are only two different courses rather than three, much lower than the seven from

Fall 2003. Second, several new items were added and others were revised. It is unclear how this would change the factor analysis solution because the total sample size is not large enough to perform proper analysis.

3.2 Confidence Intervals and Hypothesis Tests on Alpha

The results for the confidence intervals and hypothesis tests are shown in Table 30 (next page, due to its size). Based on a significance level of 0.05, Table 29 (below) summarizes the conclusions of these tests. It is encouraging that none of the courses have an alpha significantly less than 0.60; it is disappointing that none of the classes have an alpha significantly greater than 0.80. Seven of the 15 courses have an alpha that is not significantly different than 0.80. A liberal interpretation could lead to the conclusion that the SCI meets the widely-accepted value of 0.80 for these courses.

Table 29: Summary of Hypothesis Tests on Alpha

Observed Alpha range	**Conclusion of Hypothesis Test**	**Courses**
< 0.60	Alpha less than 0.60	None
	Alpha not less than 0.60 and less than 0.80 (** used as example below table)	Su03 Ex.#1, Su03 REU, F03 Ex.#2b, F03 Ex.#3, F03 DOE
0.60 to 0.80	Alpha not greater than 0.60 and less than 0.80	F02 All, F03 Ex.#2a, Sp04 Engr
	Alpha greater than 0.60 and less than 0.80	None
	Alpha not greater than 0.60 and not less than 0.80	F03 Math, Sp04 Ma.#2
	Alpha greater than 0.60 and not less than 0.80	F03 Engr, F03 Ex.#1, Sp04 Ma.#1
> 0.80	Alpha not greater than 0.80 and greater than 0.60	Su03 Engr, Su03 Math
	Alpha greater than 0.80	None

** Two hypothesis tests: 1. H_0: $\alpha = 0.60$ not rejected; 2. H_0: $\alpha = 0.80$ rejected

Table 30: Confidence Intervals and Hypothesis Tests for Alpha, Post-Tests only

Course	**Alpha**	**Lower C.I.**	**Upper C.I.**	**F, p for H_0: $\alpha = 0.60$**	**F, p for H_0: $\alpha = 0.80$**
F02 All	0.6114	0.5236	0.6899	$F_{173,5363} = 1.03$ ($p = 0.3821$)	$F_{173,5363} = 0.51$ ($p < 0.0001$)
Su03 Engr	0.8100	0.6818	0.9040	$F_{23,736} = 2.11$ ($p = 0.0019$)	$F_{23,736} = 1.05$ ($p = 0.0819$)
Su03 Math	0.8587	0.7267	0.9459	$F_{13,416} = 2.83$ ($p = 0.0007$)	$F_{13,416} = 1.42$ ($p = 0.1486$)
Su03 Ex.#1	0.5781	0.3597	0.7493	$F_{37,1184} = 0.95$ ($p = 0.4406$)	$F_{37,1184} = 0.47$ ($p = 0.0030$)
Su03 REU	0.5983	0.3453	0.7874	$F_{26,832} = 1.00$ ($p = 0.5292$)	$F_{26,832} = 0.50$ ($p = 0.0163$)
F03 Engr	0.7496	0.6421	0.8372	$F_{53,1749} = 1.60$ ($p = 0.0047$)	$F_{53,1749} = 0.80$ ($p = 0.1523$)
F03 Math	0.7232	0.5088	0.8742	$F_{18,594} = 1.45$ ($p = 0.1044$)	$F_{18,594} = 0.72$ ($p = 0.2107$)
F03 Ex.#1	0.7314	0.6018	0.8346	$F_{42,1386} = 1.49$ ($p = 0.0237$)	$F_{42,1386} = 0.74$ ($p = 0.1149$)
F03 Ex.#2a	0.6452	0.4945	0.7683	$F_{53,1749} = 1.13$ ($p = 0.2480$)	$F_{53,1749} = 0.56$ ($p = 0.0046$)
F03 Ex.#2b	0.5843	0.3957	0.7363	$F_{47,1551} = 0.96$ ($p = 0.4535$)	$F_{47,1551} = 0.48$ ($p = 0.0011$)
F03 Ex.#3	0.5424	0.3021	0.7302	$F_{36,1188} = 0.87$ ($p = 0.3176$)	$F_{36,1188} = 0.44$ ($p = 0.0014$)
F03 DOE	0.5623	0.2807	0.7717	$F_{25,825} = 0.91$ ($p = 0.4136$)	$F_{25,825} = 0.46$ ($p = 0.0099$)
Sp04 Engr	0.6562	0.4566	0.8086	$F_{30,1020} = 1.16$ ($p = 0.2504$)	$F_{30,1020} = 0.58$ ($p = 0.0350$)
Sp04 Ma.#1	0.7460	0.5985	0.8586	$F_{30,1020} = 1.57$ ($p = 0.0260$)	$F_{30,1020} = 0.79$ ($p = 0.2133$)
Sp04 Ma.#2	0.7211	0.5561	0.8465	$F_{29,986} = 1.43$ ($p = 0.0649$)	$F_{29,986} = 0.72$ ($p = 0.1358$)

3.3 Guttman Coefficients

The six Guttman reliability estimates are shown in Table 31 for Fall 2003.

Table 31: Guttman Reliability Estimates, Fall 2003 Post-Test

Course	L_1	L_2	L_3	L_4	L_5	L_6
Engr	0.7275	0.7772	0.7496	0.8429	0.7527	0.9143
Math	0.7019	0.7750	0.7232	0.9094	0.7400	1
External #1	0.7099	0.7636	0.7314	0.8826	0.7376	0.9373
External #2a	0.6262	0.6844	0.6452	0.8100	0.6582	0.8613
External #2b	0.5671	0.6387	0.5843	0.8173	0.6035	0.8749
External #3	0.5265	0.6133	0.5424	0.7683	0.5739	0.915
DOE	0.5458	0.6363	0.5623	0.7991	0.5914	1

The estimate based on regression (L_6) is the highest for all courses. Three classes have the maximum value of 1, but this is because the regression equation is over-specified due to having more items than students. The External #2b L_6 may also be a poor estimate because the matrix of scores is nearly singular, which makes regression difficult. Two questions which were missed by all or nearly all students were omitted from the analysis to alleviate this problem. The same is true for External #3, and two questions which were missed by all students were deleted. L_6 appears to be the most sensitive to the assumption that the population of students is indefinitely large (Guttman's Assumption B). The split-half coefficient (L_4) also tends to be high because it is the largest L_4 from 100 random splits. If more splits were run (e.g., 1000), it is likely that even higher values could be found.

Of the remaining estimates (L_1, L_2, L_3, L_5), L_2 is the highest for all classes. This matches Guttman's observation that L_2 will tend to be higher than L_3 if there are negative covariances. Most SCI questions have between 10 and 20 negative covariances. The next-highest is L_5 because it is also based on covariances. L_1 and L_3 are smallest. L_1 will

always be the smallest because the other values are calculated by adding an additional term to L_1. The following relationship holds for all classes:

$$L_1 < L_3 < L_5 < L_2.$$

(L_4 and L_6 omitted, see below)

The Guttman coefficients confirmed the proposals made by Guttman about the relative magnitude of the six estimates. As a high reliability is usually desired, L_2 is the most favorable of the six. L_4 and L_6 should be higher, but L_4 is based on merely finding a proper split, which is likely to have little real meaning, and L_6 will be near 1 unless the number of students is much larger than the number of items.

4. Discrimination

4.1 Ferguson's Delta

The discriminatory power of the SCI is shown in Table 32, as measured by Ferguson's delta.

Table 32: Discriminatory Power, Post-Tests

Course	**Ferguson's Delta**
Fall 02 All courses	0.944
Summer 03 Engr	0.941
Summer 03 Math #1	0.936
Summer 03 REU	0.938
Summer 03 Ext. #1	0.926
Fall 03 Engr	0.964
Fall 03 Math #2	0.918
Fall 03 External #1	0.944
Fall 03 External #2a	0.943
Fall 03 External #2b	0.931
Fall 03 External #3	0.920
Fall 03 DOE	0.911
Spring 04 Engr	0.944
Spring 04 Math #1	0.950
Spring 04 Math #2	0.949

All classes tested show a discriminatory power above the recommended 0.90. Approximately half of the classes fall within a range of 0.930 to 0.945. The highest course is Fall 03 Engr (0.964), and the lowest is Fall 03 DOE (0.911). It is not a surprise that the highest is the class which most closely matches that target audience, while the lowest is the class which is not part of the intended audience.

4.2 Discriminatory Index

A broad view of item discrimination indices for the instrument is presented in Table 33. The application of these results to specific questions is discussed in the Item Analysis section.

Table 33: Item Discriminatory Index, Number of Questions in Each Range, all semesters

Course	**Poor (< 0.20)**	**Moderate (0.20 to 0.40)**	**Good (≥ 0.40)**
Fall 02 All courses	9	17	6
Summer 03 Engr	8	11	14
Summer 03 Math #1	7	5	21
Summer 03 REU	16	7	10
Summer 03 Ext. #1	15	9	9
Fall 03 Engr	11	8	15
Fall 03 Math #2	11	13	10
Fall 03 External #1	10	15	9
Fall 03 External #2a	13	12	9
Fall 03 External #2b	15	11	8
Fall 03 External #3	10	14	10
Fall 03 DOE	14	14	6
Spring 04 Engr	14	12	9
Spring 04 Math #1	11	14	10
Spring 04 Math #2	8	15	12

The discriminatory indices improved from Fall 2002 to Summer 2003 in Engr and Math courses at OU, with more questions rated in the good category. The REU and External #1 each had nearly half the items rated poor for Summer 2003. For Fall 2003, results are relatively constant across the different universities. The best result is the Engr

course (15 items good), but even that is a slight decline from Summer 2003 due to more questions being rated poor. The other groups are similar to the Summer 2003 External #1. For Spring 2004, the results are comparable to Fall 2003. Engr is the poorest (fewest good, most poor) and similar to the External groups from Fall 2003. Math #1 is almost identical to Fall 2003 Math #2, the only difference being due to an extra item on the Spring SCI. Spring 2004 Math #2 is one of the top classes overall, with only eight items rated poor.

4.3 Point-biserial Correlation

Table 34 compares the number of items in the poor, moderate, and good ranges for the discriminatory indices and point-biserial values.

Table 34: Comparison between Discriminatory Index and Point-biserial Correlation, Spring 2004 Post

Course / Statistic	**Poor (< 0.20)**	**Moderate (0.20 to 0.40)**	**Good (≥ 0.40)**
Engr. / r_{pbis}	11	13	11
Engr / disc. index	14	12	9
Math #1 / r_{pbis}	12	13	10
Math #1 / disc. index	11	14	10
Math #2 / r_{pbis}	10	12	13
Math #2 / disc. index	8	15	12

It is clear from the table above that there is approximately the same number of items in each category, but this does not show if the same items are rated good by both statistics. The following matrices examine this possibility.

Table 35: Detailed Breakdown of Discriminatory Index and Point-biserial Correlation, Spring 2004 Post

	Engr Disc. Index					Math #1 Disc. Index					Math #2 Disc. Index		
	Poor	Mod.	Good	r_{pbis}		Poor	Mod.	Good	r_{pbis}		Poor	Mod.	Good
Poor	10	1	0		Poor	10	2	0		Poor	7	2	0
Mod.	4	8	1		Mod.	1	11	1		Mod.	0	7	6
Good	0	3	8		Good	0	1	9		Good	1	6	6

Most items fall in the same category for both metrics. Math #2 is somewhat of an exception, but this is primarily due to questions between 0.30 and 0.50 which are moderate on one measure and high on the other. To alleviate the effect of the arbitrary categories, the correlation between discriminatory index and point-biserial value was calculated for each class, shown below.

Table 36: Correlation between Discriminatory Index and Point-biserial Correlation, Spring 2004 Post (p-value in parenthesis)

Course	**Correlation (p-value)**
Engr	$r = 0.891$ ($p < 10^{-12}$)
Math #1	$r = 0.891$ ($p < 10^{-12}$)
Math #2	$r = 0.787$ ($p < 10^{-7}$)

The point-biserial correlation (r_{pbis}) is highly similar to the discriminatory index. Questions tend to fall in the same category (poor, moderate, good) on both, and the two metrics are highly-correlated. Evaluating both metrics is therefore somewhat redundant, but they can be checked for thoroughness. The decision to focus on one rather than the other seems to be primarily a personal preference.

5. Item Analysis

This section documents the changes to all SCI items. These changes are based on validity, reliability, discrimination, or some combination thereof. The analyses are presented in chronological order beginning with the Fall 2002 pilot instrument. The numbering of the items is provided as a reference point. Correct answers are marked **.

5.1 Description of Questions and Changes

Fall 2002 #1, Summer 2003 #1, Fall 2003 #16, Spring 2004 #9

> The following are temperatures for a week in August: 94, 93, 98, 101, 98, 96, and 93. By how much could the highest temperature increase without changing the median?
>
> a) Increase by 8°
> b) Increase by 2°
> c) It can increase by any amount. **
> d) It cannot increase without changing the median.

The SCI pilot test and focus groups show that students understand the question and utilize logic that is consistent with what the researchers anticipated. Choice D was the most common incorrect choice, and focus groups commented that remembering to order the number before finding the median is essential. The placement of the largest number (101) in the middle of the original list makes this crucial.

The question is generally reliable as measured by alpha-if-deleted (Fall 2002, +0.185, rank 9th). The discriminatory index is high for Fall 2002 (0.56) and Summer 2003 (0.75) Engr class. The most notable short-coming is with the Summer 2003 External #1. The results are presented in Table 37a.

Table 37a: Item Analysis Statistics for Median question, Summer 2003

Course	**Alpha-if-deleted**	**Rank**	**Overall alpha**	**Difference (+ is good)**	**Discriminatory Index**
Engr	0.7968	6th	0.8100	+0.0132	0.75
Math	0.8605	29th	0.8587	-0.0018	0.25
REU	0.5955	22nd	0.5983	+0.0028	0.16
External #1	0.5746	19th	0.5781	+0.0035	0.14
Combined	0.6942	11th	0.7039	+0.0097	0.33

For the Fall 2003 administration, the question fares even better than on the Summer 2003 administration. All courses are reliable by alpha-if-deleted. The two introductory statistics courses at OU (Engr and Math) have extremely high discriminatory

indices. The three External introductory engineering statistics courses (#1, #2a, #2b) have adequate discriminatory indices, but the introductory course at a two-year college (#3) and the DOE course are slightly lacking, at the bottom end of the moderate range. The Spring 2004 results are close to the low-end results from Fall 2003.

Table 37b: Item Analysis Statistics for Median question, Fall 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7327	5th	0.7496	+0.0169	0.63
Math	0.7045	6th	0.7232	+0.0187	0.67
External #1	0.7179	9th	0.7314	+0.0135	0.33
External #2a	0.6311	8th	0.6452	+0.0141	0.28
External #2b	0.5729	11th	0.5843	+0.0114	0.31
External #3	0.5315	14th	0.5424	+0.0109	0.20
DOE	0.5619	23rd	0.5623	+0.0004	0.21
Combined	0.7081	3rd	0.7252	+0.0171	0.59

Table 37c: Item Analysis Statistics for Median question, Spring 2004

Course	Alpha-if-deleted	Rank	Overall alpha	Difference (+ is good)	Discriminatory Index
Engr	0.6377	8th	0.6562	+0.0185	0.32
Math #1	0.7394	14th	0.7460	+0.0066	0.18
Math #2	0.7165	23rd	0.7211	+0.0046	0.24
Combined	0.6984	12th	0.7079	+0.0095	0.31

The topic (data summary) is of extreme importance in the faculty survey, with a mean score of 3.65 out of 4 (2nd highest overall). The question also satisfies the AP Statistics topic "Measuring center: median and mean." Results from the Summer 2003 Pre-Post tests show that some students are gaining this knowledge during their Introductory Statistics course. The results are summarized in Table 38.

Table 38: Knowledge Gain on Median question, Summer 2003

Course	Pre-Test % Correct	Post-Test % Correct	Gain
Engr	44%	67%	23%
Math	53%	79%	26%

Analysis of the answer distributions show that very few students change from being correct on the Pre-Test to incorrect on the Post-Test. For Engr, 6 of 9 students who

were correct on the Pre-Test were also correct on the Post-Test, while 9 of 13 students who were incorrect on the Pre-Test responded correctly on the Post-Test. For the Math course, all 7 students who were correct on the Pre-Test were also correct on the Post-Test, while 3 of the 5 students incorrect on the Pre-Test responded correctly on the Post-Test. This is summarized in the matrices below.

	Engr Pre-Test					Math Pre-Test				
Post	A	B	**C**	D		A	B	**C**	D
A	0	0	**0**	1		0	0	**0**	1
B	0	1	**1**	0		0	0	**0**	0
C	**2**	**2**	**6**	**5**		**0**	**0**	**7**	**3**
D	0	1	**2**	1		0	0	**0**	1

Note: This only includes students who took Pre- and Post-tests. C is the correct answer.

For the SCI, it is important that students show the ability to gain knowledge on questions which are typically covered in an Introductory Statistics course. While many students will be familiar with the median from previous experience, it is a topic which will almost assuredly be covered by the instructor in the course and it is therefore expected that students should have an increased knowledge at the end of the course.

These factors illustrate that the question as originally written is meeting its intended purpose. The question will continue to be monitored along these lines for future administrations.

Fall 2002 #2, Summer 2003 #2, Fall 2003 #12, Spring 2004 #24

> The heights of 5 giraffes are 15 feet, 10 feet, 17 feet, 13 feet, and 16 feet. If the measurements are changed to inches, how will the standard deviation change?
>
> a) Increase by 12
> b) Decrease by 12
> c) Increase by factor of 12 **
> d) Decrease by factor of 12
> e) It won't change

Statistically, the question was moderately successful. The question was positive on alpha-if-deleted (+0.0085, rank 17th), and the discriminatory index was 0.36. Based on focus groups, minor changes were made. The choices containing "decrease" were changed because focus groups felt it was obvious that the answer should be "increase" and choices B and D were too easy to eliminate. These answers were also chosen by few students (5% B, 2% D). The question stem was not changed. No changes were made based on the Summer 2003 results. The new choices are listed below.

a) Increase by 12
b) Increase by a factor of 12 **
c) Increase by a factor of $\sqrt{12}$
d) Increase by a factor of 144
e) It won't change

Based on objective metrics, the question varies from very good to poor depending on the course. For Summer 2003, only External #1 is poor. The Engr class is adequate, while the Math and REU have very good metrics. The Math course even has a perfect discriminatory index, although it should be noted that this was a small class (14 students). The Fall 2003 results are similarly mixed. Three classes are good (DOE, External #1 & #3); two are barely adequate (Engr, Math); two are very poor (External #2a & #2b). The Spring 2004 results are again mixed. The Engr class is very good, while both Math courses are very poor. The combined data are poor for all semesters. The results are summarized in Tables 39.

Table 39a: Item-Analysis Statistics for Change of Units question, Summer 2003

Course	**Alpha-if-deleted**	**Rank**	**Overall Alpha**	**Difference (+ is good)**	**Discriminatory Index**
Engr	0.8088	21st	0.8100	+0.0012	0.32
Math	0.8482	8th	0.8587	+0.0105	1.00
REU	0.5814	11th	0.5983	+0.0169	0.48
External #1	0.5869	29th	0.5781	-0.0088	0.05
Combined	0.7013	27th	0.7039	+0.0026	0.19

Table 39b: Item-Analysis Statistics for Change of Units question, Fall 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7525	26th	0.7496	-0.0029	0.19
Math	0.7195	21st	0.7232	+0.0037	0.21
External #1	0.7252	15th	0.7314	+0.0062	0.44
External #2a	0.6562	31st	0.6452	-0.0110	-0.01
External #2b	0.6012	30th	0.5843	-0.0169	-0.06
External #3	0.5027	4th	0.5424	+0.0397	0.60
DOE	0.5470	14th	0.5623	+0.0153	0.36
Combined	0.7301	32nd	0.7252	-0.0049	0.21

Table 39c: Item Analysis Statistics for Change of Units question, Spring 2004

Course	Alpha-if-deleted	Rank	Overall alpha	Difference (+ is good)	Discriminatory Index
Engr	0.6237	2nd	0.6562	+0.0325	0.62
Math #1	0.7511	29th	0.7460	-0.0051	0.00
Math #2	0.7505	35th	0.7211	-0.0294	-0.34
Combined	0.7160	34th	0.7079	-0.0081	0.08

There is concern about the content validity of this item. Students should encounter changing units in high school and college-freshman-level science or engineering courses. Therefore, the question is not based on statistics knowledge, *per se*. However, it is still imperative to recall that standard deviation has the same units as the measurements, whereas variance has squared units (choice D).

Fall 2002 #3, Summer 2003 #3, Fall 2003 #20, Spring 2004 #34

> You are rolling dice. You roll 2 dice and compute the mean of the number rolled, then 6 dice and compute the mean, then 10 dice and compute the mean. One of the rolls has an average of 1.5. Which trial would you be most surprised to find this result?
>
> a) Rolling 2 dice
> b) Rolling 6 dice
> c) Rolling 10 dice **
> d) This is possible for any of the trials.
> e) There is no way this can happen.

This question has undergone very minor changes. Choice D was deleted because it was only chosen by 10% of students in Fall 2002 and because it technically is true. The

phrase "most surprised" was italicized because focus groups mentioned that the question may get lost with a long story.

The question does not perform highly for all classes, but on average it is a good item. For Fall 2002, the results were relatively neutral (discriminatory index 0.28, alpha-if-deleted +0.0049, rank 21st). For Summer 2003, the Engr and Math courses have poor alpha-if-deleted, but the Engr class shows improvement in the discriminatory index compared to Fall 2002. For Fall 2003, the question has high discriminatory indices (≥ 0.40) for all but two courses. The reliability is similar, and the item ranks 1st for one course (External #2a). The results are summarized in Tables 40. For Spring 2004, the Engr and Math #1 courses are very good on alpha-if-deleted and the discriminatory index, but Math #2 is poor on both.

Table 40a: Item Analysis Statistics for Rolling Dice question, Summer 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.8106	26th	0.8100	-0.0006	0.34
Math	0.8601	27th	0.8587	-0.0014	0.00
REU	0.5732	7th	0.5983	+0.0251	0.45
External #1	0.5524	6th	0.5781	+0.0257	0.42
Combined	0.6980	19th	0.7039	+0.0059	0.39

Table 40b: Item Analysis Statistics for Rolling Dice question, Fall 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7341	8th	0.7496	+0.0155	0.50
Math	0.6993	4th	0.7232	+0.0239	0.50
External #1	0.7348	31st	0.7314	-0.0034	0.19
External #2a	0.6278	1st	0.6452	+0.0174	0.44
External #2b	0.5546	7th	0.5843	+0.0297	0.46
External #3	0.5168	9th	0.5424	+0.0256	0.40
DOE	0.5848	33rd	0.5623	-0.0225	-0.20
Combined	0.7120	10th	0.7252	+0.0132	0.50

Table 40c: Item Analysis Statistics for Rolling Dice question, Spring 2004

Course	Alpha-if-deleted	Rank	Overall alpha	Difference (+ is good)	Discriminatory Index
Engr	0.6244	3rd	0.6562	+0.0318	0.55
Math #1	0.7262	7th	0.7460	+0.0198	0.55
Math #2	0.7218	29th	0.7211	-0.0007	0.26
Combined	0.6932	6th	0.7079	+0.0147	0.41

Fall 2002 #4, Summer 2003 #4, Fall 2003 #29, Spring 2004 #13

> You are dialing into the OU Modem Pool at 10 pm. It takes an average of 25 attempts before connecting. You have attempted 15 dials. How much longer do you expect to wait?
>
> a) 15
> b) 25 **
> c) 10
> d) There is no way to know

For Summer 2003, the answers were re-arranged from largest to smallest. From focus group comments, answer D was slightly changed because the original could technically be true but it was not the best answer. The question was also made more generic by removing the reference to OU. No changes were made for Fall 2003. For Spring 2004, the last sentence was changed to "How many more attempts do you anticipate to have to dial?" because students may get confused about "expected value" as opposed to a Poisson distribution. The updated question is shown below.

> You are dialing into your local internet service provider at 9 pm. It takes an average of 25 attempts before connecting. You have attempted 15 dials. How many more attempts do you anticipate to have to dial?
>
> a) 10
> b) 15
> c) 25 **
> d) There is no way to estimate

Psychometrically, this question is undesirable because it is missed by almost every student regardless of course. From an instructional point-of-view, it is instructive to

know that students do not recognize a situation where the memory-less property of the Poisson distribution should be applied. The top class is Fall 2003 External #1, with 30% correct. The Engr classes at OU are around 25%, and the others are under 10%. The objective statistics are not very meaningful for the low courses, but the question performs well for courses with more acceptable percent correct. This pattern has been apparent since the Fall 2002 pilot study (alpha-if-deleted -0.0027, rank 23rd, discrimination 0.08).

Table 41a: Item Analysis Statistics for Memory-Less Property question, Summer 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7918	3rd	0.8100	+0.0182	0.71
Math	0.8674	32nd	0.8587	-0.0087	-0.25
REU	missed by all	n/a	0.5983	n/a	0
External #1	missed by all	n/a	0.5781	n/a	0
Combined	0.6995	23rd	0.7039	+0.0044	0.11

Table 41b: Item Analysis Statistics for Memory-Less Property question, Fall 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7336	7th	0.7496	+0.0160	0.44
Math	missed by all	n/a	0.7232	n/a	0
External #1	0.7050	1st	0.7314	+0.0264	0.67
External #2a	0.6382	16th	0.6452	+0.0070	0.13
External #2b	0.5890	25th	0.5843	-0.0047	0.03
External #3	missed by all	n/a	0.5424	n/a	0
DOE	0.5627	25th	0.5623	-0.0004	0.11
Combined	0.7192	18th	0.7252	+0.0060	0.26

Table 41c: Item Analysis Statistics for Memory-Less Property question, Spring 2004

Course	Alpha-if-deleted	Rank	Overall alpha	Difference (+ is good)	Discriminatory Index
Engr	0.6434	13th	0.6562	+0.0128	0.22
Math #1	0.7537	33rd	0.7460	-0.0077	-0.09
Math #2	missed by all	n/a	0.7211	n/a	0
Combined	0.7086	30th	0.7079	-0.0007	0.04

There is a concern that this topic is not covered or emphasized in all courses. Future teaching style surveys may help determine actual content covered and pedagogy.

Fall 2002 #5, Summer 2003 #5, Fall 2003 #4, Spring 2004 #20

> The mean height of American college men is 70 inches, with standard deviation 3 inches. The mean height of American college women is 65 inches, with standard deviation 4 inches. You conduct an experiment at OU measuring the height of 100 American men and 100 American women. Which result would most surprise you?
>
> a) A man with height 79 inches
> b) A woman with height 77 inches
> c) The average height of OU women is 68 inches
> d) The average height of OU men is 73 inches **
> e) I am not surprised by anything

For Summer 2003, choice E was removed because it sounds too silly and the word "most" was italicized in the question (focus group comments). In the choices, the "A" was replaced with "One" on choices A and B to be more clear. References to OU were removed. The modified choices are shown below.

> a) One man with height 79 inches
> b) One woman with height 77 inches
> c) The average height of women at your university is 68 inches
> d) The average height of men at your university is 73 inches **

For Fall 2003, choice B was changed to 74 because A and B were both exactly three standard deviations above the mean. In a focus group, an advanced student mentioned that he could eliminate both since there would not be two correct answers. The choices appear below.

> a) One man with height 79 inches
> b) One woman with height 74 inches
> c) The average height of women at your university is 68 inches
> d) The average height of men at your university is 73 inches **

This item has improved its discriminatory index (Fall 2002, 0.14), and alpha-if-deleted has always been very good (Fall 2002, +0.0162, rank 6^{th}). Only one subsequent class has a low discriminatory index (Spring 2004 Engr, 0.10). Focus group comments show that students possess the proper knowledge to answer the question, but arriving at

the correct answer takes more thought than students may be willing to expend when a grade is not at stake.

Table 42a: Item Analysis Statistics for Height question, Summer 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7914	2nd	0.8100	+0.0186	0.73
Math	0.8447	1st	0.8587	+0.0140	1.00
REU	0.5623	2nd	0.5983	+0.0360	0.73
External #1	0.5767	23rd	0.5781	+0.0014	0.33
Combined	0.6820	1st	0.7039	+0.0219	0.62

Table 42b: Item Analysis Statistics for Height question, Fall 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7290	2nd	0.7496	+0.0206	0.63
Math	0.7306	31st	0.7232	-0.0074	0.38
External #1	0.7198	11th	0.7314	+0.0116	0.32
External #2a	0.6527	29th	0.6452	-0.0075	0.24
External #2b	0.5852	20th	0.5843	-0.0009	0.32
External #3	0.5296	13th	0.5424	+0.0128	0.30
DOE	0.5401	4th	0.5623	+0.0222	0.36
Combined	0.7161	13th	0.7252	+0.0091	0.45

Table 42c: Item Analysis Statistics for Height question, Spring 2004

Course	Alpha-if-deleted	Rank	Overall alpha	Difference (+ is good)	Discriminatory Index
Engr	0.6563	24th	0.6562	-0.0001	0.10
Math #1	0.7205	1st	0.7460	+0.0255	0.82
Math #2	0.7122	13th	0.7211	+0.0089	0.37
Combined	0.6943	7th	0.7079	+0.0136	0.51

Fall 2002 #6, Summer 2003 #6, Fall 2003 #11, Spring 2004 #27

In question #5, which sampling method would NOT introduce bias?

a) You measure the OU basketball teams
b) You use a random number table to select students based on their OU ID **
c) You measure international students
d) You ask your friends as a way to get started
e) Any method will have bias
f) None of the methods will have bias

This question is one of the easiest on the SCI. Several classes have had all students answer it correctly. Minor changes to the answers have not changed this fact. Choice E was deleted because it might technically be true. For Summer 2003, the word "random" was also inserted into the incorrect choices because the word "random" in B was considered too big of a hint by focus groups. For Fall 2003, different randomization techniques were added. The stem was also modified so that it did not refer directly back to the previous question. The current question is shown below.

> In order to determine the mean height of American college students, which sampling method would *not* introduce bias?
>
> a) You randomly select from the university basketball team
> b) You use a random number table to select students based on their student ID **
> c) You flip a coin to select from a list of international students
> d) You roll a pair of dice to select from among your friends
> e) None of the methods will have bias

This question is difficult to interpret psychometrically because it may be too easy. This prevents the discriminatory index from being high, and makes alpha-if-deleted very close to the overall alpha. The highest discriminatory index is 0.38 (Engr Summer 2003), and only two courses has alpha-if-deleted above +0.01 (REU Summer 2003, DOE Fall 2003). For Fall 2002, the discriminatory index was 0.24 and alpha-if-deleted was -0.0043 (rank 22nd). The item analysis statistics for Summer and Fall 2003 are shown below.

Table 43a: Item Analysis Statistics for Bias question, Summer 2003

Course	**Alpha-if-deleted**	**Rank**	**Overall Alpha**	**Difference (+ is good)**	**Discriminatory Index**
Engr	0.8025	14th	0.8100	+0.0075	0.38
Math	all correct	n/a	0.8587	n/a	0
REU	0.5812	10th	0.5983	+0.0171	0.14
External #1	0.5732	18th	0.5781	+0.0049	0.07
Combined	0.6959	13th	0.7039	+0.0080	0.21

Table 43b: Item Analysis Statistics for Bias question, Fall 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7424	18th	0.7496	+0.0072	0.25
Math	0.7261	29th	0.7232	-0.0029	0.00
External #1	0.7255	17th	0.7314	+0.0059	0.15
External #2a	0.6411	16th	0.6452	+0.0041	0.05
External #2b	all correct	n/a	0.5843	n/a	0
External #3	0.5352	16th	0.5424	+0.0072	0.20
DOE	0.5484	16th	0.5623	+0.0139	0.14
Combined	0.7199	20th	0.7252	+0.0053	0.19

The Spring 2004 results are an apparent improvement despite no changes to the question. All three courses are comparable to the best from previous semesters. The difference seems to be due to more students missing the question, with the percent correct around 80% for all three compared to usually around 90% from the previous semesters. This allows the discriminatory index to take on higher values, and the item's increased variance allows it to have a bigger influence on alpha.

Table 43c: Item Analysis Statistics for Bias question, Spring 2004

Course	Alpha-if-deleted	Rank	Overall alpha	Difference (+ is good)	Discriminatory Index
Engr	0.6384	10th	0.6562	+0.0178	0.27
Math #1	0.7311	10th	0.7460	+0.0149	0.36
Math #2	0.7037	6th	0.7211	+0.0174	0.44
Combined	0.6923	5th	0.7079	+0.0156	0.38

Fall 2002 #7, Summer 2003 #7, Fall 2003 #10, Spring 2004 #23

Which statistic would you expect to have a normal distribution?

I) Height of men
II) Age of pennies in circulation
III) Age of college freshmen

a) I only **
b) II only
c) III only
d) I & II
e) II & III
f) I & III
g) All 3
h) None

For Summer 2003, the least-chosen answers were deleted to get down to five choices (C 5%, E 5%, H 2%). The remaining choices are shown below.

a) I only **
b) II only
c) I & II
d) I & III
e) All 3

For Fall 2003, focus groups thought that the "age of pennies" may have been too confusing, so it was changed to "shoe size of men." This changed the correct answer. Age was also specified to be in Years (III) to avoid confusion. The new choices are below.

I) Height of women
II) Shoe size of men
III) Age in years of college freshmen

a) I only
b) II only
c) I & II **
d) I & III
e) All 3

The item tends to be successful on the objective metrics. The Fall 2002 results were low on the discriminatory index (0.13) but acceptable on reliability (alpha-if-deleted +0.0109, rank 14th). All subsequent classes have been positive on alpha-if-deleted except Spring 2004 Engr, and only two groups have a low discriminatory index (Summer 2003 REU, Spring 2004 Engr).

Table 44a: Item Analysis Statistics for Normal Distribution question, Summer 2003

Course	**Alpha-if-deleted**	**Rank**	**Overall Alpha**	**Difference (+ is good)**	**Discriminatory Index**
Engr	0.7987	10th	0.8100	+0.0113	0.50
Math	0.8524	15th	0.8587	+0.0063	0.59
REU	0.5939	20th	0.5983	+0.0044	0.11
External #1	0.5700	14th	0.5781	+0.0081	0.25
Combined	0.6954	12th	0.7039	+0.0085	0.31

Table 44b: Item Analysis Statistics for Normal Distribution question, Fall 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7390	11th	0.7496	+0.0106	0.50
Math	0.7121	10th	0.7232	+0.0111	0.55
External #1	0.7181	10th	0.7314	+0.0133	0.48
External #2a	0.6434	25th	0.6452	+0.0018	0.31
External #2b	0.5800	17th	0.5843	+0.0043	0.29
External #3	0.5120	6th	0.5424	+0.0304	0.30
DOE	0.5406	5th	0.5623	+0.0217	0.50
Combined	0.7108	6th	0.7252	+0.0144	0.63

Table 44c: Item Analysis Statistics for Normal Distribution question, Spring 2004

Course	Alpha-if-deleted	Rank	Overall alpha	Difference (+ is good)	Discriminatory Index
Engr	0.6654	30th	0.6562	-0.0092	0.19
Math #1	0.7302	8th	0.7460	+0.0158	0.55
Math #2	0.7151	19th	0.7211	+0.0060	0.28
Combined	0.7008	19th	0.7079	+0.0071	0.37

Fall 2002 #8, Summer 2003 #8, Fall 2003 #19, Spring 2004 #17

A researcher reports a 95% confidence interval for the mean, which of the following must be true?

a) You are 95% confident you performed the measurements correctly
b) 95% of the measurements can be considered valid
c) There is a 95% chance that the population mean will be between the upper and lower limits
d) Your experimental mean will fall in a range that contains the true mean 95% of the time **
e) None of the above

For Summer 2003, choice A was deleted because it was not chosen (2%), and C was changed because it was too similar and hard to distinguish from D. The new choices appear below.

a) It is probable that 95% of the confidence intervals will be identical
b) 95% of the measurements can be considered valid
c) 95% of the measurements will be between the upper and lower limits
d) Your experimental mean will fall in a range that contains the population mean 95% of the time **
e) None of the above

For Fall 2003, choice E was deleted because focus groups said they tend to automatically delete this type of answer, which agrees with Gibb's rules. All of the answers were made to start with a number. The phrase "of the confidence interval" was added to B (formerly C) to be more specific and make it about the same length as D. Choice A was deleted because it was not chosen by anyone in Engr or Math; it was replaced with a new D.

a) 95% of the measurements can be considered valid
b) 95% of the measurements will be between the upper and lower limits of the confidence interval
c) 95% of the time, your experimental mean will fall in a range that contains the population mean **
d) 5% of the measurements should be considered outliers

For Fall 2002, the question performed very poorly (discriminatory index 0.11, alpha -0.0092, rank 29th). The changes appear to be successful because there was has been improvement in the discriminatory index and alpha-if-deleted for all classes. The item is among the best for Fall 2003 Engr, External #1, and External #2b, as well as the Spring 2004 Math courses. The results are shown in Tables 45.

Table 45a: Item Analysis Statistics for Confidence Interval question, Summer 2003

Course	**Alpha-if-deleted**	**Rank**	**Overall Alpha**	**Difference (+ is good)**	**Discriminatory Index**
Engr	0.7982	9th	0.8100	+0.0118	0.75
Math	0.8539	17th	0.8587	+0.0048	0.50
REU	0.5868	13th	0.5983	+0.0115	0.32
External #1	0.5826	27th	0.5781	-0.0045	0.19
Combined	0.6967	15th	0.7039	+0.0072	0.33

Table 45b: Item Analysis Statistics for Confidence Interval question, Fall 2003

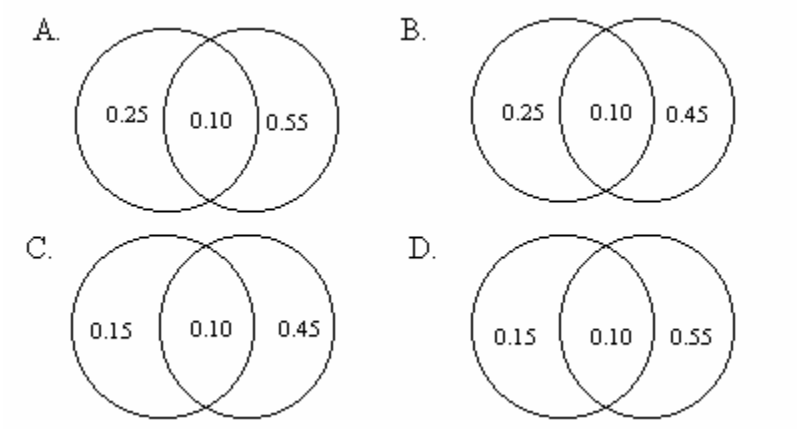
Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7322	3rd	0.7496	+0.0174	0.63
Math	0.7139	14th	0.7232	+0.0093	0.38
External #1	0.7062	2nd	0.7314	+0.0252	0.60
External #2a	0.6398	20th	0.6452	+0.0054	0.35
External #2b	0.5298	1st	0.5843	+0.0545	0.83
External #3	0.5327	15th	0.5424	+0.0097	0.50
DOE	0.5424	7th	0.5623	+0.0199	0.38
Combined	0.7090	5th	0.7252	+0.0162	0.55

Table 45c: Item Analysis Statistics for Confidence Interval question, Spring 2004

Course	Alpha-if-deleted	Rank	Overall alpha	Difference (+ is good)	Discriminatory Index
Engr	0.6473	15th	0.6562	+0.0089	0.39
Math #1	0.7254	5th	0.7460	+0.0206	0.73
Math #2	0.7024	5th	0.7211	+0.0187	0.60
Combined	0.6895	3rd	0.7079	+0.0184	0.51

Fall 2002 #9, deleted

The union of A and B = 0.80. The intersection of A and B = 0.10. A = 0.25. Which diagram correctly illustrates these conditions?



Note: D is the correct answer

This question had a negative effect on alpha (overall alpha 0.6114, alpha-if-deleted 0.6174); it is one of just 8 of 32 questions to be unreliable by this measure. The discriminatory index was 0.30, which falls between the low and high ranges.

Aside from choice C (chosen by 2%), the answer distribution is indicative of guessing (36% A, 28% B, 35% D). The item fits only loosely into the AP Statistics category "Exploring Data" and the faculty survey category "Methods of Displaying Data." This implies that the item does not conform to content validity. This item was not discussed in focus groups because the above considerations had already rendered it inappropriate.

Fall 2002 #10, Summer 2003 #9, Fall 2003 #31, Spring 2004 #8

> A student scored in the 90th percentile in his Chemistry class. Which is always true?
>
> a) His grade will be an A
> b) He earned 90% of the total possible points
> c) His grade is higher than 90% of his classmates **
> d) None of these are always true

For Summer 2003, answer C was changed to be more technically correct. This is important because C is the correct answer, and students should be not be misled into answering D. The new choice is shown below.

> c) His grade is at least as high as 90% of his classmates **

For Fall 2003, answer B was changed so that it could possibly be more appealing to students who remember the "at least" part of percentiles but are unsure of how it applies. The new choice is shown below.

> b) He earned at least 90% of the total possible points

Statistically, this item is mixed between being very good for some classes and very poor for others. For Fall 2002, the item ranks 4th on alpha (+0.0221) and has a discriminatory index of 0.45 (high range). For Summer 2003, the item has a very high discriminatory index for the Engr course and a very good alpha-if-deleted for the REU group, but it performs poorly for the Math and External #1 courses. The results are

similarly polarized for the Fall 2003 courses, but the results are adequate for Spring 2004. It is unclear why this item tends to behave in this manner. The combined data are acceptable, with the lowest discriminatory index 0.21 (Summer 2003) and all alpha-if-deleted on the positive side.

Table 46a: Item Analysis Statistics for Percentile question, Summer 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.8002	11th	0.8100	+0.0098	0.75
Math	0.8602	28th	0.8587	-0.0015	0.00
REU	0.5706	5th	0.5983	+0.0277	0.32
External #1	0.5976	32nd	0.5781	-0.0195	0.19
Combined	0.6982	20th	0.7039	+0.0057	0.21

Table 46b: Item Analysis Statistics for Percentile question, Fall 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7354	9th	0.7496	+0.0142	0.50
Math	0.7202	22nd	0.7232	+0.0030	0.17
External #1	0.7336	26th	0.7314	-0.0022	0.02
External #2a	0.6341	12th	0.6452	+0.0111	0.22
External #2b	0.5921	27th	0.5843	-0.0078	-0.05
External #3	0.5141	8th	0.5424	+0.0283	0.50
DOE	0.5440	9th	0.5623	+0.0183	0.61
Combined	0.7177	15th	0.7252	+0.0075	0.36

Table 46c: Item Analysis Statistics for Percentile question, Spring 2004

Course	Alpha-if-deleted	Rank	Overall alpha	Difference (+ is good)	Discriminatory Index
Engr	0.6495	18th	0.6562	+0.0067	0.43
Math #1	0.7360	12th	0.7460	+0.0100	0.45
Math #2	0.7158	21st	0.7211	+0.0053	0.23
Combined	0.7003	17th	0.7079	+0.0079	0.41

Fall 2002 #11, Summer 2003 #11, Fall 2003 #32, Spring 2004 #35

When calculating a confidence interval on a given population, using a larger sample size will make the confidence interval:

a) Smaller **
b) Larger
c) No change
d) It depends on the confidence level

For Summer 2003, the item was not changed. For Fall 2003, the phrase "confidence level" on choice D was changed to "significance level" to help avoid confusion between "confidence interval" and "confidence level." For Spring 2004, there was no change.

The item statistics tend to be acceptable but are poor for several courses. There is a slight decline from Fall 2002 (+0.0112, rank 11th, discrimination 0.32) and Summer 2003 compared to Fall 2003, but it is unlikely that the minor change would cause this. It would be wise to construct a similar item and compare the item analysis statistics. The better question could be kept. This item covers a different aspect of confidence intervals than Fall 2002 #8, but it is worthwhile to note that #8 fares much better statistically.

Table 47a: Item Analysis Statistics for Confidence Interval question, Summer 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.8153	31st	0.8100	-0.0053	0.11
Math	0.8472	5th	0.8587	+0.0115	1.00
REU	0.5780	8th	0.5983	+0.0203	0.34
External #1	0.5786	24th	0.5781	-0.0005	0.13
Combined	0.7006	26th	0.7039	+0.0033	0.30

Table 47b: Item Analysis Statistics for Confidence Interval question, Fall 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7613	33rd	0.7496	-0.0117	0.19
Math	0.7121	10th	0.7232	+0.0111	0.38
External #1	0.7252	15th	0.7314	+0.0062	0.36
External #2a	0.6574	32nd	0.6452	-0.0122	0.06
External #2b	0.5738	12th	0.5843	+0.0105	0.23
External #3	0.5402	20th	0.5424	+0.0022	0.20
DOE	0.5797	30th	0.5623	-0.0174	0.25
Combined	0.7248	26th	0.7252	+0.0004	0.31

Table 47c: Item Analysis Statistics for Confidence Interval question, Spring 2004

Course	Alpha-if-deleted	Rank	Overall alpha	Difference (+ is good)	Discriminatory Index
Engr	0.6382	9th	0.6562	+0.0180	0.48
Math #1	0.7555	34th	0.7460	-0.0095	0.09
Math #2	0.7128	15th	0.7211	+0.0083	0.27
Combined	0.7045	22nd	0.7079	+0.0034	0.33

Fall 2002 #12, Summer 2003 #12, Fall 2003 #24, Spring 2004 #4

Which would be more likely to have 70% boys born on a given day: A small rural hospital or a large urban hospital?

a) Rural **
b) Urban
c) Equally likely
d) Both are impossible

For Summer 2003, focus groups commented that D is too obviously wrong because anything is possible. The item was not changed for Fall 2003 and Spring 2004. The new choice follows.

d) Both are extremely unlikely

This item is strong statistically across all courses, from Fall 2002 (+0.0174, rank 3rd, discrimination 0.37) onward. This shows that the change to D did not have an adverse effect. All courses rate positive on alpha-if-deleted (maximum rank 2nd, minimum 17th). The discriminatory index also tends to be in the high range (maximum 1.00, minimum 0.25). Focus group remarks by incorrect students indicate that they tend to focus on location rather than size.

Table 48a: Item Analysis Statistics for Hospital problem, Summer 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.8056	15th	0.8100	+0.0044	0.43
Math	0.8478	6th	0.8587	+0.0109	0.75
REU	0.5697	4th	0.5983	+0.0286	0.61
External #1	0.5529	7th	0.5781	+0.0252	0.54
Combined	0.6862	2nd	0.7039	+0.0177	0.54

Table 48b: Item Analysis Statistics for Hospital problem, Fall 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7323	4th	0.7496	+0.0173	0.63
Math	0.6860	2nd	0.7232	+0.0372	1.00
External #1	0.7174	8th	0.7314	+0.0140	0.38
External #2a	0.6335	11th	0.6452	+0.0117	0.41
External #2b	0.5481	4th	0.5843	+0.0362	0.59
External #3	0.5024	3rd	0.5424	+0.0400	0.30
DOE	0.5469	13th	0.5623	+0.0154	0.25
Combined	0.7054	2nd	0.7252	+0.0198	0.59

Table 48c: Item Analysis Statistics for Hospital question, Spring 2004

Course	Alpha-if-deleted	Rank	Overall alpha	Difference (+ is good)	Discriminatory Index
Engr	0.6312	6th	0.6562	+0.0250	0.44
Math #1	0.7408	17th	0.7460	+0.0052	0.27
Math #2	0.7141	17th	0.7211	+0.0070	0.50
Combined	0.6962	10th	0.7079	+0.0117	0.43

Fall 2002 #13, Summer 2003 #14, Fall 2003 #13, Spring 2004 #31

If P (A|B) = 0.70, what is P (B|A)?

a) 0.70
b) 0.30
c) 1.00
d) 0
e) Not enough information **
f) Other: _____

This question had a negative effect on alpha (overall alpha 0.6114, alpha-if-deleted 0.6153). It is one of just 8 of 32 questions to be unreliable by this measure. The discriminatory index was 0.16, which was the 8th worst for the 32-item Fall 2002 SCI and in the low range. The researchers felt the question was too symbol-oriented, which could confuse some students. The correct answer also may be an option which students would naturally want to disregard (Gibb's Categorical Exclusive), making the problem unfair.

The topic (conditional probability) was considered too important to omit. It is listed explicitly in the AP Statistics outline. Conditional probability scored 2.85 out of 4

on the faculty survey (mean for all topics 2.63, median 2.62). Therefore, a new question was devised which was less formulaic and more focused on the concept.

> In a manufacturing process, the error rate is 1 in 1000. However, errors often occur in bursts. Given that the previous output contained an error, what is the probability that the next unit will also contain an error?
>
> a) Less than 1 in 1000
> b) Greater than 1 in 1000 **
> c) Equal to 1 in 1000
> d) Insufficient information

Focus group comments reveal that students have an understanding of the problem's purpose (i.e., conditional probability or a "non-memory-less" property). One student correctly chose B because the problem did not say memory-less, while another made the connection that the bursts would "throw off the odds" (direct quote). Several students felt the notion of a burst was not well-defined, which led to choice D. This potential problem must be monitored on future administrations. The question was not further revised for the Fall 2003 SCI.

From the Summer administration, the question generally had a positive effect on alpha. The exception is the Math course, which has been cited for possible differences in teaching method and topics covered. The answer distribution for the Math course indicates possible guessing (33% B, 42% C, 25% D). The discriminatory index displays the same basic pattern as alpha-if-deleted and is at an acceptable level. From Fall 2003, the results are basically the same with the most notable short-coming being the introductory course at External #1. The External #3 and DOE courses are not as much of a concern because they do not match the target audience. The Spring 2004 results are comparable, although the Math courses are poor on alpha-if-deleted.

Table 49a: Item Analysis Statistics for Conditional Probability problem, Summer 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7977	4th	0.8100	+0.0123	0.45
Math	0.8626	31st	0.8587	-0.0039	0.00
REU	0.5899	17th	0.5983	+0.0084	0.34
External #1	0.5297	2nd	0.5781	+0.0484	0.55
Combined	0.6901	4th	0.7039	+0.0138	0.42

Table 49b: Item Analysis Statistics for Conditional Probability problem, Fall 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7357	10th	0.7496	+0.0139	0.50
Math	0.7170	17th	0.7232	+0.0062	0.24
External #1	0.7342	29th	0.7314	-0.0028	0.19
External #2a	0.6282	7th	0.6452	+0.0170	0.40
External #2b	0.5649	8th	0.5843	+0.0194	0.40
External #3	0.5550	26th	0.5424	-0.0126	0.10
DOE	0.5907	34th	0.5623	-0.0284	-0.05
Combined	0.7198	19th	0.7252	+0.0054	0.35

Table 49c: Item Analysis Statistics for Conditional Probability question, Spring 2004

Course	Alpha-if-deleted	Rank	Overall alpha	Difference (+ is good)	Discriminatory Index
Engr	0.6262	4th	0.6562	+0.0300	0.44
Math #1	0.7515	31st	0.7460	-0.0055	0.27
Math #2	0.7199	28th	0.7211	+0.0012	0.29
Combined	0.7032	21st	0.7079	+0.0047	0.29

Fall 2002 #14, Summer 2003 #15, Fall 2003 #14, deleted

You perform a hypothesis test and calculate a p-value of 0.05. What does this mean?

a) There is a 5% possibility the observed value is due to chance **
b) There is a 5% possibility the null hypothesis is true
c) There is a 95% possibility the null hypothesis is true
d) None of the above

For Summer 2003, choice D was changed to something people might choose. The new D is wrong because it says "largest" instead of "smallest."

d) 0.05 is the largest level of significance which could lead to rejection of the null hypothesis

For Fall 2003, a new choice was added which was the opposite of A (the correct answer). All of the choices are shown below.

a) There is a 5% possibility the observed value is due to chance **
b) There is a 95% possibility that the observed value is due to chance
c) There is a 5% possibility the null hypothesis is true
d) There is a 95% possibility the null hypothesis is true
e) 0.05 is the largest level of significance which could lead to rejection of the null hypothesis

For Spring 2004, the item was deleted and replaced by a question about interpreting p-value (knowing when to reject).

Focus groups indicate that many students are guessing and that choice E was too misleading. Normally, only a few people get it right in each class on the Post. Often, the gain is negative, which means people guessed correctly on the Pre but do not get it right when they should know it. The item depends too much on how students learn the definition rather than conceptual understanding.

This question is generally poor with alpha and discrimination, although it has improved since the Fall 2002 administration (alpha -0.0010, rank 25th, discrimination 0.09). A similar item (Fall 2002 #26) tends to have better statistics, although it too has the same problem with the negative gains.

Table 50a: Item Analysis Statistics for p-value question, Summer 2003

Course	**Alpha-if-deleted**	**Rank**	**Overall Alpha**	**Difference (+ is good)**	**Discriminatory Index**
Engr	0.8058	16th	0.8100	+0.0042	0.14
Math	0.8512	13th	0.8587	+0.0075	0.50
REU	0.6191	32nd	0.5983	-0.0208	-0.04
External #1	0.5752	20th	0.5781	+0.0029	0.08
Combined	0.6993	22nd	0.7039	+0.0046	0.20

Table 50b: Item Analysis Statistics for p-value question, Fall 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7545	29th	0.7496	-0.0049	0.19
Math	0.7114	9th	0.7232	+0.0118	0.40
External #1	0.7322	25th	0.7314	-0.0008	0.20
External #2a	0.6401	22nd	0.6452	+0.0051	0.34
External #2b	0.5703	10th	0.5843	+0.0140	0.36
External #3	0.5124	7th	0.5424	+0.0300	0.70
DOE	0.5406	5th	0.5623	+0.0217	0.23
Combined	0.7223	22nd	0.7252	+0.0029	0.33

Fall 2002 #15, Summer 2003 #16, Fall 2003 #1, Spring 2004 #19

Which is true of a t-distribution?

a) It is shaped like a t
b) It describes a population
c) It is used when the population standard deviation is not known **
d) It has the same basic shape as a normal distribution but has skinnier tails
e) b & d are both true
f) c & d are both true

For Summer 2003, choice A was deleted because it was not chosen (2%) and focus groups thought it was too silly. The phrase "skinnier tails" in D was made clearer by using the term "less area." There were no changes for Fall 2003 or Spring 2004. The new choices are shown below.

a) It describes a population
b) It is used when the population standard deviation is not known **
c) It has the same basic shape as a normal distribution but has less area in the tails
d) a & c are both true
e) b & c are both true

This item has poor item analysis statistics for most classes. This was apparent from the pilot study (alpha-if-deleted -0.0193, rank 31st, discriminatory index 0.15), and the changes were not successful at changing this. The highest discriminatory index is

0.35 (Spring 2004 Engr), and it is negative for several classes. The item is also negative on alpha-if-deleted for over half the courses.

This item is a candidate for replacement. The concept relates closely to hypothesis testing, which is broken down into several areas on the faculty survey. All of the areas where a t-test could be used rank above the overall mean.

Table 51a: Item Analysis Statistics for t-distribution question, Summer 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.8076	19th	0.8100	+0.0024	0.23
Math	missed by all	n/a	0.8587	n/a	0
REU	0.6073	28th	0.5983	-0.0090	-0.02
External #1	0.5948	31st	0.5781	-0.0167	0.02
Combined	0.7139	33rd	0.7039	-0.0097	0.02

Table 51b: Item Analysis Statistics for t-distribution question, Fall 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7589	31st	0.7496	-0.0093	0.00
Math	0.7026	5th	0.7232	+0.0206	0.29
External #1	0.7171	7th	0.7314	+0.0143	0.32
External #2a	0.6537	30th	0.6452	-0.0085	0.12
External #2b	0.6025	31st	0.5843	-0.0182	-0.03
External #3	0.5795	32nd	0.5424	-0.0371	-0.10
DOE	0.5826	32nd	0.5623	-0.0203	0.23
Combined	0.7289	28th	0.7252	-0.0037	0.13

Table 51c: Item Analysis Statistics for t-distribution question, Spring 2004

Course	Alpha-if-deleted	Rank	Overall alpha	Difference (+ is good)	Discriminatory Index
Engr	0.6484	17th	0.6562	+0.0078	0.35
Math #1	0.7508	28th	0.7460	-0.0048	-0.09
Math #2	0.7111	11th	0.7211	+0.0100	0.30
Combined	0.7071	27th	0.7079	+0.0008	0.22

Fall 2002 #16, Summer 2003 #17, Fall 2003 #7, deleted

> 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

For Summer 2003, the phrase "very accurate" was italicized to be more noticeable. For Fall 2003, choice E was deleted because it was consistently not chosen (0% usually). For Spring 2004, the item was deleted because it depends too much on knowing what percent chance of rain means. There were also copyright concerns because it came out of a journal (Konold, 1995). It was replaced by a question about rain in two different cities (Spring 2004 #21).

Statistically, the question was acceptable and sometimes very good (Fall 2002, alpha-if-deleted +0.0225, rank 2nd, discriminatory index 0.43). The results from Summer and Fall 2003 are shown in Tables 52.

Table 52a: Item Analysis Statistics for Chance of Rain problem, Summer 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.8098	23rd	0.8100	+0.0002	0.20
Math	0.8541	18th	0.8587	+0.0046	0.50
REU	0.5782	9th	0.5983	+0.0201	0.48
External #1	0.5710	15th	0.5781	+0.0071	0.32
Combined	0.6978	18th	0.7039	+0.0061	0.39

Table 52b: Item Analysis Statistics for Chance of Rain problem, Fall 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7422	17th	0.7496	+0.0074	0.50
Math	0.6838	1st	0.7232	+0.0394	0.86
External #1	0.7272	20th	0.7314	+0.0042	0.45
External #2a	0.6317	9th	0.6452	+0.0135	0.47
External #2b	0.5527	5th	0.5843	+0.0316	0.57
External #3	0.5733	30th	0.5424	-0.0309	-0.20
DOE	0.5550	20th	0.5623	+0.0073	0.23
Combined	0.7135	11th	0.7252	+0.0117	0.49

Fall 2002 #17, Summer 2003 #19, Fall 2003 #33, Spring 2004 #1

> You are a doctor testing a blood-born disease. You know the overall population has a rate of 2% positive. All positives are accurately detected. You also know that the test returns a positive result for 5% of people who do not have the disease. What is the probability that a patient will test positive?
>
> a) 0.02
> b) 0.05
> c) 0.02 + 0.05
> d) 0.02 + 0.05*0.98 **
> e) 0.05 – 0.02

For Summer 2003, choice E was replaced because it was only chosen by 6% and focus groups felt that it made no sense why anyone would choose it. The new choice looks more plausible.

> e) (0.02 + 0.05)*0.98

For Fall 2003, choice B was deleted because it was not chosen in Engr or Math classes.

The item analysis statistics are generally acceptable, although they started poorly on the pilot study (discriminatory index 0.30, alpha-if-deleted -0.0102, rank 30th). There are still classes where the item performs poorly, but it is usually acceptable for the target courses. As more items are added or revised, this item should be monitored to see if it is still borderline acceptable or if it becomes poor for most classes.

Table 53a: Item Analysis Statistics for False Positives problem, Summer 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7955	5th	0.8100	+0.0145	0.73
Math	0.8580	21st	0.8587	+0.0007	0.25
REU	0.5941	21st	0.5983	+0.0042	0.14
External #1	0.5833	28th	0.5781	-0.0052	0.20
Combined	0.6977	17th	0.7039	+0.0062	0.42

Table 53b: Item Analysis Statistics for False Positives problem, Fall 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7463	22nd	0.7496	+0.0033	0.44
Math	0.7338	32nd	0.7232	-0.0106	-0.10
External #1	0.7135	4th	0.7314	+0.0179	0.56
External #2a	0.6388	18th	0.6452	+0.0064	0.37
External #2b	0.5765	14th	0.5843	+0.0078	0.26
External #3	0.5033	5th	0.5424	+0.0391	0.50
DOE	0.5625	24th	0.5623	-0.0002	0.23
Combined	0.7157	12th	0.7252	+0.0095	0.48

Table 53c: Item Analysis Statistics for False Positives question, Spring 2004

Course	Alpha-if-deleted	Rank	Overall alpha	Difference (+ is good)	Discriminatory Index
Engr	0.6617	25th	0.6562	-0.0055	0.21
Math #1	0.7443	21st	0.7460	+0.0017	0.27
Math #2	0.7381	34th	0.7211	-0.0170	-0.19
Combined	0.7149	33rd	0.7079	-0.0070	0.09

Fall 2002 #18, Summer 2003 #20, Fall 2003 #23, Spring 2004 #33

For the past 100 years, the average high temperature on October 1 is 78° with a standard deviation of 5°. What is the probability that the high temperature on October 1, 2003, will be between 73° and 83°?

a) 0.50
b) 0.68 **
c) 0.95
d) 0.997
e) Other: _____

For Summer 2003, a new choice E was added based on write-ins to "Other."

e) 1.00

For Fall 2003, choice A was deleted because it was chosen by generally less than 10% and there is no statistical reason for it to be 0.50. There were no changes for Spring 2004.

On the pilot study, the item had a moderately low discriminatory index (0.25) but a good alpha-if-deleted (+0.0210, rank 8th). On subsequent administrations, the item fared very well for most courses. The results are shown in the following tables.

Table 54a: Item Analysis Statistics for Normal Distribution question, Summer 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7884	1st	0.8100	+0.0216	1.00
Math	0.8559	19th	0.8587	+0.0028	0.50
REU	0.6153	31st	0.5983	-0.0170	0.07
External #1	0.5752	20th	0.5781	+0.0029	0.01
Combined	0.6913	6th	0.7039	+0.0126	0.37

Table 54b: Item Analysis Statistics for Normal Distribution question, Fall 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7430	20th	0.7496	+0.0066	0.38
Math	0.6907	3rd	0.7232	+0.0325	0.83
External #1	0.7077	3rd	0.7314	+0.0237	0.66
External #2a	0.6022	2nd	0.6452	+0.0430	0.63
External #2b	0.5300	2nd	0.5843	+0.0543	0.76
External #3	0.5282	12th	0.5424	+0.0142	0.20
DOE	0.5446	10th	0.5623	+0.0177	0.13
Combined	0.7037	1st	0.7252	+0.0215	0.64

Table 54c: Item Analysis Statistics for Normal Distribution question, Spring 2004

Course	Alpha-if-deleted	Rank	Overall alpha	Difference (+ is good)	Discriminatory Index
Engr	0.6359	7th	0.6562	+0.0203	0.46
Math #1	0.7236	3rd	0.7460	+0.0224	0.73
Math #2	0.7114	12th	0.7211	+0.0097	0.39
Combined	0.6877	2nd	0.7079	+0.0202	0.58

Fall 2002 #19, deleted

You believe in global warming. You use the date October 1 as your reference (see above problem). What will be your alternate hypothesis to test global warming?

a) H_0: mean = 78
b) H_0: mean > 78
c) H_1: mean = 78
d) H_1: mean $\neq$ 78
e) H_1: mean > 78 **
f) H_1: mean < 78

This item was poor from a theoretical point-of-view because it is not possible to use just one data point to test a hypothesis. There is also an issue of independence if global warming is, in fact, happening. It was decided devise a new hypothesis question at a later date (eventually the bottling problem). The objective statistics were acceptable but not great (discriminatory index 0.31, alpha-if-deleted +0.0102, rank 19th).

Fall 2002 #20, Summer 2003 #24, Fall 2003 #14, Spring 2004 #3

Which of the following could never be considered a population?

a) The students in your statistics class
b) The football teams in the Big 12
c) The players on a football team
d) Three randomly selected Wal-Mart stores **

Focus group comments provided useful insights with this question. Specifically, incorrect answers were eliminated by logic that did not match the researchers' goal for this question. For choice A, the concept of bias was mentioned by a student who felt that you would not want to conduct an experiment on a group that you are closely associated with. The choice was changed to "a physics class." Several students felt that the number of items in the choice was important, and this led to at least one student correctly choosing D because it is the smallest number.

The researchers felt that choice D was too obvious because it was the only choice that contains the word "random." Choice C was modified to help eliminate this problem.

This question was kept in a similar format for three reasons. First, it was a reliable question in terms of alpha-if-deleted (0.5970, overall alpha 0.6114). Secondly, the question's discriminatory index is 0.36, which ranks 9th best out of 32 items. Finally, the question fit into the AP Statistics category "Populations, samples, and random selection."

Which of the following could never be considered a population?

a) The students in a physics class
b) The football teams in the Big 12
c) The players on a randomly selected football team
d) 100 randomly selected Wal-Mart stores **

This new version of the item has potential problems with reliability and the discriminatory index. The results are summarized in Table 55a.

Table 55a: Item Analysis Statistics for Population question, Summer 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.8128	28th	0.8100	-0.0028	0.09
Math	0.8611	30th	0.8587	-0.0024	0.25
REU	0.6050	27th	0.5983	-0.0067	0.16
External #1	0.5799	26th	0.5781	-0.0018	0.25
Combined	0.7063	30th	0.7039	-0.0024	0.16

Focus group comments revealed minor problems that required revision. One student chose D because it is the only option that is not people. Choice A, the least-selected option, was changed so that it did not relate to people.

Upon further inspection, it became clear that choice D looks different from the incorrect options because it is the only choice that does not begin with "The" (similar to Gibb's Precise criteria). Further modification was made so that the answers looked more uniform.

Which of the following could never be considered a population?

a) Four-door cars produced in a factory in Detroit
b) Football teams in the Big 12
c) Players on a randomly selected football team
d) One hundred randomly selected Wal-Mart stores **

Results from the Fall 2003 Post-Test indicate that the item's reliability has improved due to the changes discussed above. The results are summarized for Fall 2003 in Table 55b.

Table 55b: Item Analysis Statistics for Population question, Fall 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7545	29th	0.7496	-0.0049	0.19
Math	0.7114	9th	0.7232	+0.0118	0.40
External #1	0.7322	25th	0.7314	-0.0008	0.20
External #2a	0.6401	22nd	0.6452	+0.0051	0.34
External #2b	0.5703	10th	0.5843	+0.0140	0.36
External #3	0.5124	7th	0.5424	+0.0300	0.70
DOE	0.5406	5th	0.5623	+0.0217	0.23
Combined	0.7223	22nd	0.7252	+0.0029	0.33

The first five groups match the target audience of Introductory Statistics courses (although one is in the Mathematics department). While the results are mixed, it is an improvement over the Summer administration. With a maximum discriminatory index of 0.40 and two courses with a negative difference, the question may still need improvement. It is interesting that the question performs very well for the lower-level External #3 and the advanced DOE courses. The combined data is acceptable, but the alpha-if-deleted is barely on the positive side.

For Spring 2004, the question is very good for the Engr class. The Math courses are poorer but very similar to the best classes from Summer 2003 and still an improvement over the poorest classes from Summer. The combined data is slightly better than the combined data from the Fall.

Table 55c: Item Analysis Statistics for Population question, Spring 2004

Course	Alpha-if-deleted	Rank	Overall alpha	Difference (+ is good)	Discriminatory Index
Engr	0.6184	1st	0.6562	+0.0378	0.73
Math #1	0.7473	26th	0.7460	-0.0013	0.27
Math #2	0.7195	26th	0.7211	+0.0016	0.18
Combined	0.7005	18th	0.7079	+0.0074	0.40

Fall 2002 #21, Summer 2003 #25, Fall 2003 #6, Spring 2004 #5

A fair coin is flipped four times in a row, each time landing with heads up. What is the most likely outcome if the coin is flipped a fifth time?

a) tails, because even though for each flip heads and tails are equally likely, since there have been four heads, tails is slightly more likely
b) heads, because this coin apparently likes to fall heads up
c) tails, because in any sequence of tosses, there should be about the same number of heads and tails
d) heads and tails are equally likely because each toss is independent of the others **
e) tails, because there have been so many heads, we are due a tail

The two original Head/Tail problems (second item follows) were adapted from Konold, *et al.* (1993). These basic ideas were further expanded on by Hirsch and O'Donnell (2001), who developed a multiple choice instrument to identify students who held the "representativeness misconception." The instrument, published in the article, contained variations on the same theme. The questions were repetitive and asked the students to first select an answer and then to select a reason.

For Summer 2003, the phrase "most likely" was italicized. The reason was removed from D so that it is not leading, based on focus group comments. For Fall 2003, choice B to was changed to "it has a pattern of falling heads up" instead of it "likes to", which sounds a little silly. Choice C was deleted because it was not chosen. There were no changes for Spring 2004. The updated answers follow.

a) Tails, because even though for each flip heads and tails are equally likely, since there have been four heads, tails is slightly more likely
b) Heads, because this coin has a pattern of landing heads up
c) Tails, because in any sequence of tosses, there should be about the same number of heads and tails
d) Heads and tails are equally likely **

This question is usually answered correctly by a high percent of students on both Pre and Post. This is not ideal psychometrically, but it could help keep students from getting too frustrated if every question is hard.

The item appeared to be very good on the objective metrics for the first two rounds of testing (Fall 2002, discriminatory index 0.48, alpha-if-deleted +0.0144, rank 10th). However, the item did not fare well on the Fall 2003 administration, with most classes rating poor on discrimination. Spring 2004 was also poor, primarily due to the question being too easy (all three courses above 90% correct).

Table 56a: Item Analysis Statistics for Coin Flipping problem, Summer 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.8059	17th	0.8100	+0.0041	0.59
Math	0.8534	16th	0.8587	+0.0053	0.75
REU	0.5832	12th	0.5983	+0.0151	0.18
External #1	0.5712	16th	0.5781	+0.0069	0.31
Combined	0.6985	21st	0.7039	+0.0054	0.21

Table 56b: Item Analysis Statistics for Coin Flipping problem, Fall 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7402	14th	0.7496	+0.0094	0.31
Math	0.7230	24th	0.7232	+0.0002	0.19
External #1	0.7292	22nd	0.7314	+0.0022	0.08
External #2a	0.6502	28th	0.6452	-0.0050	-0.06
External #2b	0.5879	23rd	0.5843	-0.0036	0.02
External #3	0.5441	23rd	0.5424	-0.0017	0.20
DOE	0.5544	19th	0.5623	+0.0079	0.14
Combined	0.7231	24th	0.7252	+0.0021	0.11

Table 56c: Item Analysis Statistics for Coin Flipping question, Spring 2004

Course	Alpha-if-deleted	Rank	Overall alpha	Difference (+ is good)	Discriminatory Index
Engr	0.6529	21st	0.6562	+0.0033	0.09
Math #1	all correct	n/a	0.7460	n/a	0
Math #2	0.7185	25th	0.7211	+0.0026	0.11
Combined	0.7061	25th	0.7061	+0.0018	0.07

Fall 2002 #22, Summer 2003 #26, Fall 2003 #27, Spring 2004 #16

Which of the following sequences is least likely to result from flipping a fair coin five times?

(I) HHHTH
(II) HTHTH
(III) THTTH

a) (I) because the number of heads and tails should be more equal
b) (II) because the pattern of heads and tails should be more random
c) (III) because there are too many tails relative to the number of heads
d) all the sequences are equally unlikely to occur because any sequence of five tosses has the exact same probability of occurring **
e) the sequences do not have the same probability of occurring, but we cannot say which is least likely to occur

For Summer 2003, the "because…" statement of D was removed so that it would not lead students. For Fall 2003, the reasons were removed from A, B, and C for the same reason. There were no changes for Spring 2004. The updated choices are shown below.

a) (I)
b) (II)
c) (III)
d) All the sequences are equally unlikely to occur **
e) The sequences do not have the same probability of occurring, but we cannot say which is least likely to occur

Statistically, this question is similar but more difficult than the previous question which also related to coin-flipping. For Fall 2002, the item was moderately successful (alpha-if-deleted +0.0011, rank 15th, discriminatory index 0.34). Analysis from Summer 2003 and Fall 2003 show that the changes to the question have had little effect. Most

classes rank in the teens on alpha-if-deleted and have discriminatory indices around 0.30. Spring 2004 is similar.

Table 57a: Item Analysis Statistics for Coin Sequence question, Summer 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.8244	33rd	0.8100	-0.0144	-0.11
Math	0.8451	2nd	0.8587	+0.0136	1.00
REU	0.5890	15th	0.5983	+0.0093	0.34
External #1	0.5588	8th	0.5781	+0.0193	0.41
Combined	0.6999	24th	0.7039	+0.0040	0.28

Table 57b: Item Analysis Statistics for Coin Sequence question, Fall 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7401	13th	0.7496	+0.0095	0.44
Math	0.7184	18th	0.7232	+0.0048	0.24
External #1	0.7260	19th	0.7314	+0.0054	0.34
External #2a	0.6334	10th	0.6452	+0.0118	0.37
External #2b	0.5882	24th	0.5843	-0.0039	0.28
External #3	0.5409	10th	0.5424	+0.0015	0.20
DOE	0.5332	24th	0.5623	+0.0291	0.61
Combined	0.7201	21st	0.7252	+0.0051	0.38

Table 57c: Item Analysis Statistics for Coin Sequence question, Spring 2004

Course	Alpha-if-deleted	Rank	Overall alpha	Difference (+ is good)	Discriminatory Index
Engr	0.6480	16th	0.6562	+0.0082	0.28
Math #1	0.7569	35th	0.7460	-0.0109	0.00
Math #2	0.7131	16th	0.7211	+0.0080	0.28
Combined	0.7082	29th	0.7079	-0.0003	0.26

Fall 2002 #23, deleted

In recent years, many areas of the United States have been affected by major flooding. A particular flood is called a "100-year flood". What does this mean?

a) A flood of that magnitude will occur once and only once every 100 years.
b) It is the largest flood that has occurred in the last 100 years.
c) Every year, there is a 1% chance of having a flood that big. **
d) There won't be another flood that big for 100 years.
e) If the flood did not take place for 99 years then there is a 100% chance that it will occur in the 100th year

This question was deleted because it depends too much on knowing what a 100-year flood means and interpreting it as a probability. This is a problem for the content validity of the SCI. Statistically, the question was marginally acceptable (alpha-if-deleted +0.0122, rank 13th, discriminatory index 0.22).

Fall 2002 #24, Summer 2003 #28, Fall 2003 #2, Spring 2004 #12

> A student attended college A for two semesters and earned a 3.24 GPA (grade point average). The same student then attended college B for four semesters and earned a 3.80 GPA for his work there. How would you calculate the student's GPA for all of his college work? Assume that the student took the same number of hours each semester.
>
> a) $\frac{3.24+3.80}{2}$
> b) $\frac{3.24+3.80}{6}$
> c) $\frac{3.24(2)+3.80(4)}{2}$
> d) $\frac{3.24(2)+3.80(4)}{6}$ **
> e) It is not possible to calculate the students overall GPA without knowing his GPA for each individual semester.

The inspiration for this item came from a journal article (Pollatsek, *et al.,* 1981) about student concepts on the mean. For Summer 2003, choice B was deleted because it was only chosen by 2%. No further changes have been made.

The question performs well on reliability, and the discriminatory index is usually at least 0.30. For Fall 2002, the discriminatory index was 0.45 and alpha-if-deleted was +0.0052 (rank 23rd). The data for subsequent semesters are shown in Tables 58.

Table 58a: Item Analysis Statistics for GPA question, Summer 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.8068	18th	0.8100	+0.0032	0.36
Math	0.8498	10th	0.8587	+0.0089	0.75
REU	0.5508	1st	0.5983	+0.0475	0.71
External #1	0.5668	10th	0.5781	+0.0113	0.36
Combined	0.6925	8th	0.7039	+0.0014	0.47

Table 58b: Item Analysis Statistics for GPA question, Fall 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7398	12th	0.7496	+0.0098	0.31
Math	0.7221	23rd	0.7232	+0.0011	0.33
External #1	0.7256	18th	0.7314	+0.0058	0.39
External #2a	0.6399	21st	0.6452	+0.0053	0.17
External #2b	0.5427	3rd	0.5843	+0.0416	0.46
External #3	0.5387	19th	0.5424	+0.0037	0.30
DOE	0.5429	8th	0.5623	+0.0194	0.43
Combined	0.7116	9th	0.7252	+0.0136	0.44

Table 58c: Item Analysis Statistics for GPA question, Spring 2004

Course	Alpha-if-deleted	Rank	Overall alpha	Difference (+ is good)	Discriminatory Index
Engr	0.6500	20th	0.6562	+0.0062	0.18
Math #1	0.7372	13th	0.7460	+0.0088	0.36
Math #2	0.7019	4th	0.7211	+0.0192	0.57
Combined	0.6991	14th	0.7079	+0.0088	0.38

Fall 2002 #25, Summer 2003 #29, Fall 2003 #34, Spring 2004 #22

You perform 2 hypothesis tests on the same population. The first has a p-value of 0.01; the second has a p-value of 0.02. The sample mean is equal for the 2 tests. Which test has a larger sample size?

a) First test **
b) Second test
c) Sample sizes equal because sample means equal
d) Insufficient information (describe what else you need to know): _____

For Summer 2003, choice C was slightly modified and a new D was added. Information was added to the stem based on write-in answers to D. The updated item is shown below. The item was not changed for Fall 2003 or Spring 2004.

You perform the same two significance tests on large samples from the same population. The two samples have the same mean and the same standard deviation. The first test results in a p-value of 0.01; the second, a p-value of 0.02. The sample mean is equal for the 2 tests. Which test has a larger sample size?

a) First test **
b) Second test
c) Sample sizes equal
d) Sample sizes are not equal but there is not enough information to determine which sample is larger

This question fares well on the objective metrics. For Fall 2002, the discriminatory index was 0.34 and alpha-if-deleted was +0.0096 (rank 20th). Data from Summer 2003 and Fall 2003 are better than Fall 2002 for most classes. This suggests the changes were improvements. Spring 2004 is also generally good, although one only of the three classes has a discriminatory index above 0.40. The data are shown below.

Table 59a: Item Analysis Statistics for p-value question, Summer 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.8096	22nd	0.8100	+0.0004	0.34
Math	0.8456	3rd	0.8587	+0.0131	0.75
REU	0.5894	16th	0.5983	+0.0089	0.16
External #1	0.5661	9th	0.5781	+0.0120	0.40
Combined	0.6930	9th	0.7039	+0.0109	0.45

Table 59b: Item Analysis Statistics for p-value question, Fall 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7408	15th	0.7496	+0.0088	0.50
Math	0.7153	16th	0.7232	+0.0079	0.57
External #1	0.7278	21st	0.7314	+0.0036	0.29
External #2a	0.5993	1st	0.6452	+0.0459	0.81
External #2b	0.5571	7th	0.5843	+0.0272	0.41
External #3	0.5368	18th	0.5424	+0.0056	0.20
DOE	0.5108	1st	0.5623	+0.0515	0.73
Combined	0.7108	6th	0.7252	+0.0144	0.57

Table 59c: Item Analysis Statistics for p-value question, Spring 2004

Course	Alpha-if-deleted	Rank	Overall alpha	Difference (+ is good)	Discriminatory Index
Engr	0.6402	11th	0.6562	+0.0160	0.37
Math #1	0.7223	2nd	0.7460	+0.0237	0.73
Math #2	0.7161	22nd	0.7211	+0.0050	0.26
Combined	0.6917	4th	0.7079	+0.0162	0.47

Fall 2002 #26, Summer 2003 #18, Fall 2003 #25, Spring 2004 #18

A researcher performs a t-test to test the following hypotheses:

$$H_0 : \mu \leq \mu_0$$

$$H_1 : \mu > \mu_0$$

He rejects the null hypothesis and reports a p-value of 0.1. Which of the following must be true?

a) The test statistic fell within the rejection region at the $\alpha = 0.05$ significance level.
b) The power of the test statistic used was 90%.
c) The is a 10% possibility that the observed value is due to chance **
d) The probability that the null hypothesis is not true is 0.1
e) The probability that the null hypothesis is actually true is 0.9

This question has not changed. The response pattern to this question is similar to the other question about the meaning of p-value (Fall 2002 #14) – most classes have a negative gain and very few students are correct on the post-test. The objective statistics are slightly better but still very poor for some classes. For Fall 2002, the discriminatory index was low (0.17) but alpha-if-deleted was acceptable (+0.0099, rank 16th).

Table 60a: Item Analysis Statistics for p-value question, Summer 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7946	4th	0.8100	+0.0154	0.59
Math	0.8464	4th	0.8587	+0.0123	0.75
REU	0.6210	33rd	0.5983	-0.0227	0.09
External #1	0.5481	4th	0.5781	+0.0300	0.54
Combined	0.6877	3rd	0.7039	+0.0162	0.51

Table 60b: Item Analysis Statistics for p-value question, Fall 2003

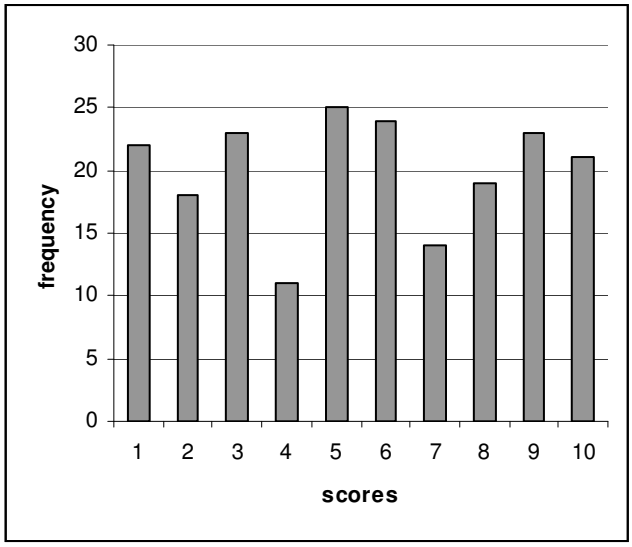
Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7495	25th	0.7496	+0.0001	0.13
Math	0.7136	13th	0.7232	+0.0096	0.29
External #1	0.7340	27th	0.7314	-0.0026	0.18
External #2a	0.6274	5th	0.6452	+0.0178	0.45
External #2b	0.5802	18th	0.5843	+0.0041	0.26
External #3	0.5181	10th	0.5424	+0.0243	0.20
DOE	0.5789	29th	0.5623	-0.0166	0.00
Combined	0.7226	23rd	0.7252	+0.0026	0.21

Table 60c: Item Analysis Statistics for p-value question, Spring 2004

Course	Alpha-if-deleted	Rank	Overall alpha	Difference (+ is good)	Discriminatory Index
Engr	0.6456	14th	0.6562	+0.0106	0.35
Math #1	0.7411	18th	0.7460	+0.0049	0.27
Math #2	0.7008	3rd	0.7211	+0.0203	0.68
Combined	0.6961	9th	0.7079	+0.0118	0.43

Fall 2002 #27, Summer 2003 #21, Fall 2003 #28, Spring 2004 #25

Consider the sample distribution below. The population from which this sample was taken most likely has what kind of distribution?



a) normal
b) exponential
c) uniform **
d) lognormal
e) bimodal

For Summer 2003, choice D was changed to "Skewed" because students may eliminate "lognormal" based on unfamiliarity (focus groups). The question changed to "This sample was *most likely* taken from what kind of population distribution?" to eliminate confusion about "population" being mentioned at beginning of question (focus groups). For Fall 2003, choice B was deleted because it was chosen by only 1 person in Engr and 0 in Math for Summer 2003. There were no additional changes for Spring 2004.

This item tends to have strong statistics. For Fall 2002, the discriminatory index was 0.45 and alpha-if-deleted was +0.0226 (rank 1st). Summer 2003 and Fall 2003 metrics are not as strong on reliability, but the discriminatory index is higher for several courses. Spring 2004 data are all positive on reliability, and two of the three courses have high discriminatory indices.

Table 61a: Item Analysis Statistics for Parent Distribution question, Summer 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7969	7th	0.8100	+0.0131	0.59
Math	0.8501	11th	0.8587	+0.0086	0.75
REU	0.6140	30th	0.5983	-0.0157	0.18
External #1	0.5696	13th	0.5781	+0.0085	0.31
Combined	0.6920	7th	0.7039	+0.0119	0.37

Table 61b: Item Analysis Statistics for Parent Distribution question, Fall 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7411	16th	0.7496	+0.0085	0.38
Math	0.7122	12th	0.7232	+0.0110	0.40
External #1	0.7340	27th	0.7314	-0.0026	0.22
External #2a	0.6192	3rd	0.6452	+0.0260	0.58
External #2b	0.5798	16th	0.5843	+0.0045	0.25
External #3	0.5550	26th	0.5424	-0.0126	-0.10
DOE	0.5558	22nd	0.5623	+0.0065	0.20
Combined	0.7168	14th	0.7252	+0.0084	0.45

Table 61c: Item Analysis Statistics for Parent Distribution question, Spring 2004

Course	Alpha-if-deleted	Rank	Overall alpha	Difference (+ is good)	Discriminatory Index
Engr	0.6499	19th	0.6562	+0.0063	0.24
Math #1	0.7249	4th	0.7460	+0.0211	0.64
Math #2	0.7197	27th	0.7211	+0.0014	0.48
Combined	0.6968	11th	0.7079	+0.0111	0.47

Fall 2002 #28, deleted

For a small sample size (< 30) the sample statistic used to calculate the confidence interval on the mean is

a) z
b) t **
c) σ
d) χ^2

The question was deleted because it was too definition-oriented and because there was another question about the t-statistic. Psychometrically, the item was marginally high on discrimination (0.36) and good on reliability (+0.0211, rank 6th).

Fall 2002 #29, deleted

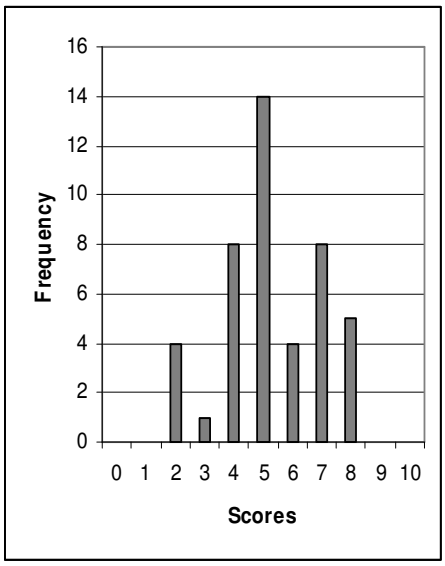
Which of the following statements is true:

a) The probability that the null hypothesis is correct is equal to α
b) If the null hypothesis is rejected, then the test proves that the alternate hypothesis is correct
c) In all statistical tests of hypothesis, the sum of type 1 and type 2 error is 1
d) The probability of rejecting the null hypothesis when the null hypothesis is false is called as the power of the statistical test. **

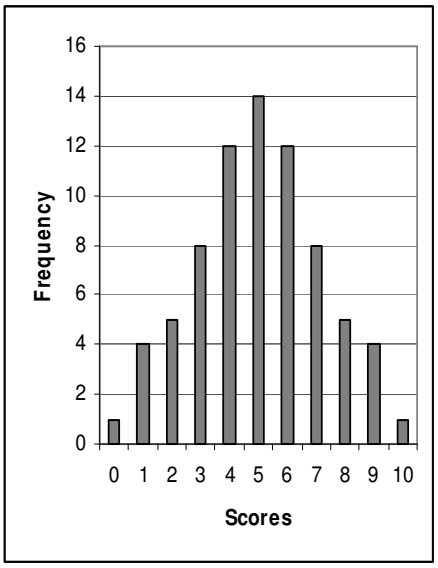
This question was deleted for similar reasons as the previous question. The decision was made to focus on improving the other questions about hypothesis testing. On objective metrics, this question fared well (discriminatory index 0.36, alpha-if-deleted +0.0221, rank 4th). This item is a good candidate for reinstatement if another question about hypothesis testing is needed.

Fall 2002 #30, Summer 2003 #32, Fall 2003 #21, Spring 2004 #30

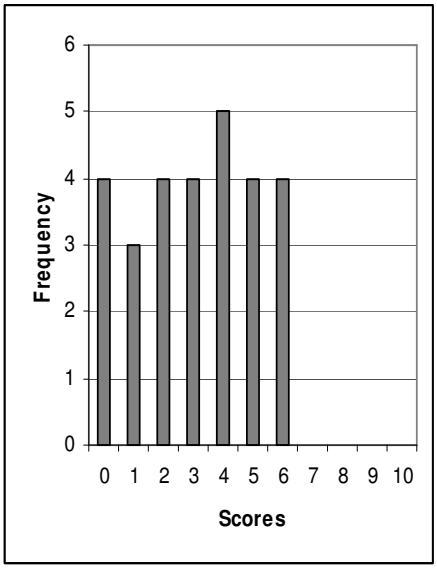
Which of the following distributions shows more variability?



I



II



III

Why did you choose the answer above?

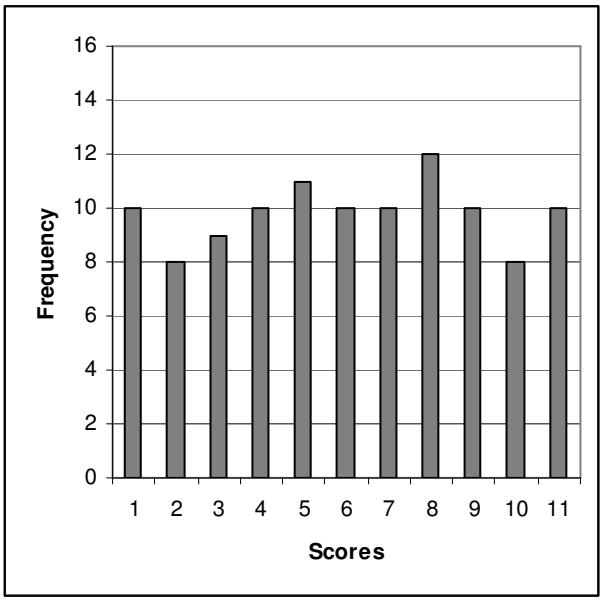
a) Because it's bumpier.
b) Because it's more spread out.
c) Because it has a large number of different scores.
d) Because the values differ more from the center.
e) Other (please specify)

_____

This question was originally two parts, but it seemed to confuse many students because they only answered the "Why" part but did not pick a graph. For Summer 2003, the "Why" part was a separate question (#33). The choices to the graphical portion were also written in the same format as all the other questions. Some students now will simply circle a graph, but having the letter options eliminates confusion.

a) I
b) II
c) III **
d) The variability is equal for all three
e) Insufficient information

For Fall 2003, choices D and E were eliminated because focus groups commented that they are not viable options. A fourth graph (below) was added, which is now the correct answer. Students who considered range to be the best simple estimate of variability would have incorrectly chosen graph II on the original question. Graph II and Graph III on the original choices were actually very close in standard deviation, but Graph III was still the correct choice. There were no changes for Spring 2004.



This is the correct answer now.

The item statistics for Fall 2002 are meaningless because many students left the question blank due to its poor construction. The discrimination index was 0.14 and alpha-if-deleted was -0.0193 (rank 28th). For later semesters, the item tends to perform well on reliability but have low-to-moderate discriminatory indices. This item is one of the most difficult on the instrument, and it is not possible to obtain high discriminatory indices with questions of extreme difficulty. Most classes are close to 20% correct, except the two Math courses which were 29% (Summer 2003) and 37% (Fall 2003). The difference is not very meaningful because these courses are much smaller than the others.

Table 62a: Item Analysis Statistics for Graphical Variability question, Summer 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.8128	7th	0.8100	-0.0028	0.21
Math	0.8505	11th	0.8587	+0.0082	0.75
REU	0.6021	30th	0.5983	-0.0038	0.21
External #1	0.5693	13th	0.5781	+0.0088	0.25
Combined	0.7028	28th	0.7039	+0.0011	0.28

Table 62b: Item Analysis Statistics for Graphical Variability question, Fall 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7483	16th	0.7496	+0.0013	0.38
Math	0.7099	12th	0.7232	+0.0133	0.40
External #1	0.7219	27th	0.7314	+0.0095	0.22
External #2a	0.6249	4th	0.6452	+0.0203	0.38
External #2b	0.5698	9th	0.5843	+0.0145	0.28
External #3	0.5429	26th	0.5424	-0.0005	-0.10
DOE	0.5488	22nd	0.5623	+0.0135	0.20
Combined	0.7178	16th	0.7252	+0.0074	0.30

Table 62c: Item Analysis Statistics for Graphical Variability question, Spring 2004

Course	Alpha-if-deleted	Rank	Overall alpha	Difference (+ is good)	Discriminatory Index
Engr	0.6415	12th	0.6562	+0.0147	0.35
Math #1	0.7513	30th	0.7460	-0.0053	0.00
Math #2	0.7153	20th	0.7211	+0.0058	0.29
Combined	0.7059	24th	0.7079	+0.0020	0.22

Fall 2002 #31, deleted

> Each year, Computerworld magazine reports the Datapro ratings of all computer software vendors. Vendors are rated on a scale from 1 to 4 (1=poor, 4=excellent) in such areas as reliability, efficiency, ease of installation, and ease of use by a random sample of software users. A software vendor wants to determine whether the product has a higher mean Datapro rating than a rival vendor's product. Which formulation of hypotheses is correct (let H_0 represent the null hypothesis and H_1 represent the alternative hypothesis):
>
> a) $H_0 : \mu_{vendor}-\mu_{rival} \leq 0 ; H_1= \mu_{vendor}-\mu_{rival} >0$ **
> b) $H_0 : \mu_{vendor}-\mu_{rival} = 0 ; H_1= \mu_{vendor}-\mu_{rival} <0$
> c) $H_0 : \mu_{vendor}-\mu_{rival} = 0 ; H_1= \mu_{vendor}-\mu_{rival} \neq 0$
> d) $H_0 : \mu_{vendor}-\mu_{rival} > 0 ; H_1= \mu_{vendor}-\mu_{rival} = 0$

This item was deleted because there are other questions about hypothesis testing and because all the symbols could be confusing. Statistically, the item was average (discriminatory index 0.30, alpha-if-deleted +0.0083, rank 18th).

Fall 2002 #32, deleted

> If a random sample of n observations, $y_1, y_2, \mathrm{K}\ y_n$, is selected from a normal distribution with mean μ and variance σ^2, then the sampling distribution of the variance has a
>
> a) Poisson distribution
> b) Chi-square distribution **
> c) Binomial distribution
> d) Exponential Distribution

The question was deleted because it sounds too much like a definition. It is also unlikely that an introductory class would cover this topic. It could be a good idea to test with an advanced class. The discriminatory index was moderate (0.30), but the alpha-if-deleted was poor (-0.0022, rank 27th).

Summer 2003 #10, Fall 2003 #3, Spring 2004 #15

For the following set of numbers, which measure will most accurately describe the central tendency?

3, 3, 4, 5, 6, 8, 10, 12, 19, 36, 83

a) Mean
b) Median **
c) Mode
d) Standard deviation

This question was added to improve the content validity. Measures of central tendency ranked 2nd on the faculty survey. There have been no changes to the item. The item usually does not score well on objective metrics. The reliability is positive for most courses but barely so, and the rank is around the middle. The discriminatory index is around 0.20 for most classes, which is the border of the low and moderate ranges. The Spring 2004 results stand out as very poor, but there is no clear reason why this is so since the question has not changed.

Table 63a: Item Analysis Statistics for Central Tendency question, Summer 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.8099	24th	0.8100	+0.0001	0.36
Math	0.8495	9th	0.8587	+0.0092	0.75
REU	0.6005	24th	0.5983	-0.0022	0.18
External #1	0.5466	3rd	0.5781	+0.0315	0.42
Combined	0.6976	16th	0.7039	+0.0063	0.36

Table 63b: Item Analysis Statistics for Central Tendency question, Fall 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7469	23rd	0.7496	+0.0027	0.31
Math	0.7357	33rd	0.7232	-0.0125	-0.24
External #1	0.7299	23rd	0.7314	+0.0015	0.21
External #2a	0.6385	17th	0.6452	+0.0067	0.26
External #2b	0.5837	19th	0.5843	+0.0006	0.20
External #3	0.5464	25th	0.5424	-0.0040	0.10
DOE	0.5401	2nd	0.5623	+0.0222	0.50
Combined	0.7234	25th	0.7252	+0.0018	0.28

Table 63c: Item Analysis Statistics for Central Tendency question, Spring 2004

Course	Alpha-if-deleted	Rank	Overall alpha	Difference (+ is good)	Discriminatory Index
Engr	0.6643	28th	0.6562	-0.0081	0.19
Math #1	0.7474	27th	0.7460	-0.0014	0.09
Math #2	0.7372	33rd	0.7211	-0.0161	-0.07
Combined	0.7147	32nd	0.7079	-0.0068	0.20

Summer 2003 #13, Fall 2003 #30, Spring 2004 #14

> An architectural firm wants to design a building that minimizes energy loss through the exterior walls. There are four types of insulation that can be used within the walls and three types of bricks which can be used on the exterior walls. What is the best analysis design to use?
>
> a) Two-factor full factorial **
> b) Two-factor full factorial without interaction
> c) Two-factor fractional factorial
> d) Two one-factor ANOVAS, one for insulation type and one for brick type

This item was added to test the SCI on advanced classes. The objective metrics are not meaningful for introductory classes because students will have to guess, but it is interesting to look at the DOE class. Several other classes have positive reliability and high discriminatory indices. The item statistics for the DOE class are acceptable but not as high as expected.

Table 64a: Item Analysis Statistics for ANOVA question, Summer 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.8101	25th	0.8100	-0.0001	0.18
Math	0.8576	20th	0.8587	+0.0011	0.50
REU	0.6036	26th	0.5983	-0.0053	0.21
External #1	0.5208	1st	0.5781	+0.0573	0.70
Combined	0.6938	10th	0.7039	+0.0101	0.42

Table 64b: Item Analysis Statistics for ANOVA question, Fall 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7647	34th	0.7496	-0.0151	-0.06
Math	0.7603	34th	0.7232	-0.0371	-0.24
External #1	0.7343	30th	0.7314	-0.0029	0.37
External #2a	0.6418	24th	0.6452	+0.0034	0.41
External #2b	0.5860	21st	0.5843	-0.0017	0.12
External #3	0.5214	11th	0.5424	+0.0210	0.40
DOE	0.5472	15th	0.5623	+0.0151	0.30
Combined	0.7357	34th	0.7252	-0.0105	0.06

Table 64c: Item Analysis Statistics for ANOVA question, Spring 2004

Course	Alpha-if-deleted	Rank	Overall alpha	Difference (+ is good)	Discriminatory Index
Engr	0.6552	23rd	0.6562	+0.0010	0.35
Math #1	0.7461	23rd	0.7460	-0.0001	0.09
Math #2	0.7181	24th	0.7211	+0.0030	0.40
Combined	0.7068	26th	0.7079	+0.0011	0.22

Summer 2003 #22, Fall 2003 #8, Spring 2004 #7

A farmer wants to know if the monthly yield of a crop is dependent on the amount of fertilizer used, the amount of water used, and the average daily high temperature. What is the best analysis design to use?

a) Three linear regressions, one for each parameter
b) Multiple regression **
c) A three-factor full factorial
d) A three-factor fractional factorial
e) Three one-factor ANOVAs, one for each parameter

This item is similar to the previous one about a factorial ANOVA. Statistically, it rates very poorly for almost every class. It is interesting that the DOE class had no students get this correct. This is most likely because they did not cover regression and opted for choices that were more familiar. The item analysis statistics are shown below.

Table 65a: Item Analysis Statistics for Multiple Regression question, Summer 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.8113	27th	0.8100	-0.0013	0.29
Math	0.8590	23rd	0.8587	-0.0003	0.50
REU	0.5915	19th	0.5983	+0.0068	0.11
External #1	0.5758	22nd	0.5781	+0.0023	0.18
Combined	0.7004	25th	0.7039	+0.0035	0.22

Table 65b: Item Analysis Statistics for Multiple Regression question, Summer 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7600	32nd	0.7496	-0.0104	0.00
Math	0.7247	28th	0.7232	-0.0015	0.24
External #1	0.7379	33rd	0.7314	-0.0065	0.06
External #2a	0.6435	26th	0.6452	+0.0017	0.15
External #2b	0.6038	32nd	0.5843	-0.0195	0.03
External #3	0.5741	31st	0.5424	-0.0317	-0.10
DOE	missed by all	n/a	0.5623	n/a	0
Combined	0.7297	30th	0.7252	-0.0045	0.15

Table 65c: Item Analysis Statistics for Multiple Regression question, Spring 2004

Course	Alpha-if-deleted	Rank	Overall alpha	Difference (+ is good)	Discriminatory Index
Engr	0.6756	35th	0.6562	-0.0194	-0.05
Math #1	0.7471	25th	0.7460	-0.0011	0.27
Math #2	0.7240	32nd	0.7211	-0.0029	0.01
Combined	0.7161	35th	0.7079	-0.0082	0.13

Summer 2003 #23, Fall 2003 #22, Spring 2004 #10

A bottling company believes a machine is under-filling 20-ounce bottles. What will be the alternate hypothesis to test this belief?

a) H_1: mean = 20
b) H_1: mean $\neq$ 20
c) H_1: mean > 20
d) H_1: mean < 20 **

This question was added to fill the coverage gap after deleting the hypothesis test questions about global warming and computer software ratings. For Spring 2004, the format of the answers was changed based on an open-ended SCI given to an advanced course. Many of the students wrote out the answers in words rather than symbols, and these were incorporated as the new choices. They are shown below.

a) On average, the bottles are being filled to 20 ounces.
b) On average the bottles are not being filled to 20 ounces.
c) On average, the bottles are being filled with more than 20 ounces.
d) On average, the bottles are being filled with less than 20 ounces. **

On the objective metrics, it seems that the item needed improvement. Most classes are negative on reliability and low to moderate on discrimination. After the changes, the Spring 2004 data for the Math courses are among the best, but the Engr course is still poor. More data are needed to determine if the changes improved the item.

Table 66a: Item Analysis Statistics for Hypothesis Definition question, Summer 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.8151	30th	0.8100	-0.0051	0.21
Math	0.8588	22nd	0.8587	-0.0001	0.25
REU	0.5730	6th	0.5983	+0.0253	0.61
External #1	0.5918	30th	0.5781	-0.0137	0.04
Combined	0.7071	31st	0.7039	-0.032	0.22

Table 66b: Item Analysis Statistics for Hypothesis Definition question, Fall 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7563	30th	0.7496	-0.0067	0.06
Math	0.7107	8th	0.7232	+0.0125	0.38
External #1	0.7348	31st	0.7314	-0.0034	0.13
External #2a	0.6397	19th	0.6452	+0.0055	0.27
External #2b	0.5956	29th	0.5843	-0.0113	0.11
External #3	0.5354	17th	0.5424	+0.0070	0.30
DOE	0.5687	28th	0.5623	-0.0064	0.05
Combined	0.7290	29th	0.7252	-0.0038	0.18

Table 66c: Item Analysis Statistics for Hypothesis Definition question, Spring 2004

Course	Alpha-if-deleted	Rank	Overall alpha	Difference (+ is good)	Discriminatory Index
Engr	0.6744	34th	0.6562	-0.0182	-0.12
Math #1	0.7396	15th	0.7460	+0.0064	0.36
Math #2	0.7148	18th	0.7211	+0.0063	0.57
Combined	0.7075	28th	0.7079	+0.0004	0.22

Summer 2003 #27, Fall 2003 #17, Spring 2004 #28

You perform a regression and obtain an R^2 value of 0.75. What does this mean?

a) The slope of the regression line is 0.75
b) The regression model accurately fits 75% of the data points
c) 75% of the variability in the data can be accounted for by the regression model **
d) You can expect 75% accuracy for future predictions using this model

This item was added because simple linear regression is one of the most important topics on the faculty survey (score 3.52 out of 4, rank 4th). The R^2 metric is an important and common tool used to assess the appropriateness of any regression model, although not restricted to simple linear regression. There is concern about the content validity of this item because students are unlikely to encounter regression in an introductory statistics course. Students have likely encountered R^2 in other classes when graphing in Excel. At OU, the topic is usually covered in a College of Engineering course on numerical computation techniques, and students who have taken this class may have an advantage. This could be a problem of the item simply not being fair to all levels of students in an introductory statistics class.

This question is moderately successful on objective metrics. Two courses have high discriminatory indices for the Summer 2003 data and all courses rate positive on alpha-if-deleted. For Fall 2003, it is interesting that the course at a two-year college had the highest discriminatory index and also that the question rated 1st on reliability. The other groups are slightly lacking compared to Summer 2003, but the item could generally be considered acceptable. The Spring 2004 metrics also moderate for the Math courses but poor for the Engr course.

Table 67a: Item Analysis Statistics for Regression question, Summer 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.8177	32nd	0.8100	+0.0032	-0.11
Math	0.8590	23rd	0.8587	+0.0089	0.50
REU	0.5885	14th	0.5983	+0.0475	0.48
External #1	0.5680	11th	0.5781	+0.0113	0.33
Combined	0.7037	29th	0.7039	+0.0002	0.28

Table 67b: Item Analysis Statistics for Regression question, Fall 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7425	19th	0.7496	+0.0071	0.38
Math	0.7194	20th	0.7232	+0.0038	0.10
External #1	0.7237	14th	0.7314	+0.0077	0.39
External #2a	0.6371	15th	0.6452	+0.0081	0.28
External #2b	0.5890	25th	0.5843	-0.0047	0.21
External #3	0.4817	1st	0.5424	+0.0607	0.50
DOE	0.5799	31st	0.5623	-0.0176	0.09
Combined	0.7187	17th	0.7252	+0.0065	0.38

Table 67c: Item Analysis Statistics for Regression question, Spring 2004

Course	Alpha-if-deleted	Rank	Overall alpha	Difference (+ is good)	Discriminatory Index
Engr	0.6688	33rd	0.6562	-0.0126	0.06
Math #1	0.7412	19th	0.7460	+0.0048	0.36
Math #2	0.7072	10th	0.7211	+0.0139	0.37
Combined	0.7022	20th	0.7079	+0.0057	0.41

Summer 2003 #30, Fall 2003 #26, Spring 2004 #2

A chemist wants to determine if there is a relationship between product purity and pressure for a chemical reaction. He regresses pressure on product purity and fails to reject the null hypothesis. This means that:

a) Pressure and product purity are statistically significantly correlated
b) Pressure does not help explain the variation in product purity **
c) There is no linear relationship between pressure and product purity
d) There is no curvilinear relationship between pressure and product purity

This is another question added to see if advanced students would show a difference. There has been one change to answer D so that it would not sound too similar to C. The new choice is the following.

d) Pressure can be used to predict product purity

Unlike the first two advanced questions, this item does rate slightly higher for the DOE class compared to most others. However, the percent correct in the DOE class is close to that for the other classes. This implies that the higher-ranking students are getting it correct rather than answers being based more on guessing.

Table 68a: Item Analysis Statistics for Regression question, Summer 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.8008	12th	0.8100	+0.0032	0.46
Math	0.8742	33rd	0.8587	+0.0089	-0.50
REU	0.5905	18th	0.5983	+0.0475	0.36
External #1	0.6000	33rd	0.5781	+0.0113	0.02
Combined	0.7097	32nd	0.7039	-0.0058	0.19

Table 68b: Item Analysis Statistics for Regression question, Fall 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7533	28th	0.7496	-0.0037	0.19
Math	0.7186	19th	0.7232	+0.0046	0.24
External #1	0.7506	34th	0.7314	-0.0192	-0.23
External #2a	0.6584	33rd	0.6452	-0.0132	0.16
External #2b	0.5802	27th	0.5843	+0.0041	0.08
External #3	0.5461	24th	0.5424	-0.0037	0.40
DOE	0.5533	18th	0.5623	+0.0090	0.36
Combined	0.7328	33rd	0.7252	-0.0079	0.10

Table 68c: Item Analysis Statistics for Regression question, Spring 2004

Course	Alpha-if-deleted	Rank	Overall alpha	Difference (+ is good)	Discriminatory Index
Engr	0.6538	22nd	0.6562	+0.0024	0.17
Math #1	0.7450	22nd	0.7460	+0.0010	0.27
Math #2	0.7048	7th	0.7211	+0.0163	0.60
Combined	0.7002	16th	0.7079	+0.0077	0.26

Summer 2003 #31, Fall 2003 #18, Spring 2004 #29

A scientist takes a set of 50 measurements. The standard deviation is reported as -2.30. Which of the following must be true?

a) Some of the measurements were zero
b) Most of the measurements were negative
c) All of the measurements less than the mean
d) All of the measurements were negative
e) The standard deviation was calculated incorrectly **

This item was added because there was not a question dealing with standard deviation. For Fall 2003, choice A was deleted and it was not replaced with a new choice. It was deleted because no one in Engr or Math picked it on the Summer 2003 post-test. There were no changes for Spring 2004.

This item rates positive on eleven of the thirteen classes from Summer and Fall 2003, and Spring 2004. It is the top question for the Fall 2003 Engr course. The discriminatory index is high for several classes but could use improvement in others.

Table 69a: Item Analysis Statistics for Standard Deviation question, Summer 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.8016	13th	0.8100	+0.0084	0.48
Math	0.8480	7th	0.8587	+0.0107	0.75
REU	0.6117	29th	0.5983	-0.0134	0.04
External #1	0.5501	5th	0.5781	+0.0280	0.49
Combined	0.6910	5th	0.7039	+0.0129	0.42

Table 69b: Item Analysis Statistics for Standard Deviation question, Fall 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7226	1st	0.7496	+0.0270	0.94
Math	0.7143	15th	0.7232	+0.0089	0.36
External #1	0.7237	6th	0.7314	+0.0077	0.39
External #2a	0.6353	14th	0.6452	+0.0099	0.17
External #2b	0.5878	22nd	0.5843	-0.0035	0.02
External #3	0.5414	22nd	0.5424	+0.0010	0.20
DOE	0.5552	21st	0.5623	+0.0071	0.16
Combined	0.7087	4th	0.7252	+0.0165	0.56

Table 69c: Item Analysis Statistics for Standard Deviation question, Spring 2004

Course	Alpha-if-deleted	Rank	Overall alpha	Difference (+ is good)	Discriminatory Index
Engr	0.6285	5th	0.6562	+0.0277	0.45
Math #1	0.7339	11th	0.7460	+0.0121	0.27
Math #2	0.6984	2nd	0.7211	+0.0227	0.79
Combined	0.6874	1st	0.7079	+0.0205	0.48

Summer 2003 #34, Fall 2003 #9, Spring 2004 #26

You have a set of 30 numbers. The standard deviation from these numbers is reported as zero. You can be certain that:

a) Half of the numbers are above the mean
b) All of the numbers in the set are zero
c) A computational error was made
d) All of the numbers in the set are equal **
e) The mean, median, and mode of these numbers are different

This item is similar to the previous item in content. The item was not changed for Fall 2003, but choice C was deleted for Spring 2004 because it was not chosen by anyone in the Fall 2003 Engr and External #1 courses.

Based on objective metrics, this item tends to fare slightly better than the previous item. The alpha-if-deleted is positive for eleven of the thirteen courses. The discriminatory index is more consistent across all courses but lacks the very high values that the previous question had for two classes. Seven classes are in the high range, compared with just four for the previous item.

Table 70a: Item Analysis Statistics for Standard Deviation question, Summer 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.8085	20th	0.8100	+0.0015	0.13
Math	0.8522	14th	0.8587	+0.0065	0.50
REU	0.5688	3rd	0.5983	+0.0295	0.43
External #1	0.5718	17th	0.5781	+0.0063	0.20
Combined	0.6965	14th	0.7039	+0.0074	0.30

Table 70b: Item Analysis Statistics for Standard Deviation question, Fall 2003

Course	Alpha-if-deleted	Rank	Overall Alpha	Difference (+ is good)	Discriminatory Index
Engr	0.7327	5th	0.7496	+0.0169	0.56
Math	0.7274	30th	0.7232	-0.0042	0.02
External #1	0.7225	13th	0.7314	+0.0089	0.31
External #2a	0.6345	13th	0.6452	+0.0107	0.40
External #2b	0.5758	13th	0.5843	+0.0085	0.40
External #3	0.4950	2nd	0.5424	+0.0474	0.70
DOE	0.5458	12th	0.5623	+0.0165	0.30
Combined	0.7111	8th	0.7252	+0.0141	0.46

Table 70c: Item Analysis Statistics for Standard Deviation question, Spring 2004

Course	Alpha-if-deleted	Rank	Overall alpha	Difference (+ is good)	Discriminatory Index
Engr	0.6628	26th	0.6562	-0.0066	0.05
Math #1	0.7404	16th	0.7460	+0.0056	0.36
Math #2	0.6956	1st	0.7211	+0.0255	0.78
Combined	0.6998	15th	0.7079	+0.0081	0.34

Fall 2003 #5, deleted

> You are waiting in line at the bank to cash your weekly paycheck. The bank claims to have an average waiting time of 10 minutes. You have already stood in line for five minutes. How much longer do you expect to wait?
>
> a) 5 minutes
> b) 10 minutes **
> c) 15 minutes
> d) There is no way to estimate

This item was added for comparison with the modem pool problem because it is the same concept. The results were very similar in that almost no one got it correct. Consequently, the discriminatory index is low for most classes. This item was deleted rather than the modem pool problem because the researchers felt this problem was less-obviously memory-less.

Table 71: Item Analysis Statistics for Waiting Time problem, Fall 2003

Course	**Alpha-if-deleted**	**Rank**	**Overall Alpha**	**Difference (+ is good)**	**Discriminatory Index**
Engr	0.7442	21st	0.7496	+0.0054	0.25
Math	missed by all	n/a	0.7232	n/a	0
External #1	0.7163	5th	0.7314	+0.0151	0.46
External #2a	0.6668	34th	0.6452	-0.0216	-0.10
External #2b	0.5779	15th	0.5843	+0.0064	0.11
External #3	0.5574	28th	0.5424	-0.0150	0.00
DOE	0.5627	25th	0.5623	-0.0004	0.11
Combined	0.7260	27th	0.7252	-0.0008	0.12

Spring 2004 #6

> For which of the following samples would you expect to calculate the smallest variance?
>
> a) An olympic sprinter's running times for 15 trials of the 200 meter dash **
> b) A high school track team's running times for the 200 meter dash
> c) An olympic sprinter's running times for 5 trials each of the 100 meter, 200 meter and 400 meter dashes
> d) An olympic track team's running times for the 100 meter, 200 meter, and 400 meter dashes

This item was added for comparison with the graphical variability question. Statistically, it is very poor, with all classes negative on reliability and poor on discrimination. The graphical variability question is not very strong, but it is better than this item. The three classes are around 75% correct, which limits the discriminatory index, but it should still be higher.

Table 72: Item Analysis Statistics for Variability question, Spring 2004

Course	Alpha-if-deleted	Rank	Overall alpha	Difference (+ is good)	Discriminatory Index
Engr	0.6649	29th	0.6562	-0.0087	0.05
Math #1	0.7516	32nd	0.7460	-0.0056	0.00
Math #2	0.7231	31st	0.7211	-0.0020	0.12
Combined	0.7126	31st	0.7079	-0.0047	0.03

Spring 2004 #11

Which of the following statistics are not affected by extreme negative outliers?

a) range
b) 1st quartile **
c) mean
d) variance

This question was also added for increased coverage on variability. This item is good on the objective metrics for the Math courses but poor for the Engr course. This item needs more testing in the Engr course before any conclusions can be made.

Table 73: Item Analysis Statistics for Outlier question, Spring 2004

Course	Alpha-if-deleted	Rank	Overall alpha	Difference (+ is good)	Discriminatory Index
Engr	0.6631	27th	0.6562	-0.0069	0.14
Math #1	0.7313	9th	0.7460	+0.0147	0.36
Math #2	0.7071	9th	0.7211	+0.0140	0.57
Combined	0.6989	13th	0.7079	+0.0090	0.34

Spring 2004 #21

A meteorologist predicts a 40% chance of rain in London and a 70% chance in Chicago. What is the most likely outcome?

a) It rains only in London
b) It rains only in Chicago
c) It rains in London and Chicago
d) It rains in London or Chicago **

This item was added as a replacement for the other problem about chance of rain (Fall 2002 #16). In this item, it should only matter that the students understand a percent chance of rain to be a probability. Students have confusion about chance of rain and sometimes view it as either/or rather than a probability. This item also has more than two viable options. Depending on how students interpret "or," this question varies in difficulty. If "or" is taken to mean just one but not both, the student may need to calculate the actual probabilities, perhaps using a Venn diagram as a guide. If "or" is taken to mean at least one, then choice D should be chosen without hesitation. Choice D is the correct answer under either interpretation.

The objective metrics for the Math courses are in the same range as those for the deleted question, but the Engr course is poor. The percent correct is around 50% on pre-test and post-test for all three courses. It is therefore somewhat questionable if the item is assessing something that is learned in the course.

Table 74: Item Analysis Statistics for Chance of Rain question, Spring 2004

Course	**Alpha-if-deleted**	**Rank**	**Overall alpha**	**Difference (+ is good)**	**Discriminatory Index**
Engr	0.6675	32nd	0.6562	-0.0113	0.10
Math #1	0.7417	20th	0.7460	+0.0043	0.45
Math #2	0.7127	14th	0.7211	+0.0084	0.37
Combined	0.7058	23rd	0.7079	+0.0021	0.41

Spring 2004 #32

An engineer performs a hypothesis test and reports a p-value of 0.03. Based on a significance level of 0.05, what is the correct interpretation?

a) The null hypothesis is true.
b) The alternate hypothesis is true.
c) Do not reject the null hypothesis.
d) Reject the null hypothesis. **
e) Accept the alternate hypothesis.

This is a replacement for the deleted p-value question that was not performing well on alpha-if-deleted and often had a negative gain (Fall 2002 #14). This new item focuses on interpreting p-value rather forcing people to recall a definition. The psychometric analysis is very good for the Math courses but poor for the Engr course, as is the case with several other new items for Spring 2004. It is still an improvement over the old item.

Table 75: Item Analysis Statistics for Interpreting p-value question, Spring 2004

Course	**Alpha-if-deleted**	**Rank**	**Overall alpha**	**Difference (+ is good)**	**Discriminatory Index**
Engr	0.6654	30th	0.6562	-0.0092	0.17
Math #1	0.7255	6th	0.7460	+0.0205	0.64
Math #2	0.7048	7th	0.7211	+0.0163	0.59
Combined	0.6948	8th	0.7079	+0.0131	0.51

Looking at the scores, it is apparent for Engr and Math #1 that the item is also an improvement because the old item had negative gains and usually around 10% correct on the post-test. Oddly, Math #2 is much different from Math #1 even though the professor is the same.

Table 76: Knowledge Gain on Interpreting p-value question, Spring 2004

Course	**Pre-Test % Correct**	**Post-Test % Correct**	**Gain**
Engr	18%	41%	+23%
Math #1	8%	52%	+44%
Math #2	34%	33%	-1%

5.2 Item Analysis Conclusions

The evolution of the items was presented along with the reasoning, whether it be qualitative or quantitative. Nearly every item has undergone at least minor changes, while some have been deleted on the basis of content validity. The instrument as a whole is only as good as its items, and the current status is that most of the items could be considered good. It is apparent that some topics are difficult to capture. For example, the item where students should recall the memory-less property is missed by almost every student in all classes despite the fact that it should be simple if the concept is understood. This raises the question of whether the topic should even be included if all it tells instructors is that students do not understand. Similar concerns exist for items which are answered correctly by nearly every student, although it is more re-assuring on the surface to know what students know. The quest will continue to write meaningful items which meet the exacting standards of validity, reliability, and discrimination which have been presented in this chapter.

CHAPTER V

Preliminary Conclusions

The data presented were drawn from the first four semesters of SCI administrations (Fall 2002, Summer 2003, Fall 2003, Spring 2004). The results indicate that the instrument has improved in terms of validity, reliability, and discriminatory power. The improvements were brought about through careful editing of questions, drawing on objective statistics (e.g., alpha-if-deleted, discriminatory index, answer distributions) and subjective analyses (e.g., focus groups, researchers' experience).

As with the Force Concept Inventory, the ultimate goal is to produce an instrument which is nationally recognized as a useful tool for improving student learning in statistics, both by identifying students' problem areas and providing feedback which professors can use to improve teaching. The progress in this project's first two years provides a solid background for this to happen.

Figure 1 depicts the first phase in the creation process of the SCI. Test Theory and Concept Inventories are considered inputs to the genesis of the project. The Statistics Concept Inventory node leads to Reliability, Validity, and Item Analysis as the families of analysis techniques. It should be acknowledged that some steps occurred before the construction of the instrument, most notably content validity considerations of a faculty survey, textbook reviews, and journal articles. The model also acknowledges the interaction between test-level analyses (Reliability, Validity) and detailed Item Analysis.

These considerations were taken in tandem to arrive at the Spring 2004 version, which changed little throughout the remainder of the project.

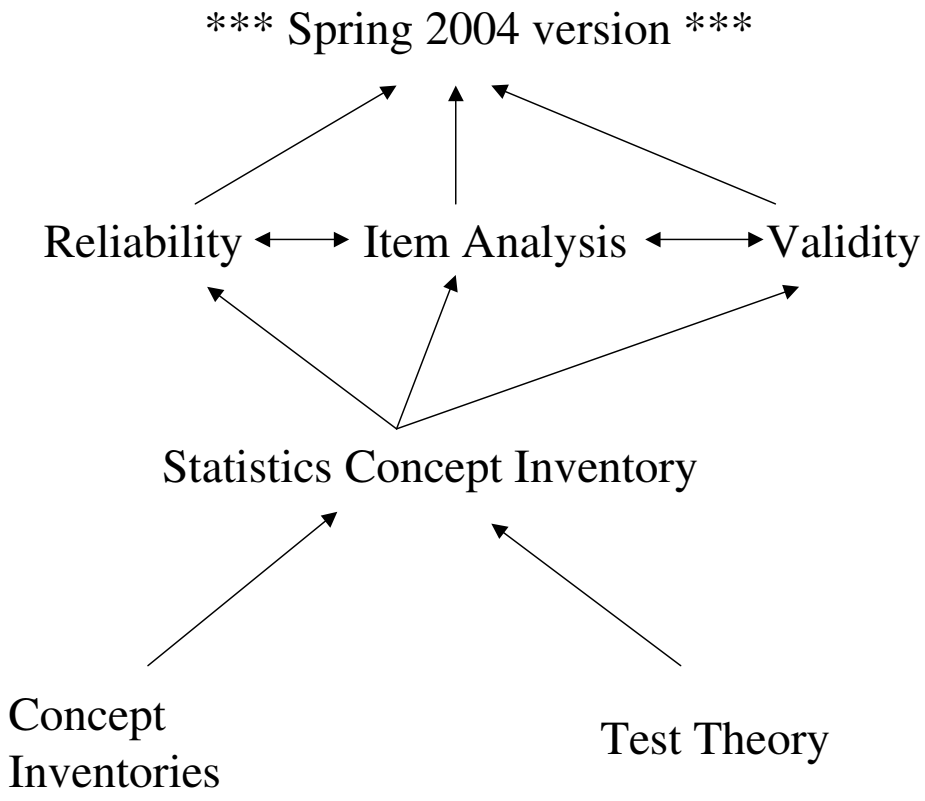


Figure 1: Creation process presented in Book One

This first book of the dissertation was presented as might be a typical Master's thesis, which this document was originally destined to be. The remainder of the dissertation is more compartmentalized, focusing on specific steps of the development such as the online test, assessing misconceptions, and validating the content. The adventure concludes with this author's proposed version of the SCI to serve as a usable classroom tool and leaves open the possibility for continued work.

References

Allen, K. 2004. Explaining Cronbach's Alpha. Available at http://coecs.ou.edu/sci under Publications. A paper is also being prepared with the findings.

Anderson, D.L., K.M. Fisher, and G.J. Norman. 2002. Development and Evaluation of the Conceptual Inventory of Natural Selection. *Journal of Research in Science Teaching*. 39 (10): 952-978.

Ausubel, D. 1968. Educational Psychology: A Cognitive View. Holt, Reinhart, and Winston: New York.

Beichner, R.J. 1994. Testing student interpretation of kinematics graphs. *American Journal of Physics*. 62 (8): 750-755.

Brown, F.G. 1983. *Principles of Educational and Psychological Testing*. Holt, Reinhart, and Winston: New York.

Brown, J.B. 1975. The Number of Alternatives for Optimum Test Reliability. *Journal of Educational Measurement*. 12(2): 109-113.

College Board. 2003. Course Description: Statistics. Retrieved December 18, 2003. http://www.collegeboard.com/prod_downloads/ap/students/statistics/ap03_statistics.pdf.

Cortina, J.M. 1993. What is Coefficient Alpha? An Examination of Theory and Applications. *Journal of Applied Psychology*. 78 (1): 98-104.

Costin, F. 1970. The Optimal Number of Alternatives in Multiple-Choice Achievement Tests: Some Empirical Evidence for a Mathematical Proof. *Educational and Psychological Measurement*. 30: 353-358.

Costin, F. 1972. Three-choice Versues Four-choice Items: Implications for Reliability and Validity of Objective Achievement Tests. *Educational and Psychological Measurement*. 32: 1035-1038.

Cronbach, L.J. 1943. On Estimates of Test Reliability. *Journal of Educational Psychology*. 34: 485-494.

Cronbach, L.J. 1946. A Case Study of the Split-Half Reliability Coefficient. *Journal of Educational Psychology*. 37: 473-480.

Cronbach, L.J. 1947. Test 'Reliability': Its Meaning and Determination. *Psychometrika*. 12 (1): 1-16.

Cronbach, L.J. 1951. Coefficient Alpha and the Internal Structure of Tests. *Psychometrika*. 16 (3): 297-334.

Crouch, C.H., and Mazur, E. 2001. Peer Instruction: Ten years of experience and results. *American Journal of Physics*. 69 (9): 970-977.

Ebel, R. 1954. Procedures for the Analysis of Classroom Tests. *Educational & Psychological Measurement*. 14: 352-364.

Ebel, R.L. 1969. Expected Reliability as a Function of Choices Per Item. *Educational and Psychological Measurement*. 29: 565-570.

Engelhardt, P.V., and R.J. Beichner. 2004. Students' understanding of direct current resitive electrical circuits. *American Journal of Physics*. 72 (1): 98-115.

Engineering Accreditation Commission. 2003. Criteria For Accrediting Engineering Programs. 2004-2005 Criteria. http://www.abet.org/criteria_eac.html.

Engineering Accreditation Commission. 2006. Criteria For Accrediting Engineering Programs, 2006-2007 Criteria. http://www.abet.org/forms.shtml, Accessed March 16, 2006.

Evans, D.L., G.L. Gray, S. Krause, J. Martin, C. Midkiff, B.M. Notaros, M. Pavelich, D. Rancour, T.R. Rhoads, P. Steif, R.A. Streveler, and K. Wage. 2003. Progress On Concept Inventory Assessment Tools. *Proceedings of the 33rd ASEE/IEEE Frontiers in Education Conference*, Session T4G-8.

Flanagan, J.C. 1937. A proposed procedure for increasing the efficiency of objective tests. *Journal of Educational Psychology*. 28: 17-21.

Garfield, J. and A. Ahlgren. 1988. Difficulties in Learning Basic Concepts in Probability and Statistics: Implications for Research. *Journal for Research in Mathematics Education*. 19 (1): 44-63.

Gary, G.L., D. Evans, P. Cornwell, F. Costanzo, and B. Self. 2003. Toward a Nationwide Dynamics Concept Inventory Assessment Test. *Proceedings of the 2003 American Society for Engineering Education Annual Conference & Exposition*. Session 1168.

Gibb, B. 1964. *Test-Wiseness as Secondary Cue Response*. Dissertation, Stanford University.

Guttman, L. 1945. A Basis for Analyzing Test-Retest Reliability. *Psychometrika*. 10 (4): 255-282.

Hake, R. 1998. Interactive-engagement vs. traditional methods: A six-thousand student survey of mechanics test data for introductory physics courses. *American Journal of Physics*. 6 (1): 64-75.

Halloun, I. and D. Hestenes. 1985. The initial knowledge state of college physics students. *American Journal of Physics*. 53 (11): 1043-1055.

Heller, P., and D. Huffman. 1995. Interpreting the Force Concept Inventory: A Reply to Hestenes and Halloun. *The Physics Teacher*. 33 (November): 503-511.

Hestenes, D., and I. Halloun. 1995. Interpreting the Force Concept Inventory: A Response to March 1995 Critique by Huffman and Heller. *The Physics Teacher*. 33 (November): 502-506.

Hestenes, D., and M. Wells. 1992. A Mechanics Baseline Test. *The Physics Teacher*. 30 (March): 159-166.

Hestenes, D., M. Wells, and G. Swackhamer. 1992. Force Concept Inventory. *The Physics Teacher*. 30 (March): 141-158.

Hirsch, L.S., and A.M O'Donnell, 2001. Representativeness in statistical reasoning: Identifying and assessing misconceptions. *Journal of Statistics Education, 9*(2).

Hopkins, K.D., J.C. Stanley, and B.R. Hopkins. 1990. *Educational and Psychological Measurement and Evaluation, 7th Edition*. Prentice Hall: Englewood Cliffs, NJ.

Huffman, D., and P. Heller. 1995. What Does the Force Concept Inventory Actually Measure?. *The Physics Teacher*. 33 (March): 138-143.

Jacobi, A., J. Martin, J. Mitchell, and T. Newell, 2003. A Concept Inventory for Heat Transfer. *Proceedings of the 33rd ASEE/IEEE Frontiers in Education Conference*. Session T3D-6.

Jordan, W., H. Cardenas, and C.B. O'Neal. 2005. Using a Materials Concept Inventory to Assess an Introductory Materials Class: Potentials and Problems. *Proceedings of the 2005 American Society for Engineering Education Annual Conference and Exposition*. Session 1064.

Kahneman, D. and A. Tversky. 1972. Subjective Probability: A Judgment of Representativeness. *Cognitive Psychology*. 3 (3): 430-453.

Kelley, T. 1939. The Selection of Upper and Lower Groups for the Validation of Test Items. *Journal of Educational Psychology*. 30: 17-24.

Kim, E., and Pak, S.-J. 2000. Students do not overcome conceptual difficulties after solving 1000 traditional problems. *American Journal of Physics*. 70 (7): 759-765.

Kline, P. 1986. A Handbook of Test Construction. Methuen & Co. Ltd: New York.

Kline, P. 1993. The Handbook of Psychological Testing. Routledge: London and New York.

Konold, C., 1995. Issues in Assessing Conceptual Understanding in Probability and Statistics. *Journal of Statistics Education*. 3 (1), online.

Konold, C., A. Pollatsek, A. Well, J. Lohmeier, and A. Lipson. 1993. Inconsistencies in Students' Reasoning About Probability. *Journal for Research in Mathematics Education*. 24 (5): 392-414.

Krause, S., J.C. Decker, and R. Griffin. 2003. Using a Materials Concept Inventory to Assess Conceptual Gain in Introductory Materials Engineering Courses. *Proceedings of the 33rd ASEE/IEEE Frontiers in Education Conference*. Session T3D-6.

Krause, S., J. Birk, R. Bauer, B. Jenkins, M.J. Pavelich. 2004a. Development, Testing, and Application of a Chemistry Concept Inventory. *Proceedings of the 34th ASEE/IEEE Frontiers in Education Conference*. Session T1G-1.

Krause, S., Tasooji, A., and Griffin, R. 2004b. Origins of Misconceptions in a Materials Concept Inventory From Student Focus Groups. *Proceedings of the 2004 American Society for Engineering Education Annual Conference & Exposition.*

Kuder, G.F., and M.W. Richardson. 1937. The Theory of the Estimation of Test Reliability. *Psychometrika*. 2 (3): 151-160.

Libarkin, J.C., and S.W. Anderson. 2005. Assessment of Learning in Entry-Level Geoscience Courses: Results from the Geoscience Concept Inventory. *Journal of Geoscience Education*. 53 (4, September): 394-401.

Lord, F.M. 1977. Optimal Number of Choices per Item – A Comparison of Four Approaches. *Journal of Educational Measurement*. 14(1): 33-38.

Maloney, D.P., T.L. O'Kuma, C.J. Hieggelke, and A.V. Heuvelen. 2000. Surveying students' conceptual knowledge of electricity and magnetism. *American Journal of Physics*. 69 (7) : S12 – S23.

Martin, J., J. Mitchell, and T. Newell. 2003. Development of a Concept Inventory for Fluid Mechanics. *Proceedings of the 33rd ASEE/IEEE Frontiers in Education Conference*.Session T3D-6.

Martin, J. K., J. Mitchell, and T. Newell. 2004. Analysis of Reliability of the Fluid Mechanics Concept Inventory. *Proceedings of the 34th ASEE/IEEE Frontiers in Education Conference*. Session T1A-1.

Michel, H., J. Jackson, P. Fortier, and H.Liu. Computer Engineering Concept Inventory. http://www.foundationcoalition.org/home/keycomponents/concept/computer.html, Accessed April 17, 2006.

Midkiff, K.C., T.A. Litzinger, and D.L. Evans. 2001. Development of Engineering Thermodynamics Concept Inventory Instruments. *Proceedings of the 31st ASEE/IEEE Frontiers in Education Conference*. Session F2A-3.

Montgomery, D. and G. Runger. 1994. Applied Statistics and Probability for Engineers. Wiley: New York.

Moore, D. 1997. The Active Practice of Statistics. W. H. Freeman and Company: New York.

Morgan, J., and J. Richardson. Strength of Materials (SOM) Concept Inventory. http://www.foundationcoalition.org/home/keycomponents/concept/strength.html, Available in either pdf or ppt, Accessed September 8, 2005.

Notaros, B. Electromagnetics Concept Inventory. http://www.foundationcoalition.org/home/keycomponents/concept/electromagnetics.html, Accessed April 17, 2006.

Novick, M.R., and C. Lewis. Coefficient Alpha and the Reliability of Composite Measurements. *Psychometrika*. 32 (1): 1-13.

Nunnally, J. 1978. Psychometric Theory. McGraw-Hill: New York.

Olds, B.M., R. Streveler, R.L. Miller, and M.A. Nelson. 2004. Preliminary Results from the Development of a Concept Inventory in Thermal and Transport Science. *Proceedings of the 2004 American Society for Engineering Education Annual Conference & Exposition*. Session 3230.

Oosterhof, A. 1996. Developing and Using Classroom Assessment. Merrill / Prentice Hall: Englewood Cliffs, New Jersey.

Pavelich, M., B. Jenkins, J. Birk, R. Bauer, and S. Krause. 2004. Development of a Chemistry Concept Inventory for Use in Chemistry, Materials and other Engineering Courses. (Draft). *Proceedings of the 2004 American Society for Engineering Education Annual Conference & Exposition*.

Pollatsek, A., S. Lima, and A.D. Well. 1981. Concept or Computation: Students' Understanding of the Mean. *Educational Studies in Mathematics*. 12: 191-204.

Ramos, R.A., and Stern, J. 1973. Item Behavior Associated with Changes in the Number of Alternatives in Multiple Choice Items. *Journal of Educational Measurement*. 10(4): 305-310.

Rebello, N.S., and Zollman, D.A. 2004. The effect of distracters on student performance on the force concept inventory. *American Journal of Physics*. 72 (1, January): 116-125.

Richardson, J., P. Steif, J. Morgan, and J. Dantzler. 2003. Development Of A Concept Inventory For Strength Of Materials. *Proceedings of 33rd ASEE/IEEE Frontiers in Education Conference*. Session T3D-29.

Rhoads, T.R., and R.J. Roedel. 1999. The Wave Concept Inventory - A Cognitive Instrument Based on Bloom's Taxonomy. *Proceedings of the 29th ASEE/IEEE Frontiers in Education Conference*. Session 13c1. Paper 13c1-14.

Roedel, R.J., S. El-Ghazaly, T.R. Rhoads, and E. El-Sharawy. 1998. The Wave Concepts Inventory – An Assessment Tool for Courses in Electromagnetic Engineering. *Frontiers In Education Conference Proceedings 1998*.

Rogers, W.T., and Harley, D. 1999. An Empirical Comparison of Three- and Four-Choice Items and Tests: Susceptibility to Testwisness and Internal Consistency Reliability. *Educational and Psychological Measurement*. 59(2): 234-247.

Rulon, P.J. 1939. A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review*. 9: 99-103.

Schau, C., T.L. Dauphinee, A. Del Vecchio and J.J. Stevens. Surveys of Attitudes Toward Statistics. [http://www.unm.edu/~cshau/downloadsats.pdf, accessed October 2, 2002]

Simoni, M.F., M.E. Herniter, and B.A. Ferguson. 2004. Concepts to Questions: Creating an Electronics Concept Inventory Exam. *Proceedings of the 2004 American Society for Engineering Education Annual Conference & Exposition*. Session 1793.

Steif, P. 2003. Comparison Between Performance on a Concept Inventory and Solving Multifaceted Problems. *Proceedings of the 33rd ASEE/IEEE Frontiers in Education Conference*. Session T3D-6.

Steif, P. 2004. Initial Data From A Statics Concept Inventory. *Proceedings of the 2004 American Society for Engineering Education Annual Conference & Exposition*. Session 1368.

Steif, P.S., and J.A. Dantzler. 2005. A Statics Concept Inventory: Development and Psychometric Analysis. *Journal of Engineering Education*. 33: 363-371.

Steif, P.S., and M. Hansen. 2006. Comparisons Between Performance in a Statics Concept Inventory and Course Examinations. *International Journal of Engineering Education*. (in press) [http://www.me.cmu.edu/people/faculty/steif/educationalresearch.htm, accessed February 8, 2006]

Stone, A., K. Allen, T.R. Rhoads, T.J. Murphy, R.L. Shehab, C. Saha. 2003. The Statistics Concept Inventory: A Pilot Study. *Proceedings of the 33rd ASEE/IEEE Frontiers in Education Conference*. Session T3D-6.

Streiner, D. L. 2003. Starting at the Beginning: An Introduction to Coefficient Alpha and Internal Consistency. *Journal of Personality Assessment*. 80 (1): 99-103.

Thompson, B., and X. Fan, 2003. Confidence Intervals About Score Reliability Coefficients. Chapter 5 in Score Reliability. B. Thompson editor. Sage Publications: Thousand Oaks, CA.

Thorndike, R.L. 1982. Applied Psychometrics. Houghton Mifflin: Boston.
Thornton, R.K., and D.R. Sokoloff. 1998. Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation and the Evaluation of Active Learning Laboratory. *American Journal of Physics*. 66(4): 338-352.

Voska, K. W., and H.W. Heikkinen, 2000. Identification and Analysis of Student Conceptions Used to Solve Chemical Equilibrium Problems. *Journal of Research in Science Teaching*. 37: 160-176.

Wage, K.E., J.R. Buck, T.B. Welch, and C.H.G. Wright. 2002. Testing and Validation of the Signals and Systems Concept Inventory. *Proceedings of the 2nd IEEE Signal Processing Education Workshop*. Session 4.6: pp. 1-6.

Wage, K.E., J.R. Buck, C.H.G. Wright, and T.B. Welch. 2005. The Signals and Systems Concept Inventory. *IEEE Transactions on Education*. 48 (3): 448-461.

Book Two

Table of Contents

List of Tables 194
List of Figures 195

VI Assessing and Improving Test Reliability 197

Abstract 197
1. Introduction 197
2. Reliability Background 199
3. About the data 200
    3.1 The Statistics Concept Inventory 200
    3.2 Data Collection 201
4. The Behavior of alpha: Empirical Data 202
    4.1 Theoretical considerations 202
    4.2 Macro view of alpha 203
    4.3 Micro view of alpha 204
    4.4 Explanation using comments from focus groups 208
    4.5 The big picture 210
5. The Behavior of alpha: Simulated Data 211
    5.1 Simulation parameters 211
    5.2 Simulation results 212
6. Conclusion 213

References 213
Acknowledgements 214
Appendix: Derivation of variance for dichotomous scoring 215

VII Online test 216

Abstract 216
1. Background 216
    1.1 Clarification 221
    1.2 Implications 221
2. Server 221
3. Database Architecture 222
    3.1 Student table 222
    3.2 Administrator table 223
    3.3 Questions table 223
    3.4 Answers table 223
    3.5 Order table 224
4. Web Interface 224
    4.1 The test 224
    4.2 Administrator 232
    4.3 High-level administrator 236
    4.4 Guest login 238

5. Online Analysis....240
5.1 Participation rate....240
5.2 Demographics....241
5.3 Reliability....242
5.4 Overall scores....244
5.5 Item scores....246
5.6 Completion time....252
5.7 Order effects....255
5.8 Student feedback....259
5.9 Benefits of the online test....260
6. Preliminary Conclusions....260
7. A Controlled Study....260
7.1 Participation rate....261
7.2 Demographics....261
7.3 Reliability....262
7.4 Overall Scores....263
7.5 Item Scores....263
7.6 Completion time....265
8. Synthesis and Comparison....267
8.1 Corollary: The Problem with Educational Research....268
References....270

VIII Self Efficacy of Statistical Reasoning Skills....271

Abstract....271
A. Literature review....271
1. Difficulties....271
2. Attitudes....271
3. Cognitive Ability: Probability....275
4. Cognitive Ability: Statistics....281
5. Teaching Strategies....292
B. SCI confidence analysis....295
1. Method....295
1.1 Motivation....295
1.2 Data collection....295
1.3 Results sample....296
1.4 Research question....298
2. Results Summary....298
2.1 Reliability....298
2.2 Results – macro....299
2.3 Over- / Under-Confidence: A Statistical Basis....305
3. Results: Over-confidence....309
4. Results: Under-confidence....314
5. Further comparisons to previous work....319
6. Conclusions....328
References....329

List of Tables

Chapter VI

Table 1: Macro view of alpha 203
Table 2: Correlation of alpha-if-deleted with various item metrics 208
Table 3: Ten worst questions in terms of "alpha if item deleted" 209
Table 4: Coefficient Alpha for the SCI across six semesters 210
Table 5: Coefficient Alpha (with Post-Test components) for the SCI 210
Table 6: Theoretical Probabilities for MatLab simulation of Alpha behavior 211
Table 7: Theoretical Probabilities for Five Additional Questions in Simulation 212
Table 8: Results of MatLab Simulation on the Behavior of Alpha 212

Chapter VII

Table 1: "Student" table fields 222
Table 2: "Admin" table fields 223
Table 3: "Questions" table fields 223
Table 4: "Answers" table fields 224
Table 5: "Order" table fields 224
Table 6: Summary of online participation rates 240
Table 7: Participation rates for paper administration at one university 241
Table 8: Paper vs. online demographics 241
Table 9: Online vs. paper reliability (α), all students 242
Table 10: Online vs. paper reliability (α), Engineering majors 243
Table 11: Scaled-up sub-test reliabilities, all students 244
Table 12: Scaled-up sub-test reliabilities, Engineering majors 244
Table 13: Online vs. paper scores, all students 245
Table 14: Online vs. paper scores, Engineering majors 245
Table 15: Online vs. paper standard deviation, all students 245
Table 16: Online vs. paper standard deviation, Engineering majors 245
Table 17: Online vs. paper statistical tests, all students 246
Table 18: Online vs. paper statistical tests, Engineering majors 246
Table 19: Summary statistics for difference in item difficulty 247
Table 20: Sample of a contingency table for response pattern comparisons 250
Table 21: Results of degree of association tests 251
Table 22: Number of significant differences in tests of association 252
Table 23: Summary statistics for online completion time (minutes) 252
Table 24: Participation, Spring 2006 261
Table 25: Participant demographics, part 1 262
Table 26: Participant demographics, part 2 262
Table 27: Reliability of online and paper versions of the SCI, with significance test 262
Table 28: Reliability of online and paper versions of the SCI, scaled-up to $k = 38$ 262
Table 29: Mean percent correct, across version 263
Table 30: Standard deviation of percent correct, across version 263
Table 31: Significance tests for differences between versions 263

Table 32: Results of degree of association tests 265
Table 33: Summary of significant differences across semesters 267

Chapter VIII

Table 1: Difficulty and Abstractness of statistics and other educational topics 272
Table 2: Reliability of confidence scale.....298
Table 3: Summary statistics for average confidence ratings across items.....299
Table 4: Confidence of correct answers.....300
Table 5: Summary statistics for average confidence ratings across items.....300
Table 6: Items in the over-confidence and under-confidence regions.....304
Table 7: Item counts by topic area and confidence region 304
Table 8: Probabilities of region dominance.....305
Table 9: Confidence Band confidence classifications 307

List of Figures

Chapter VI

Figure 1: Relationship between change in alpha and average inter-item correlation.....205
Figure 2: Relationship between average inter-item correlation and gap 206
Figure 3: Relationship between change in alpha and gap.....206
Figure 4: Relationship between change in alpha and discriminatory index 207

Chapter VII

Figure 1: Instructional email received by a student.....225
Figure 2: Login screen 225
Figure 3: Successful login.....225
Figures 4: Informed Consent Form.....226
Figure 5: Demographic questionnaire.....228
Figures 6: Question display 229
Figure 7: Feedback form after completing test (optional) 230
Figure 8: Results email (snippet) 231
Figure 9: Administrator main page.....232
Figures 10: Adding a student 233
Figure 11: Topic area selection.....235
Figure 12: Listing questions.....237
Figure 13: Administrator management page (snippet) 238
Figure 14: Guest account main menu 239
Figure 15: Sample of how questions are displayed 239
Figures 16: Histogram of item difficulty differences 247

Figures 17: Online vs. Paper fraction correct 249
Figure 18: Histogram of completion time (minutes) 253
Figures 19: Number correct vs. completion time.....254
Figure 20: Sample histogram of question counts at position 1 255
Figure 21: Sample histogram of order counts for question #1.....256
Figure 22: Fraction correct versus order position 257
Figure 23: Mean item confidence versus order position.....258
Figure 24: Mean item confidence versus order position (full confidence scale) 259
Figure 25: Online vs. Paper fraction correct 264
Figure 26: Score vs. Time, rank-order 266
Figure 27: The Problem with Educational Research 269

Chapter VIII

Figure 1: Relationship between Difficulty and Abstractness 273
Figure 2: Marble rolling apparatus 276
Figure 3: A typical family ("triples") of trinomial distributions.....282
Figure 4: Sample SCI item with confidence rating scale.....296
Figures 5: Confidence graph sample.....297
Figure 6a: Frequency distribution of average item confidence 299
Figure 6b: Frequency distribution of average item confidence, by Incorrect and Correct.....301
Figures 7: Confidence vs. fraction correct 302
Figures 8: Confidence Bands 308
Figures 9: Confidence profiles, coin flipping question.....310
Figures 10: Confidence profiles, alternative hypothesis question 311
Figures 11: Confidence profiles, hospital question.....312
Figures 12: Confidence profiles, waiting time question 314
Figures 13: Confidence profiles, correlation items.....317
Figures 14: Confidence profiles, p-value question 319
Figures 15: Confidence profiles, GPA question 321
Figures 16: Confidence profiles, card sequence question.....322
Figures 17: Confidence profiles, height question 324
Figures 18: Confidence profiles, rain question 325
Figures 19: Confidence profiles, dice rolling question.....327

CHAPTER VI

Assessing and Improving Test Reliability: An Engineer's Perspective

Abstract

Reliability is a fundamental concept of test construction. The most common measure of reliability, Cronbach's coefficient alpha, is frequently used without an understanding of how it behaves. The findings of this article indicate that, in general, questions have a low reliability (in terms of "alpha if item deleted") when students who answer correctly have lower overall scores than students who answer incorrectly. This is quantified by the "gap" between these students' overall scores. This is also shown to be highly-positively correlated with each question's average inter-item correlation and discriminatory index. Possible causes include poorly written questions (e.g., the correct answer looks different from the incorrect answers), questions where students must guess (e.g., the topic is too advanced), and questions where recalling a definition is crucial. Scores and focus group comments from the Statistics Concept Inventory (SCI) are used to make these judgments.

1. Introduction

The concept of test reliability is a cornerstone of test analysis. There are several methods for assessing reliability. The most commonly cited are test-retest, in which answer consistency is measured from one administration to the next; alternative forms, where subjects take two separate tests which are nearly identical in every aspect; and internal consistency, which is related to inter-item correlations and measures the extent to which the test questions are highly correlated with each other. This article investigates internal consistency as measured by coefficient alpha.

There have been several attempts in recent years to shed light on coefficient alpha (Cortina, 1993; Streiner, 2003). While informative, these are often written from theoretical viewpoints. Little work has been presented which details how alpha behaves for real test data on a question-by-question basis. This article presents the background information for coefficient alpha and utilizes data from an instrument, the Statistics Concept Inventory (Stone, *et al.,* 2003), as an illustration of how alpha behaves.

The concept inventory movement was spurred by the development and successful implementation of the Force Concept Inventory (FCI) (Halloun and Heststnes, 1985; Hestenes, *et al.,* 1992). The FCI was developed as a pre-post test to identify student misconceptions of Newtonian force when entering a physics course and check for gains upon completing the course. After many rounds of testing, it was discovered that students gain the most conceptual knowledge in interactive engagement courses, as opposed to traditional lectures (Hake, 1998). The success of the FCI prompted researchers to develop instruments in other fields. In light of recent ABET accreditation standards, which focus on outcomes rather than simply fulfilling seat time requirements, many engineering topics have begun development concept inventories (Evans, *et al.,* 2003).

Many engineering concept inventories are in their early development phase, and discussions with authors indicate a lack of understanding about test reliability. This article sheds light on test reliability in a practical manner such that it can be understood and applied by those with little knowledge of psychometrics.

2. Reliability Background

One of the first attempts to quantify test reliability was formulated by Kuder and Richardson (1937). They comment that a reliability coefficient based on test-retest will often result in a reliability that is too high due to material remembered on the second administration. Further, increasing the time between administrations is impractical because subjects may gain knowledge in the interim.

Kuder and Richardson focus on the concept of a split-half coefficient, in which the test is split in two parts and a correlation is calculated between those two parts. A test of length k has $({}_kC_{k/2}) \div 2$ ways to be split in two; the term is divided by 2 to remove redundancy. For a test with 10 items, there are 126 combinations. Each split-half will result in a different reliability. There are potential problems with deciding how to split the test and which is the most appropriate split. Depending on the split, the calculated reliability may be higher or lower than the "true" reliability. To overcome these problems, the authors derive several equations which arrive at unique values of the reliability coefficient.

The most often-cited result from Kuder and Richardson (KR) is their equation 20, sometimes called KR-20 and later dubbed "alpha" by Cronbach (1951). The KR-20 is so commonly used because it assumes dichotomous scoring (i.e., 0 for incorrect, 1 for correct), which is how most achievement tests are scored. The formula is given in equation (1). The expression $\sum p_i q_i$ can be substituted in place of $k\overline{pq}$ to give a more general result.

$$r_{tt} = \frac{k}{k-1}\frac{\sigma_t^{\,2} - k\overline{pq}}{\sigma_t^{\,2}} \tag{1}$$

where: r_{tt} is the reliability of the test
k is the number of questions
σ_t^2 is the total score variance for the test
p_i is the proportion of students who answer each item correctly
q_i is the proportion of students who answer each item incorrectly
$\overline{pq}$ is the average p multiplied by the average q for the test, equivalent to assuming p is constant across all items.

Equation 1 was generalized by Cronbach (1951) as shown in equation 2, which allows any equally-weighted scoring method for test items. Although commonly referred to as Cronbach's alpha or coefficient alpha, this expression was derived independently by Guttman (1945) as well and is sometimes referred to as Guttman-Cronbach alpha in psychometric literature.

$$\alpha = \frac{k}{k-1}\left(\frac{\sigma_t^{\,2} - \Sigma V_i}{\sigma_t^{\,2}}\right) = \frac{k}{k-1}\left(1 - \frac{\Sigma V_i}{\sigma_t^{\,2}}\right) \tag{2}$$

where: α is Cronbach's coefficient alpha (same meaning as r_{tt})
k is the number of questions on the test
$\sum V_i$ is the sum of the individual item variances
σ_t^2 is the total score variance for the test (denoted as V_t by Cronbach)

For dichotomously scored items, V_i reduces to $p_i q_i$ and the KR-20 equation is obtained. The derivation of this relationship is given in an Appendix.

3. About The Data

3.1 The Statistics Concepts Inventory

The illustrative data were obtained using the Statistics Concepts Inventory (SCI). The SCI is a multiple choice instrument developed to assess student understanding of fundamental statistics concepts. The test was piloted during the Fall 2002 semester at the

University of Oklahoma (OU) (Stone, *et al.,* 2003). The pilot version was constructed by first identifying topics to include using a modified Delphi approach. Questions and multiple-choice answers were written by searching statistics textbooks and educational literature for examples which covered these topics. The researchers also used personal experience to develop additional questions.

The revision process included focus groups, analysis of correct and incorrect answer distributions, and expert opinions. Several new questions were generated as well. The data in this article were gathered from the second version of the SCI, which has 33 questions and was administered during summer 2003. Two sample questions from the test are given below:

1. The following are temperatures for a week in August: 94, 93, 98, 101, 98, 96, and 93. By how much could the highest temperature increase without changing the median?

   a) Increase by 8°
   b) Increase by 2°
   c) It can increase by any amount (*correct*)
   d) It cannot increase without changing the median.

2. A researcher performs a t-test to test the following hypotheses:

   $$H_0 : \mu \leq \mu_0$$
   $$H_1 : \mu > \mu_0$$

   He rejects the null hypothesis and reports a p-value of 0.10. Which of the following must be correct?

   a) The test statistic fell within the rejection region at the significance level
   b) The power of the test statistic used was 90%
   c) Assuming the null is true, there is a 10% possibility that the observed value is due to chance (*correct*)
   d) The probability that the null hypothesis is not true is 0.10

3.2 Data Collection

The data in this study were gathered from four sources: 1) a statistics class in the College of Engineering at OU, with students having a background of at least three semesters of Calculus; 2) a statistics class in the Department of Mathematics at OU, primarily consisting of engineering students with a similar background as (1); 3) two

groups of undergraduates participating in a summer research program in OU's School of Industrial Engineering, with backgrounds ranging from no statistics experience to several semesters of statistics; 4) a statistics class in the College of Engineering at a four-year university outside OU, with a similar background to (1) and (2). The number of students in each group ranged from 14 to 39. Groups (1) and (2) took the instrument as a pre- and post-test. The data in this article are from the post-test.

4. The Behavior Of Alpha: Empirical Data

4.1 Theoretical Considerations

Statistical packages such as SPSS™ and SAS™ report "alpha if item deleted" which shows how alpha would change if a certain question were omitted. A "good" question will have a lower "alpha if item deleted" because deleting that question will lower the overall alpha.

The simplest way to explain how a question will have a negative effect on alpha (i.e, higher "alpha if item deleted") is to consider Cronbach's definition of alpha (equation 2). A "bad" question will lower the overall test variance (σ_t^2). This happens when students with low overall scores perform better on a question than students with high overall scores. This "squishes" the class together (smaller variance). When σ_t^2 decreases, the ratio $\sum V_i / \sigma_t^2$ increases. This ratio is subtracted from 1, which lowers alpha.

For each item, the effect on total score variance is estimated by subtracting the average total score of those who answer the item incorrectly from the average total score of those who answer correctly. We call this value the "gap."

The average inter-item correlation is also considered a measure of a question's reliability (Kuder and Richardson, 1937). The inter-item correlation is the Pearson

correlation coefficient (r) computed with the 0-1 scores for a pair of items. The average inter-item correlation is each item's average inter-item correlation with the other *k-1* items. If a question has negative or low positive (close to zero) inter-item correlations, it does not "fit" with the rest of the questions. This will be shown to relate to which students are answering a question correctly.

It is even possible for the overall alpha to be negative. For example, if every student received the same total score on a test, σ_t^2 would be zero. As the test variance approaches zero, the ratio $\sum V_i / \sigma_t^2$ approaches infinity. When the calculation 1 - $\sum V_i / \sigma_t^2$ is performed, alpha will approach negative infinity. (Note: $\sum V_i$ would only be zero also if every question were answered either correctly or incorrectly by every student.)

4.2 A macro view of alpha

To understand the behavior of coefficient alpha, the components of Cronbach's formula need to be analyzed first. Table 1 shows how alpha and its components vary across the four groups used in this study:

Table 1: Macro view of alpha

Group	**n (students)**	**k (items)**	**Overall α**	σ_t^2	$\sum V_i$	**Range***
OU Math	14	33	0.8587	38.06	6.37	21
OU Engr	24	33	0.8100	31.16	6.68	15
OU REU	27	33	0.5983	14.99	6.29	16
Outside	39	33	0.5781	14.69	6.46	17

* Range is the maximum score minus the minimum score

With the sum of individual variances ($\sum V_i$) approximately constant over the four groups, total test variance (σ_t^2) is seen as the most important component of alpha. For these groups, alpha varies inversely with the number of students for this data, but this is a coincidence when viewing the magnitude of the changes (going down the chart, alpha decreases by 0.05 then 0.21, but n increases by 10 then just 3). This pattern is not seen on

data from subsequent administrations (refer to Table 4 near the end for more data). The range is included as a simplified estimate of variance, but it lacks explanatory power for alpha aside from the highest alpha having the highest range.

4.3 A micro view of alpha

The data used in this section are for the summer research students (group (3) in section 3.2). They were selected for further illustration because a focus group was conducted with over half of these students, which allowed additional insight to be gained about why questions may be performing poorly in terms of reliability. Because the data for the other three groups are similar, their presentation would not add to the discussion or change the result except to reinforce the generalities of the data from group (3).

Each question's effect on alpha is measured by the change in alpha, found by subtracting "alpha if item deleted" (as reported by SPSS™) from the overall alpha. A "good" question will have a lower "alpha if item deleted" because removing that question would lower alpha; thus, the change in alpha will be positive.

The most direct way to explain alpha is to first compare the change in alpha to the average inter-item correlation for each question as shown in Figure 1.

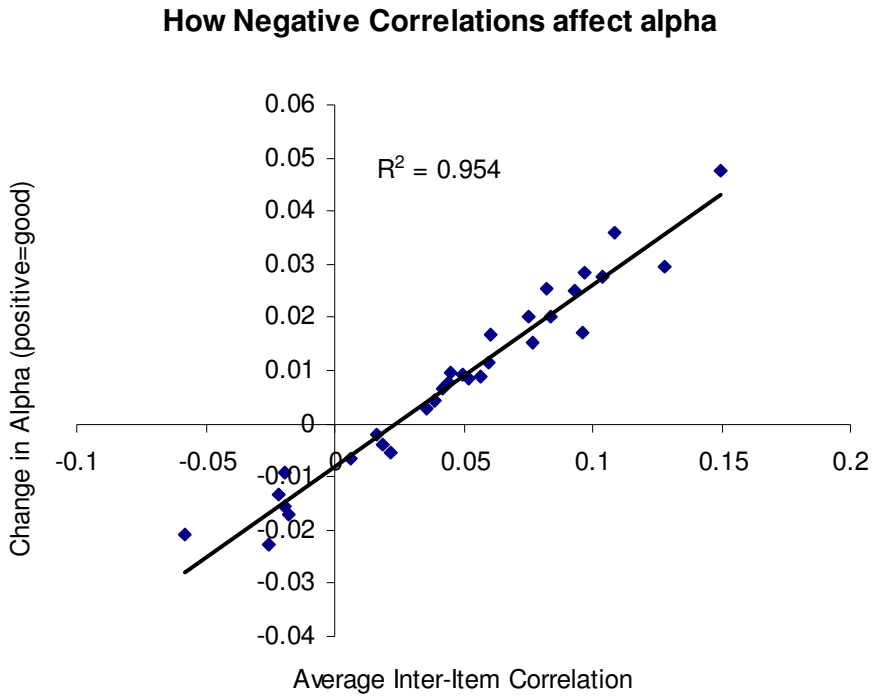


Figure 1: Relationship between change in alpha and average inter-item correlation

Because both axes represent measures of a test's reliability, a strong correlation is expected. It is also important to show why certain questions have poor correlations. This is presented in terms of average inter-item correlation vs. gap and change in alpha vs. gap, shown in Figures 2 and 3, respectively. Gap is calculated using total score rather than percentage. Using percentage will change the scale of the x-axis, but the correlation will not change. On this 33 question test, one point of gap corresponds to three percent.

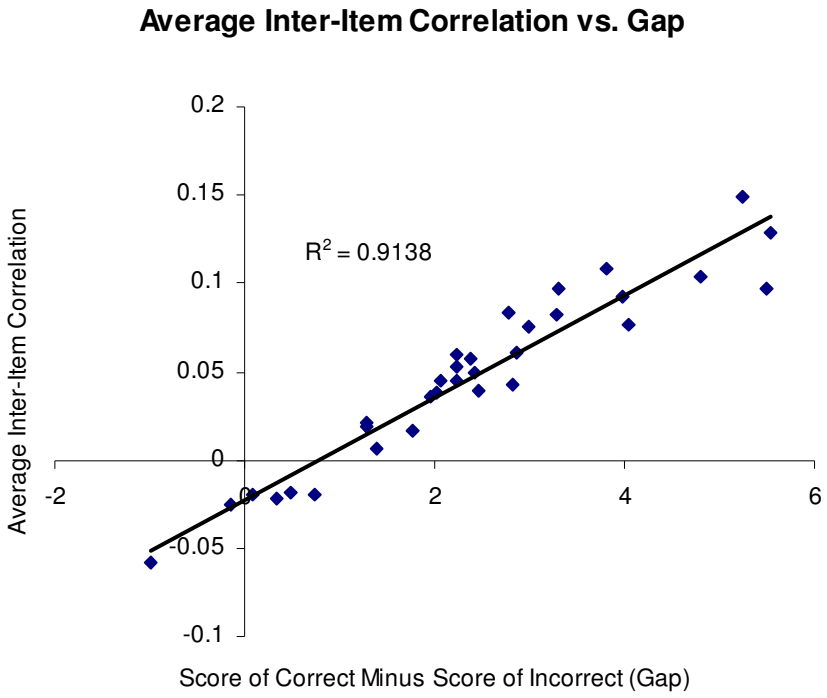


Figure 2: Relationship between average inter-item correlation and gap

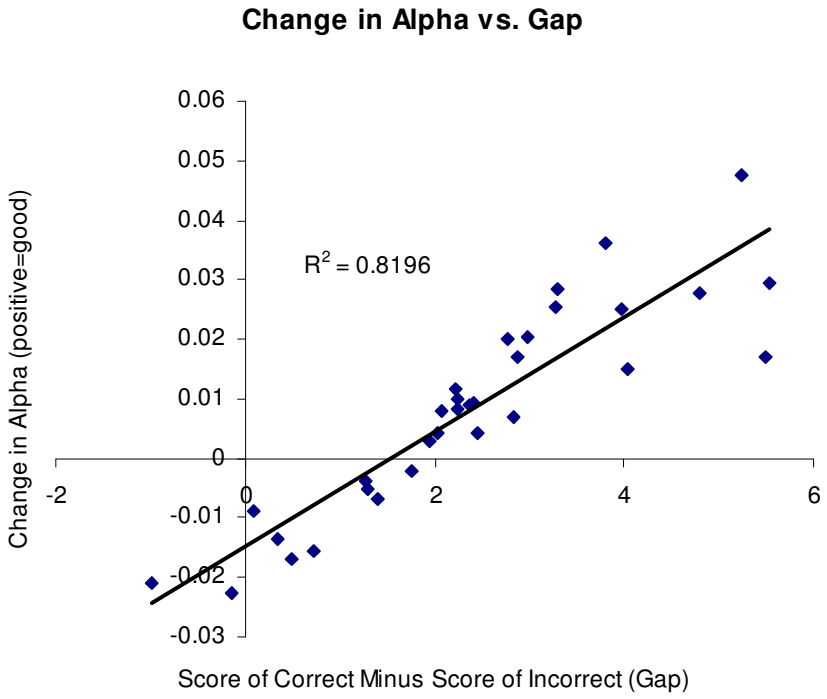


Figure 3: Relationship between change in alpha and gap

These plots continue to show strong relationships between the variables. This matches the theoretical explanation presented in section 4.1. Specifically, questions with a low or negative "gap" are those which lower the variance of the overall test score. Low total variance has been shown to be both mathematically and empirically the crucial component of coefficient alpha. Combined with what has been presented about the mathematical behavior of alpha, these graphs imply that a question's average inter-item correlation and, more directly, a question's "gap" are plausible causes of a question behaving poorly as measured by "alpha if item deleted."

Another measure to quantify the appropriateness of a question is the discriminatory index (Kelley, 1939). This statistic is calculated by comparing the average score of the top quartile to the bottom quartile. (Example: 4th Q 60% of students correct, 1st Q 25% of students correct → Discriminatory index 0.60 – 0.25 = 0.35) This statistic can also be shown to correlate highly with alpha if item deleted as shown in Figure 4.

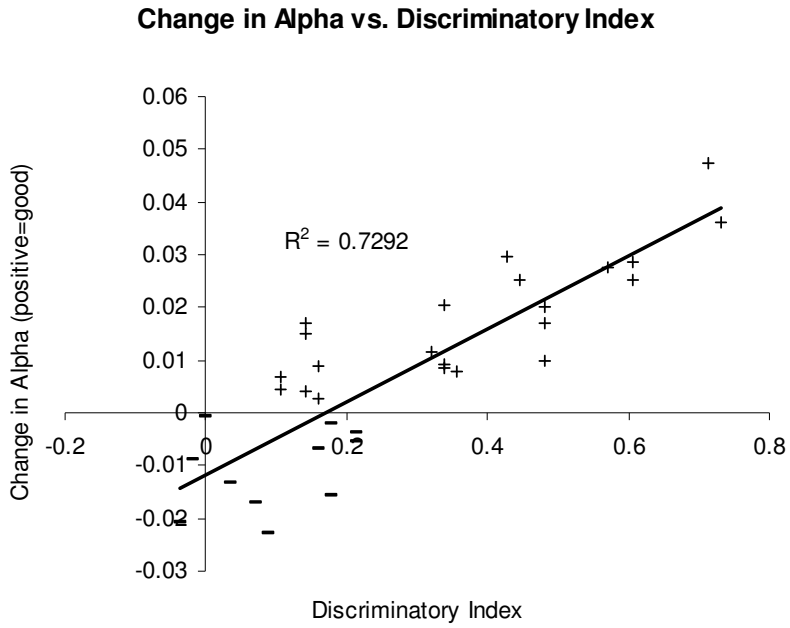


Figure 4: Relationship between change in alpha and discriminatory index

For this group of students, discriminatory index does not correlate as strongly as the "gap." Because the discriminatory index only accounts for half of the subjects, this is not surprising. However, for two other groups analyzed, change in alpha correlates more strongly with discriminatory index than with "gap." The lack of a consistent pattern limits further conclusions. Table 2 shows the correlations of alpha with the various other measures presented previously for three courses.

Table 2: Correlation of alpha-if-deleted with various item metrics

Course	$\bar{r}$	**Gap**	**Disc. Index**	**n**
OU REU	0.977	0.905	0.854	27
OU Engr	0.991	0.877	0.918	24
OU Math	0.973	0.889	0.935	14
Outside	0.982	0.970	0.905	38

Key : $\bar{r}$ average inter-item correlation
n students in each class

4.4 Explanation using comments from focus groups

Over half of the students who took the test attended a focus group where questions were discussed in detail. This allowed more scrutiny on a question-by-question basis. Using the comments of the focus groups, important information can be obtained about what makes a question "bad" in terms of alpha. Table 3 (next page) presents the ten worst questions in terms of "alpha if item deleted" (marked by a minus sign in Figure 4).

By evaluating these questions in such a manner, it is important to remember that alpha is a property of this set of scores and not of the test itself. The overall alpha and the "bad" questions will vary from class to class. This could be partly due to chance but also could indicate that one professor covered a topic whereas another did not or that topics were covered in different manners with varying results. These variations bring to light the difficulty of defining a target population and finding a representative, consistent sample. While the SCI has a target of statistical beginners, specifically those who are engineering

majors, the varied backgrounds and classroom exposure make finding the precise target audience (i.e., those who have been exposed to all concepts) impractical.

Table 3: Ten worst questions in terms of "alpha if item deleted"

Rank	Question Topic	Possible problem	Student comments
33	Meaning of p-value	Too many symbols; depends on remembering the definition	Most students guessed
32	Meaning of p-value	Again, depends on remembering the definition. One answer was very nearly correct but wrong by one word.	Several students guessed; a strong distracter threw off others
31	68-95-99 rule for normal	Requires remembering a rule	People who got it correct say they just remember the rule
30	Parent distribution of a sample	*No useful comments*	n/a
29	Calculating standard deviation	Depends on attention to detail	They think it is easy as long as you read carefully
28	t-distribution	*No useful comments*	n/a
27	Sample vs. population	Poorly written: incorrect choices looked different from correct choice	One student chose the correct answer for incorrect reasons
26	Design of experiment	Advanced topic	Most students guessed
25	Variability of a histogram	Students do not understand the graphs	Most students discussed lack of understanding
24	Central tendency	Term "central tendency" possibly confusing	One mentioned being confused by the term

Note: These questions ranked highest on alpha if deleted, therefore are considered the "worst" questions relative to the remaining questions.

The comments indicate that questions on which students guessed had a negative impact on alpha. This makes sense in light of the other data presented because one expects a question on which students guess to have a "gap" near zero. It is also likely that these questions measure some attribute other than statistical reasoning, such as test-taking ability or memory. This is plausible when compared with the effect that negatively

correlated items have on alpha. These items appear not to be measuring the same construct. When this happens, inter-correlations among items tend to be smaller. In other words, these questions are not *internally consistent* with the rest of the test.

4.5 The Big Picture

The reliability analysis is conducted after each round of test administration and used to guide revisions of the SCI. Table 4 shows the pre-test and post-test coefficient alpha for the combined course data from each semester. Moving down the chart, the test shows an increasing trend on the post-test, indicating the revisions are successful. The pre-test consistently has an alpha in the 0.69 range for the past four semesters. Moving across the chart for each semester, there is usually an increase in alpha from pre-test to post-test. This is consistent with the in-depth analysis because a pre-test will be subject to more guessing and test-taking tricks than a post-test which should more accurately assess the knowledge of the students.

Table 4: Coefficient Alpha for the SCI across six semesters

Semester	**Pre-Test Alpha**	**Post-Test Alpha**
Fall 2002	n/a	0.6494
Summer 2003	0.7434	0.6965
Fall 2003	0.6915	0.7031
Spring 2004	0.6979	0.7203
Fall 2004	0.6943	0.6692
Spring 2005	0.6852	0.7600

Table 5: Coefficient Alpha (with Post-Test components) for the SCI across six semesters

Semester	**Post-Test**					
	Alpha	**n (students)**	**k (items)**	σ_t^2	$\sum V_i$	**Range**
Fall 2002	0.5957	174	32	15.11	6.39	22
Summer 2003	0.6965	66	33	18.91	6.64	19
Fall 2003	0.7031	241	34	18.10	6.63	21
Spring 2004	0.7203	91	35	18.72	7.14	18
Fall 2004	0.6692	107	37	16.76	7.85	19
Spring 2005	0.7600	59	39	22.29	8.22	19

5. The Behavior of Alpha: Simulated Data

To further test the findings about what makes questions unreliable (i.e., people with low overall scores answer correctly and people with high overall scores answer incorrectly), a simulation was run using random numbers generated in MatLab.

5.1 Simulation parameters

The simulation consisted of 10 students, who ranged in ability from 0.1 to 1.0. The first test contained 10 questions, with two questions from each of 5 levels of difficulty: -0.4, -0.2, 0, 0.2, and 0.4 (hardest to easiest). The difficulty of each question is added to the student's ability to give a probability of answering a question correctly. For example, a student with an ability of 0.3 answering a question with difficulty of 0.4 will have a 0.7 (70%) probability of answering correctly. The probability is compared to the corresponding random number (uniform distribution, 0 to 1) to determine if the student receives a 0 or 1 score for that question. (Note: If a probability is less than 0 or greater than 1, the probability is taken to be 0.0 or 1.0, respectively.) The theoretical probabilities and corresponding true scores are given below:

Table 6: Theoretical Probabilities for MatLab simulation of Alpha behavior

Student	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	TrueScore
1	0.5	0.5	0.3	0.3	0.1	0.1	0	0	0	0	1.8
2	0.6	0.6	0.4	0.4	0.2	0.2	0	0	0	0	2.4
3	0.7	0.7	0.5	0.5	0.3	0.3	0.1	0.1	0	0	3.2
4	0.8	0.8	0.6	0.6	0.4	0.4	0.2	0.2	0	0	4
5	0.9	0.9	0.7	0.7	0.5	0.5	0.3	0.3	0.1	0.1	5
6	1	1	0.8	0.8	0.6	0.6	0.4	0.4	0.2	0.2	6
7	1	1	0.9	0.9	0.7	0.7	0.5	0.5	0.3	0.3	6.8
8	1	1	1	1	0.8	0.8	0.6	0.6	0.4	0.4	7.6
9	1	1	1	1	0.9	0.9	0.7	0.7	0.5	0.5	8.2
10	1	1	1	1	1	1	0.8	0.8	0.6	0.6	8.8

To test the effect of questions where low students answer correctly (and vice versa), two additional sets of questions were added. The first set had three questions of

difficulty -0.4, -0.2, and 0 (13 total questions). The second set added questions of difficulty 0.2 and 0.4 (15 total questions). For these additional questions (first 10 remain unchanged), the probabilities are reversed so that the lowest student has an ability of 1.0 and the highest has an ability of 0.1 only on the additional questions. The probabilities of a correct answer for the five new questions are given below:

Table 7: Theoretical Probabilities for Five Additional Questions in Simulation

Student	Prob11	Prob12	Prob13	Prob14	Prob15
1	1	1	1	0.8	0.6
2	1	1	0.9	0.7	0.5
3	1	1	0.8	0.6	0.4
4	1	0.9	0.7	0.5	0.3
5	1	0.8	0.6	0.4	0.2
6	0.9	0.7	0.5	0.3	0.1
7	0.8	0.6	0.4	0.2	0
8	0.7	0.5	0.3	0.1	0
9	0.6	0.4	0.2	0	0
10	0.5	0.3	0.1	0	0

For each of the three data sets (10 questions, 13 questions, 15 questions), 100 runs were performed in MatLab.

5.2 Simulated Results

The results of the MatLab simulation on the behavior of alpha are shown in Table 8.

Table 8: Results of MatLab Simulation on the Behavior of Alpha

Measure	**10 Question Test**	**13 Question Test**	**15 Question Test**
Mean alpha	0.8071	0.4220	0.0865
Median alpha	0.8197	0.4805	0.1783
St. dev. of alpha	0.0586	0.2547	0.3869
Minimum alpha	0.6278	-0.8264	-1.5517
Maximum alpha	0.9025	0.7766	0.6750

The 10 Question Test has the highest alpha by all the metrics shown above and also the smallest standard deviation from the 100 runs. When questions were added on which the low students were more likely to answer correctly than the high students, alpha of the simulated test was lowered. The variance of the alpha runs also increases when

these “bad” questions are added. This implies that “bad” questions not only lower alpha but also make it more variable, which could further reduce the utility of coefficient alpha for tests which have a high proportion of poor items. These findings from the simulated data match those from the real test data, further strengthening the evidence.

6. Conclusion

Coefficient alpha is an important tool in the assessment of test reliability. This paper provides insight into the behavior of alpha from a theoretical vantage and extends this to data from a real test. High variance of scores is the key component needed to attain a high coefficient alpha. Focus groups conducted with students after taking the test show that there are several possible causes for questions that adversely affect alpha – guessing, the use of test-taking skills, or when recalling a definition is necessary. In general, these “bad” questions do not conform to the material on the test and have high “alpha if item deleted” values, which is highly correlated with average inter-item correlations and the discriminatory index.

This paper provides insight for using coefficient alpha as a tool to aid in the revision of questions and thus improving the overall reliability of a test. Coefficient alpha and “alpha if item deleted” identify which questions are not conforming to the overall conceptual framework of the test. The results presented here indicate that these measures can be used to aide in item revision or deletion, especially when coupled with focus group discussion. Coefficient alpha and “alpha if item deleted” should simply be considered tools in the test-writer’s toolbox. Other judgments, such as those based on validity, are still important in evaluating the appropriateness of test items.

References

Cortina, J.M. 1993. What is Coefficient Alpha? An Examination of Theory and Applications. *Journal of Applied Psychology*. 78 (1): 98-104.

Cronbach, L.J. 1951. Coefficient Alpha and the Internal Structure of Tests. *Psychometrika*. 16 (3): 297-334.

Evans, D.L., G.L. Gray, S. Krause, J. Martin, C. Midkiff, B.M. Notaros, M. Pavelich, D. Rancour, T.R. Rhoads, P. Steif, R.A. Streveler, and K. Wage. 2003. Progress On Concept Inventory Assessment Tools. *Proceedings of the 33rd ASEE/IEEE Frontiers in Education Conference*, Session T4G-8.

Guttman, L. 1945. A Basis for Analyzing Test-Retest Reliability. *Psychometrika*. 10 (4): 255-282.

Hake, R. 1998. Interactive-engagement vs. traditional methods: A six-thousand student survey of mechanics test data for introductory physics courses. *American Journal of Physics*. 6 (1): 64-75.

Halloun, I. and D. Hestenes. 1985. The initial knowledge state of college physics students. *American Journal of Physics*. 53 (11): 1043-1055.

Hestenes, D., M. Wells, and G. Swackhamer. 1992. Force Concept Inventory. *The Physics Teacher*. 30 (March): 141-158.

Kelley, T. 1939. The Selection of Upper and Lower Groups for the Validation of Test Items. *Journal of Educational Psychology*. 30: 17-24.

Kuder, G.F., and M.W. Richardson. 1937. The Theory of the Estimation of Test Reliability. *Psychometrika*. 2 (3): 151-160.

Stone, A., K. Allen, T.R. Rhoads, T.J. Murphy, R.L. Shehab, C. Saha. 2003. The Statistics Concept Inventory: A Pilot Study. *Proceedings of the 33rd ASEE/IEEE Frontiers in Education Conference*. Session T3D-6.

Streiner, D. L. 2003. Starting at the Beginning: An Introduction to Coefficient Alpha and Internal Consistency. *Journal of Personality Assessment*. 80 (1): 99-103.

Acknowledgements

The author acknowledges the contributions and suggestions of Teri Reed Rhoads, Teri J. Murphy, Robert A. Terry, and Andrea Stone, as well as anonymous reviewers from the Journal of Engineering Education. Their feedback has enhanced the preceding exposition.

Appendix: Derivation of variance for dichotomous scoring

This relationship is derived using the basic definition of population variance:

$$\sigma^2 = V_i = \frac{\Sigma(x_i - \mu)^2}{n}$$

where: x_i are the individual observations (0 or 1)
μ is the population mean (p_i for each question)
n is the total number of observations (students)

For dichotomously scored data, the sum portion of the variance equation can be broken down into the 0 and 1 scores:

For 0 scores on a question: $\Sigma(x_i - \mu)^2 = (0 - p_i)^2 q_i n = p_i^2 q_i n$

The term $(0 - p_i)^2$ represents the fact that 0 is the value of each observation (x_i) and that the overall mean for each question is p_i. The term $q_i n$ accounts for summing all incorrect scores for that question (the proportion incorrect multiplied by the total number).

For the correct students, the same logic holds in calculating *Vi*, but now each x_i is 1 and the total number of correct students is $p_i n$.

For 1 scores on same question: $\Sigma(x_i - \mu)^2 = (1 - p_i)^2 p_i n = q_i^2 p_i n$

Combining the 0 and 1 portions and dividing by n yields the total variance for an individual question (V_i):

$$V_i = \frac{p_i^2 q_i n + q_i^2 p_i n}{n}$$

The next step is to divide out the *n*'s and re-arrange the numerator:

$$V_i = p_i q_i (p_i + q_i)$$

The term $p_i + q_i$ is the sum of the proportion correct plus the proportion incorrect. This must total 1. Therefore, the final result for each question's variance is:

$$V_i = p_i q_i$$

CHAPTER VII

Development and Equivalency of the Online SCI

Abstract

Begun in Fall 2004, an online version of the Statistics Concept Inventory was designed to streamline data collection and reduce class-time burdens associated with administration, thus enabling greater dissemination. Data from the Fall 2005 post-test identified potential differences between online and paper versions, primarily in the Inferential sub-test. However, confounding was present with the paper administration predominantly at one university. A more controlled (and necessarily smaller) study from the Spring 2006 pre-test showed differences to be minimal. Previous research suggests that equivalency between computerized and paper tests is certainly attainable, especially for highly computer-literate subjects on a non-speeded test. This chapter concludes that any potential differences between online and paper versions of the SCI are mitigated by the enhanced dissemination using the online version.

1. Background

Mead and Drasgow (1993) conducted a meta-analysis of previously-published studies on computerized testing. Studies were included based on the following criteria: non-clinical population; at least high school age but not geriatric; cognitive ability test; reliability was reported or available elsewhere; comparison between paper-and-pencil and comparably-constructed computerized test. A total of 25 studies were analyzed; 11 of these used the ASVAB.

Across administration mode, a correlation of 0.91 (disattenuated; 0.90 when restricted to [0,1] range) was found between test scores. Stratifying across test type, timed

power tests (0.97; 0.95) had extremely high correlations while speeded tests (0.72) were less-highly correlated. The scaled mean differences indicate that computerized tests are slightly more difficult but not appreciably (d_i = -0.04 for all tests; d_i = -0.03 for timed power tests). The authors conclude that there is "little effect of medium of administration for power tests" (p. 456) but caution should be used in claiming that a computerized version of a test is constructed to be equivalent to its paper-and-pencil version.

Goldberg and Pedulla (2002) investigated mode effects on the GRE in paper-and-pencil and computerized (non-adaptive) forms. The computerized forms were presented both with and without editorial control (the ability to review and change answers), giving three testing conditions in total. For both computerized versions, items were presented one-at-a-time and respondents were asked to verify their answers before proceeding to the next item. Subjects consisted of 222 "traditional-aged" college juniors and seniors at two liberal arts colleges in the Northeast; this restrictive subject pool and a gender imbalance (72% female) hinder the external validity of the study.

A MANCOVA was performed with the three GRE sub-tests (Verbal, Quantitative, Analytical) as dependent variables. Independent variables were test mode, computer familiarity (trichotomized as low-medium-high from a self-report scale), and the mode × familiarity interaction. Undergraduate GPA and SAT scores were included as covariates, but SAT scores were dropped from the analysis because they interacted significantly with test mode.

A significant main effect (Wilk's λ) was found for both test mode ($p < 0.001$) and computer familiarity ($p \leq 0.004$). For the subtests, only Quantitative had a significant interaction ($p = 0.042$). Quantitative and Analytical had significant main effects on

computer familiarity, while all three sub-tests had significant main effects across mode. *Post hoc* comparisons show that the paper-and-pencil participants out-performed the computerized-without-editorial-control (CWEC) group by approximately 10% on each of the sub-tests. Adjusting for GPA, there is little difference by computer familiarity on the Verbal sub-test, however the Analytical sub-test shows scores approximately 10% higher for those with high computer familiarity compared to low familiarity. Quantitative is more difficult to interpret because there is an interaction between mode and familiarity. The overall results are comparable to Analytical, but the CWEC mode apparently mitigated differences due to familiarity; small sample sizes (some as small as 5 students) limit the generalizability at this level.

The GRE is a timed test. In this study, the analytical subtest was found to be "highly speeded in all three test modes" (p. 1065) and especially in the CWEC mode. This essentially means that the poorer performance on the computerized tests may be due to time restraints rather than any other variable such as computer familiarity, test design, or content.

Clariana and Wallace (2002) compared performance on a 100-item multiple choice test in paper (n = 51 students) and computerized (54) modes. The computerized version displayed one question per screen and allowed subjects to review and change previous answers. The paper test had a fixed order, whereas the computer version was randomized. An attitudinal survey of preferences for distance learning was included to explain possible differences between presentation modes.

A highly-significant difference was found ($p < 0.001$) between modes, with computer (mean 83.0) out-performing paper (76.2). Several 2×2 ANOVA were

conducted: gender, computer familiarity, and competitiveness showed no mode effects. However, students with high content familiarity performed especially well on the computerized version. A series of correlations were calculated between the attitudinal measure and the test scores. For the paper version, the strongest correlations were found among qualities labeled by the authors as "egocentric" (e.g. not caring how others perform in the class). For the computer version, only one significant correlation was found between the attitudinal measures and test score (r = 0.29, "I work harder than others to stand out from the crowd"; considered a sign of "competitiveness").

Spray, *et al.* (1989) compared examinee performance and item characteristics on paper-and-pencil and computerized versions of a Marine Corps test relating to ground radio repair. This subject pool had more computer experience than was common at the time (1989), which makes the results more generalizable to the present. Data were analyzed in a 2×2 ANOVA with medium (computer, paper) and class, for each of three content units. The number of classes was 11, 16, and 9, for the three content units; the number of subjects (computer, paper) were 113, 121; 179, 172; and 96, 82.

The results show no interactive effects between medium and class. For two of the three content units, there was no significant effect for medium (there was a significant main effect for all three units by class). The third content unit showed significantly higher scores for the paper-and-pencil version, although the mean was only 1 point higher out of 25. A Kolmogorov-Smirnov test on the cumulative score distributions for each content unit revealed no significant differences.

Analysis of the logistic item curves showed that only four items of 230 yielded different models at a significance level of 0.01, although the power is limited because

most items were answered by less than 100 examinees. Additional attention was paid to the third content area because it was the only one which had a significant medium effect in the ANOVA. Eleven of the 75 items showed significant differences at 0.10, but no distinguishing features could be found between any of these items. Moreover, analyzing only items answered by at least 50 examinees, the item difficulties have nearly identical distributions across media. Therefore, the authors conclude that the paper and computer versions of this instrument are equivalent.

Cole, *et al.* (2001) compared paper and web-based versions of the Force Concept Inventory. The FCI was administered as both pre-test ($n_{paper} = 614$, $n_{web} = 559$) and post-test ($n_{paper} = 407$, $n_{web} = 518$) to students enrolled in introductory physics at a "medium sized university in the midwest" (p. 6). To achieve a balanced design, subjects were randomly assigned to either paper or web groups based on the last digit of their student ID numbers. The group which took the paper pre-test took the web post-test, and vice versa; further control was maintained by having students complete a science attitudinal survey in the opposite medium at both pre- and post-administrations.

While the scores on the pre-test were lower, the conclusions drawn from the comparisons are equivalent to the post-test; therefore, only post-test results are discussed. The FCI proved to be nearly equally reliable across versions ($\alpha_{paper} = 0.87$; $\alpha_{web} = 0.89$). A 2×2 ANOVA analyzed gender and administration effects; age and ethnicity were not considered because the majority of students were Caucasians between 18 and 22. Most importantly, administration and interaction effects were non-significant, although a significant effect was found due to gender. Item means were compared, and only one was found to have a significant difference across medium (based on a critical value of 0.01),

out of 30 total items. Response patterns were analyzed via χ^2, and two items were found to differ significantly along these lines. Given these minimal differences, the authors conclude that their web-based version of the FCI is equivalent to the paper version.

1.1 Clarification

Many large-scale computerized tests are presented in an *adaptive* form: from a common item pool, the item order is tailored to the individual examinee based on his responses. The Graduate Record Examination (GRE) is an adaptive test which many people have some familiarity with. The SCI is *not* an adaptive test.

1.2 Implications

The studies reviewed certainly demonstrate that it is possible to construct a computerized test such that it is equivalent to a paper version. Because the SCI is designed as a power test rather than a speeded test, the process becomes easier. Further, all participants are college students and most commonly engineering majors, which suggests that computer familiarity should not limit the equivalence.

2. Server

The SCI is housed in Carson Engineering Center on a desktop computer, running Microsoft Windows XP with an Intel Pentium III (801 MHz) processor. The server runs Apache version 2.0.55 [`http://httpd.apache.org/`], which is a free HTTP server package. The database runs mySQL version 5.0.16 [`http://www.mysql.com/`]. The web interface was programmed in the PHP language version 5.1.1 [`http://www.php.net/`], which allows dynamic creation of webpages including database access. Apache, mySQL, and PHP are all free programs with a wealth of support information available online.

3. Database Architecture

3.1 Student table

The fields of the "Student" table are shown in the table below. The email address, provided by the instructor, servers as the primary key (read: unique identifier) for this table. All other information is optional. The student demographics are entered by the student when taking the SCI; these are the fields from "Gender" down to "Math_dif" in Table 1. Several fields contain information about the student's login behavior: "Taken" (number of logins), "IRB" (answer to ICF; 0 = "No", 1 = "Yes"), "StartTime" and "StopTime", "Feedback" (provided by student), and "Finished" (sets to 1 when SCI is completed).

Table 1: "Student" table fields

Field	Type	Null	Key	Default	Extra
LastName	varchar(20)	NO			
FirstName	varchar(20)	NO			
Email	varchar(100)	NO	PRI		
Section	varchar(50)	NO			
Passcode	varchar(250)	NO		0	
Taken	int(1)	NO		0	
Gender	char(1)	NO		.	
Race	varchar(10)	NO		.	
Year	varchar(5)	NO		.	
Enroll	varchar(6)	NO		.	
Major_gen	varchar(10)	NO		.	
Major_spec	varchar(100)	NO		.	
Exp_hs1	int(1)	NO		0	
Exp_hs2	int(1)	NO		0	
Exp_col1	int(1)	NO		0	
Exp_col2	int(1)	NO		0	
Exp_col3	int(1)	NO		0	
Exp_none	int(1)	NO		0	
First_time	int(1)	NO		1	
Math_alg	int(1)	NO		0	
Math_somecalc	int(1)	NO		0	
Math_allcalc	int(1)	NO		0	
Math_lin	int(1)	NO		0	
Math_dif	int(1)	NO		0	
IRB	int(1)	NO		0	
StartTime	datetime	NO		0000-00-00 00:00:00	
StopTime	datetime	NO		0000-00-00 00:00:00	
Feedback	text	NO			
Finished	int(1)	YES		NULL	

3.2 Administrator table

Table 2 shows the fields of the administrator ("Admin") table. "Email" is the primary key and is used as the login name for the administrator functions. The "Section" field ties the students to that administrator account. The four sub-test areas are set to 0 (off) or 1 (on) as requested by the instructor.

Table 2: "Admin" table fields

Field	Type	Null	Key	Default	Extra
Email	varchar(200)	NO	PRI		
Passcode	varchar(250)	NO		0	
Section	varchar(50)	NO			
Probability	int(1)	NO		0	
Descriptive	int(1)	NO		0	
Inferential	int(1)	NO		0	
Graphical	int(1)	NO		0	

3.3 Questions table

Table 3 shows the "Questions" table. The "Number" serves as the primary key. The numbering is maintained to correspond with the paper version of the SCI. Each question has space for up to six multiple choice options, although most questions use four. The "Topic" is set to correspond to one of the four sub-tests (*cf.* Table 2).

Table 3: "Questions" table fields

Field	Type	Null	Key	Default	Extra
Question	text	NO			
ChoiceA	varchar(200)	NO			
ChoiceB	varchar(200)	NO			
ChoiceC	varchar(200)	NO			
ChoiceD	varchar(200)	YES		NULL	
ChoiceE	varchar(200)	YES		NULL	
ChoiceF	varchar(200)	YES		NULL	
Topic	varchar(30)	NO			
Number	int(2)	NO	PRI	0	
Correct	char(1)	NO			
Graphic	blob	YES		NULL	
Description	varchar(255)	NO		.	

3.4 Answers table

Table 4 shows the fields of the "Answers" table. This is where student answers are stored. An auto-increment integer ("Key") serves as the primary key. The "Email" field corresponds to the "Student" table, and "Number" corresponds to the "Questions"

table. The student's "Answer" and "Confidence" are saved based on responses to the items.

Table 4: "Answers" table fields

Field	Type	Null	Key	Default	Extra
Number	tinyint(3)	NO		0	
Answer	char(2)	NO			
Email	varchar(200)	NO			
Key	int(10)	NO	PRI	NULL	auto_increment
confidence	tinyint(3)	NO		0	

3.5 Order table

The "Order" table is used when each student's random question order is generated. The "Sequence" field ranges from 1 to the maximum number of items (e.g., 38 if taking all four sub-tests). The "Qnum" field defines which question is displayed at the appropriate point in the sequence.

Table 5: "Order" table fields

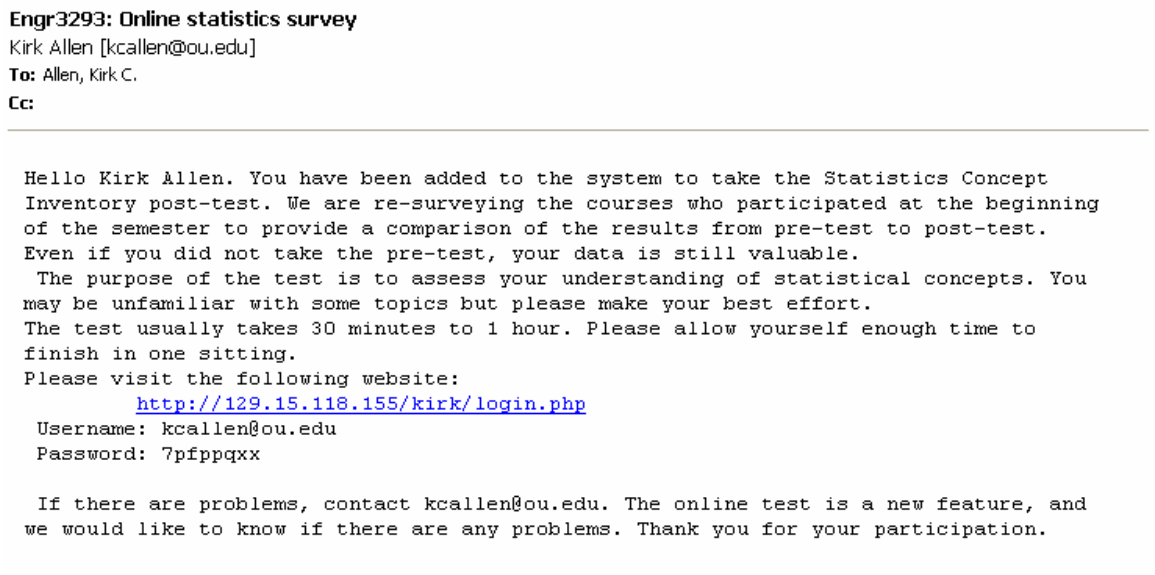
Field	Type	Null	Key	Default	Extra
Email	varchar(100)	NO	PRI		
Sequence	int(11)	NO	PRI	0	
Qnum	int(11)	YES		NULL	

4. Web Interface

4.1 The Test

General process

This section describes the online test interface as viewed by a student who is taking the test. After an instructor has decided to use the SCI, the administrator functions are used to add students and send instruction emails (described more fully in the *Administrator* section which follows). A sample of the instructional email for the Fall 2005 post-test is shown in Figure 1.



Engr3293: Online statistics survey
Kirk Allen [kcallen@ou.edu]
To: Allen, Kirk C.
Cc:

Hello Kirk Allen. You have been added to the system to take the Statistics Concept Inventory post-test. We are re-surveying the courses who participated at the beginning of the semester to provide a comparison of the results from pre-test to post-test. Even if you did not take the pre-test, your data is still valuable.
The purpose of the test is to assess your understanding of statistical concepts. You may be unfamiliar with some topics but please make your best effort.
The test usually takes 30 minutes to 1 hour. Please allow yourself enough time to finish in one sitting.
Please visit the following website:
http://129.15.118.155/kirk/login.php
Username: kcallen@ou.edu
Password: 7pfppqxx

If there are problems, contact kcallen@ou.edu. The online test is a new feature, and we would like to know if there are any problems. Thank you for your participation.

Figure 1: Instructional email received by a student

When the student visits the login page as directed in the email, the following screen appears (Figure 2). The input is checked against the "Student" table, and correct information passes the student onto the test, as indicated in Figure 3. Incorrect entry will result in Figure 2 being re-displayed with a note that the previous entry was incorrect.

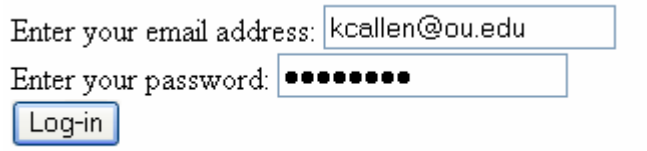


Figure 2: Login screen

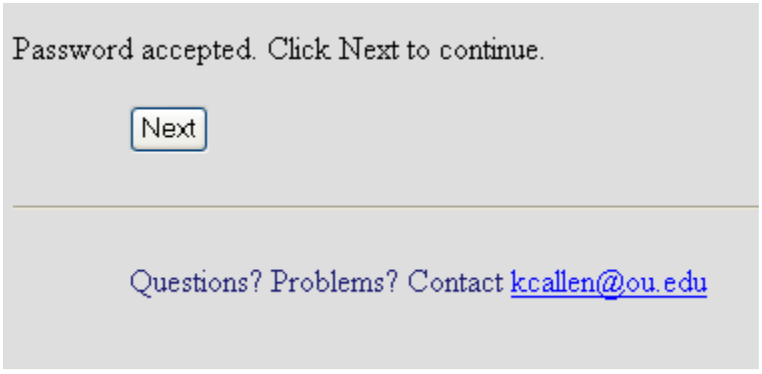


Figure 3: Successful login

At this point, a random question order is generated based on the student's section and stored in the "q_order" table; this random ordering is designed to prevent collaboration between students. The first screen which appears to the student is the informed consent form (ICF, Figures 4a and 4b). This is the same as the paper version of the ICF.

INFORMED CONSENT FORM FOR RESEARCH BEING CONDUCTED UNDER THE AUSPICES OF THE UNIVERSITY OF OKLAHOMA-NORMAN CAMPUS

"The Statistics Concepts Inventory (SCI): A Cognitive Achievement Tool in Engineering Statistics"

Principle Investigator: Teri Reed Rhoads, Industrial Engineering, 325-3419, teri.rhoads@ou.edu

Co-Investigator: Teri J Murphy, Mathematics, 325-4071, tjmurphy@math.ou.edu

Collaborators: Kirk Allen, Industrial Engineering, kcallen@ou.edu

Andrea Stone, Mathematics, adstone@ou.edu

Description of the Study: We are working to develop an assessment tool that will be used to measure student understanding of probability and statistics concepts. To this end, we ask you to complete the following questionnaire, responding to each item as accurately and honestly as you can. The questionnaire may take up to 35 minutes to complete. The questionnaire consists of two parts: a series of questions that will ask you about statistics concepts and a few questions about your academic background. In addition (if applicable), we ask that you allow your instructor to forward us your final course grade so that it can be used in analyzing this tool.

Potential Benefits and Risks to You: Once the development of the SCI is completed, it will be used to assess student achievement and improve instruction in statistics courses. This study is educational research and does not have any known associated risks beyond daily interactions between students and instructors. As with all testing situations, students may feel a certain amount of stress while completing the instrument.

Conditions of Participation: Participation is voluntary. You may discontinue participation at any time. Declining to participate or discontinuing participation will not result in any penalty. You must be 18 years of age or older to participate.

Confidentiality: All findings will be presented in aggregate form with no identifying information to ensure your confidentiality. Your name will not be used at any time. Instead, you will be assigned an ID code that will be used to link data. A list of participants will be avaialble to your instructor for the purpose of verifying participation. Your instructor will not know whether you elected to participate as a research subject or simply to take the test. Your instructor will not know whether you participated until after course grades have been reported.

Contacts for Questions About the Study: If you have questions about the research, you may contact Dr. Teri Rhoads at (405) 325-3419 or teri.rhoads@ou.edu. If you have questions regarding your rights as a research participant, please call the Office of Research Administration at (405) 325-4757 or irb@ou.edu.

Figure 4a: Informed Consent Form, top portion

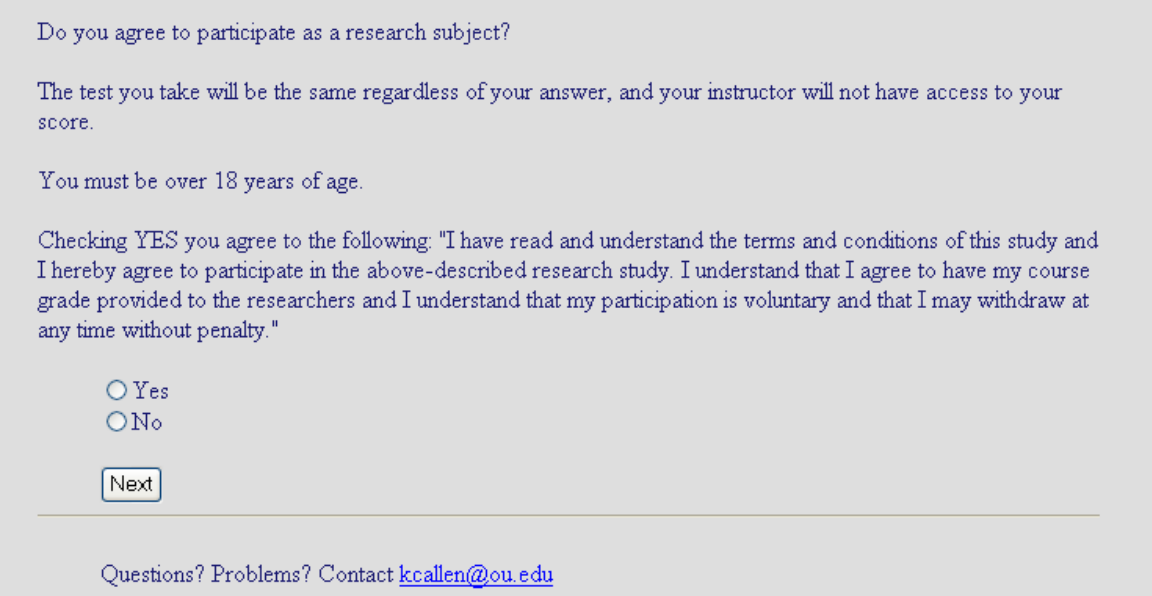


Figure 4b: Informed Consent Form, bottom portion

The student next views the demographics questionnaire (Figure 5). All responses are optional, and incomplete data is stored as "." in the database. Additionally, the starting time is recorded upon completing the demographics questionnaire.

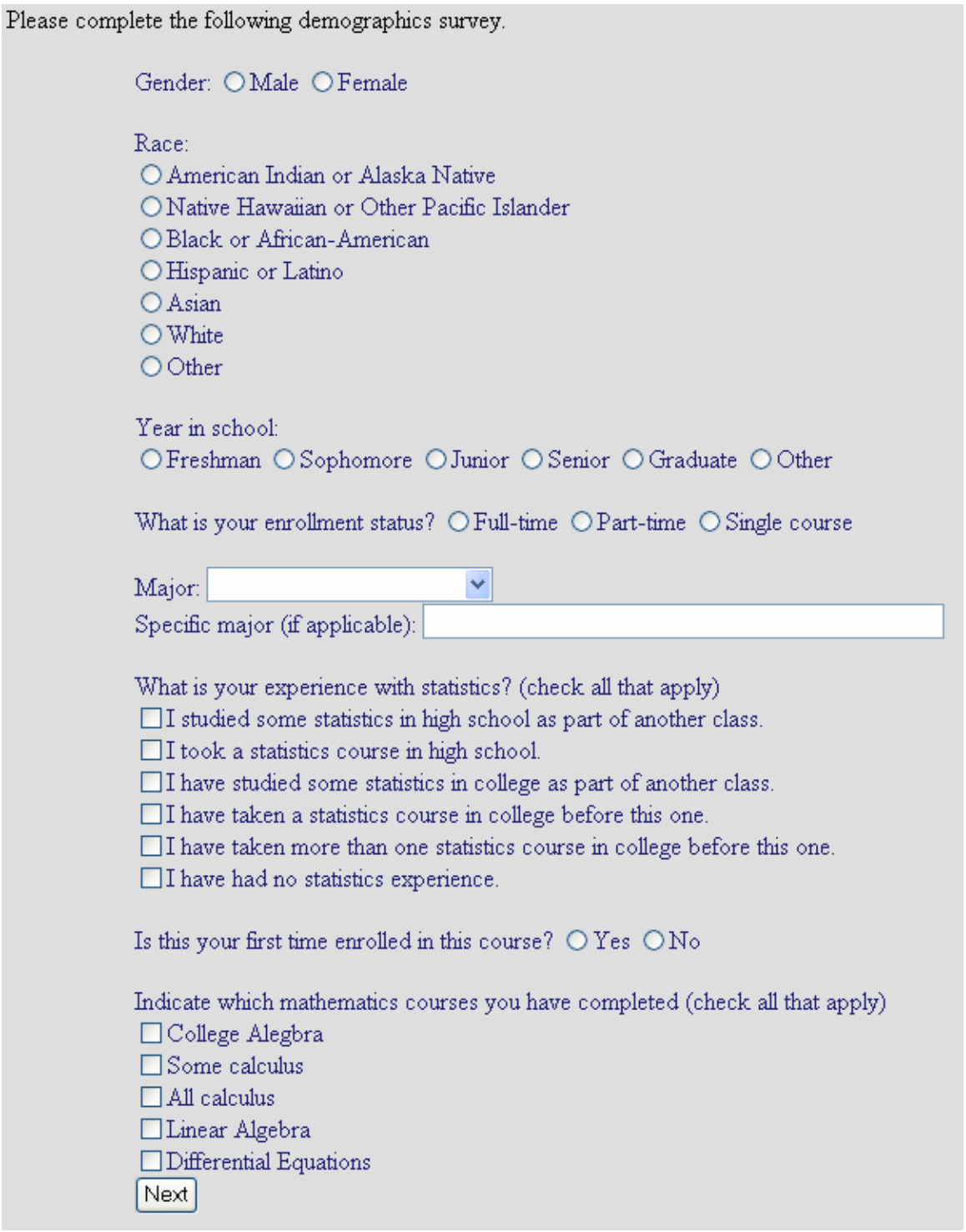


Figure 5: Demographic questionnaire

The student then proceeds to the test questions (Figures 6a and 6b). The question number (as known to the student, not based on the "Number" field of the "Questions" table) is displayed at the top to give the student an idea of his progress through the test.

The text of the question is displayed next, and the choices are marked with radio buttons for selection. Upon clicking "Answer," the student is shown his selection and given two options: 1) "Change Answer", for example if an incorrect button was clicked; or 2) rate his confidence in the answer and proceed to the next question (if a confidence is not selected, a default value of zero is stored in the database but ignored for data analysis).

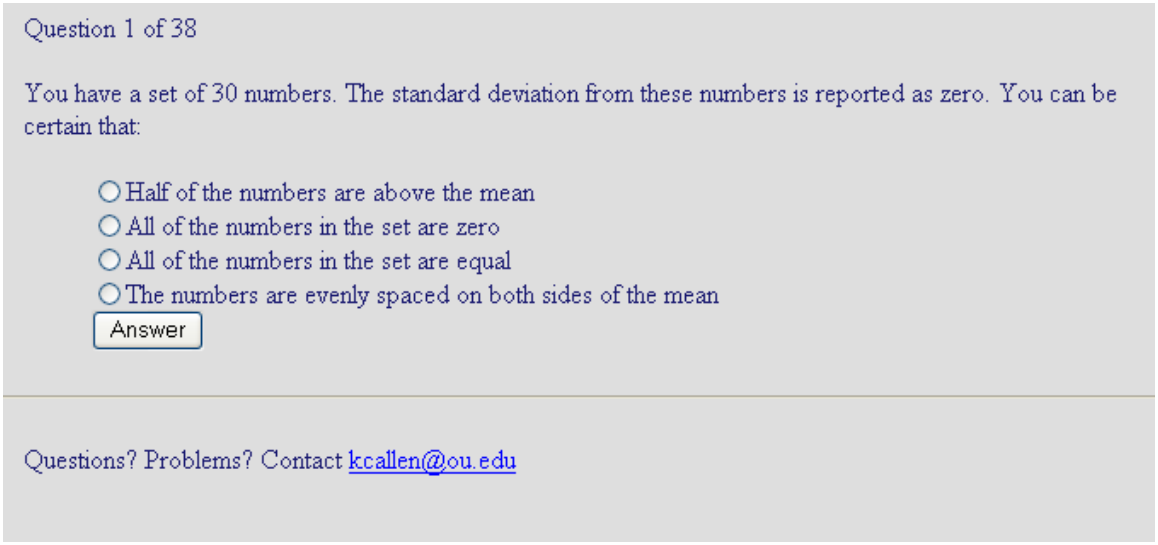


Figure 6a: Question display, prior to answering

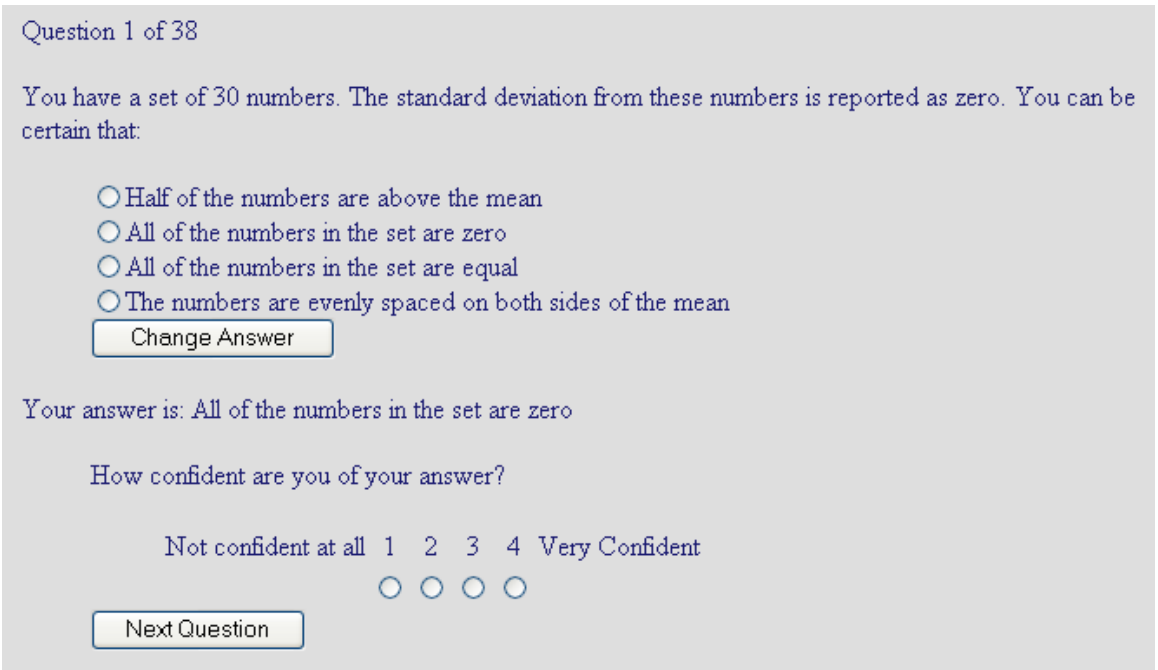


Figure 6b: Question display, after answering

After completing all questions, the student is directed to a feedback form (Figure 7) which is used to gather qualitative opinions of the testing system and allows reporting of any errors encountered. This form is optional. Additionally, the stopping time is recorded when this page is displayed.

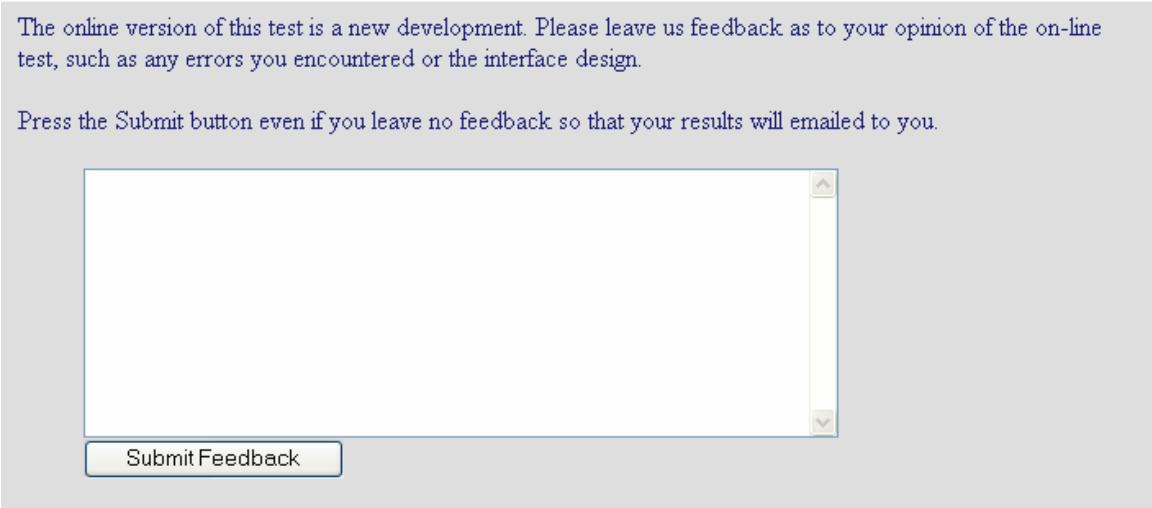


Figure 7: Feedback form after completing test (optional)

At this point, the student has completed the SCI. An email is generated and sent to the student containing his answers and the number correct (Figure 8). The letter answers are irrelevant unless the instructor plans to review the results, but they are included for thoroughness. The number correct gives the student a general idea of his results, but no diagnostic information can be gathered from his answers.

Results from online statistics survey
Kirk Allen [kcallen@ou.edu]
To: Allen, Kirk C.
Cc:

Here are your results. If your instructor plans to review the results, you may want to bring this to class.

The questions as you saw them were displayed in a random order. The numbers you see here are the actual question numbers, completely different from the order you saw them.

Number correct: 11

1) C

2) C

3) D

4) D

5) C

Figure 8: Results email (snippet)

Further comments

While students are encouraged to complete the test in one sitting, the system is been designed so that a student can stop in the middle and return to the same point. The early version of the online test allowed a single login, which proved irksome in resetting students who stopped in the middle such as due to a computer freeze-up. The total number of logins is recorded in the "Taken" field of the "Student" table to allow analysis of student login patterns.

Several steps have been taken to insure the privacy of the test items. First is the random question order, previously described. Second, the copy function has been disabled so that students cannot save the questions into a word processing program. Third, an attempt to print test items will output only a blank page. A screen capture appears to be the only mechanism for saving items, but this is likely too tedious for a student to attempt on a test which is not high-stakes.

4.2 Administrator

The administrator function (Figure 9) is designed so that a class instructor can perform the necessary functions of managing a class taking the online SCI. In practice, all administering has been done by the author thus far. A screenshot of the main menu is shown below. The relevant submenus are described in the sections that follow.

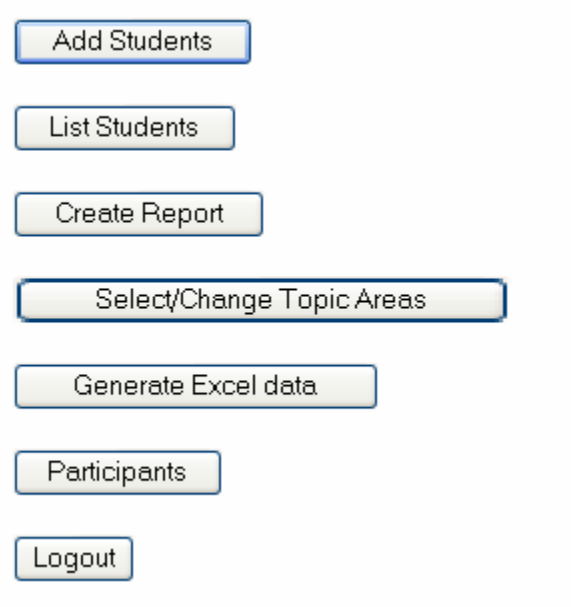


Figure 9: Administrator main page

Adding students

A student is added to the database by minimally inputting his email address (Figures 10a and 10b). Space is provided for first and last names as well, but in some instances student names are not provided by instructors. If left blank, the default name is "Statistics Student." To ease data entry, space is provided so that the email server is automatically appended to the student email alias for the most common servers. For example, if "kcallen" is entered in the "OU email" box, the address will automatically become "kcallen@ou.edu"; the functionality is provided for "@purdue.edu" as well. The "other email" box requires the input of an entire email address (e.g.,

"kcallen@hotmail.com"). When the "Submit" button is clicked, the student is inserted to the Student table as a member of the administrator's section. An eight-character password is randomly generated. Using the PHP Mailer [http://phpmailer.sourceforge.net/] class, an email containing the login instructions is sent to the provided address. Screen-shots below demonstrate adding a student. A sample email was provided as Figure 1.

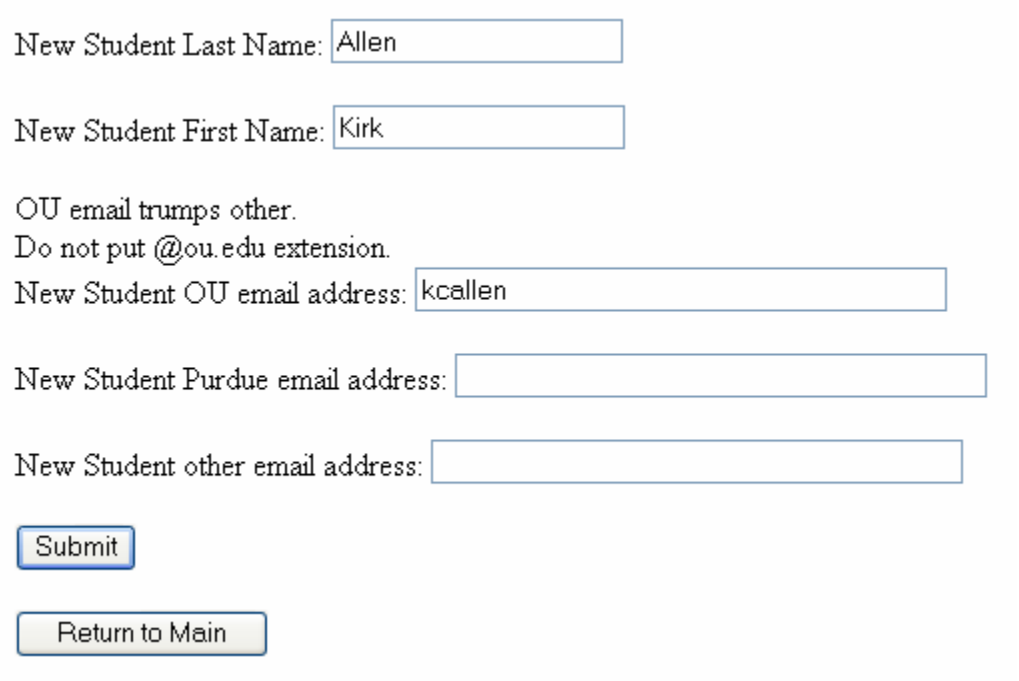


Figure 10a: Adding a student, prior to clicking "Submit"

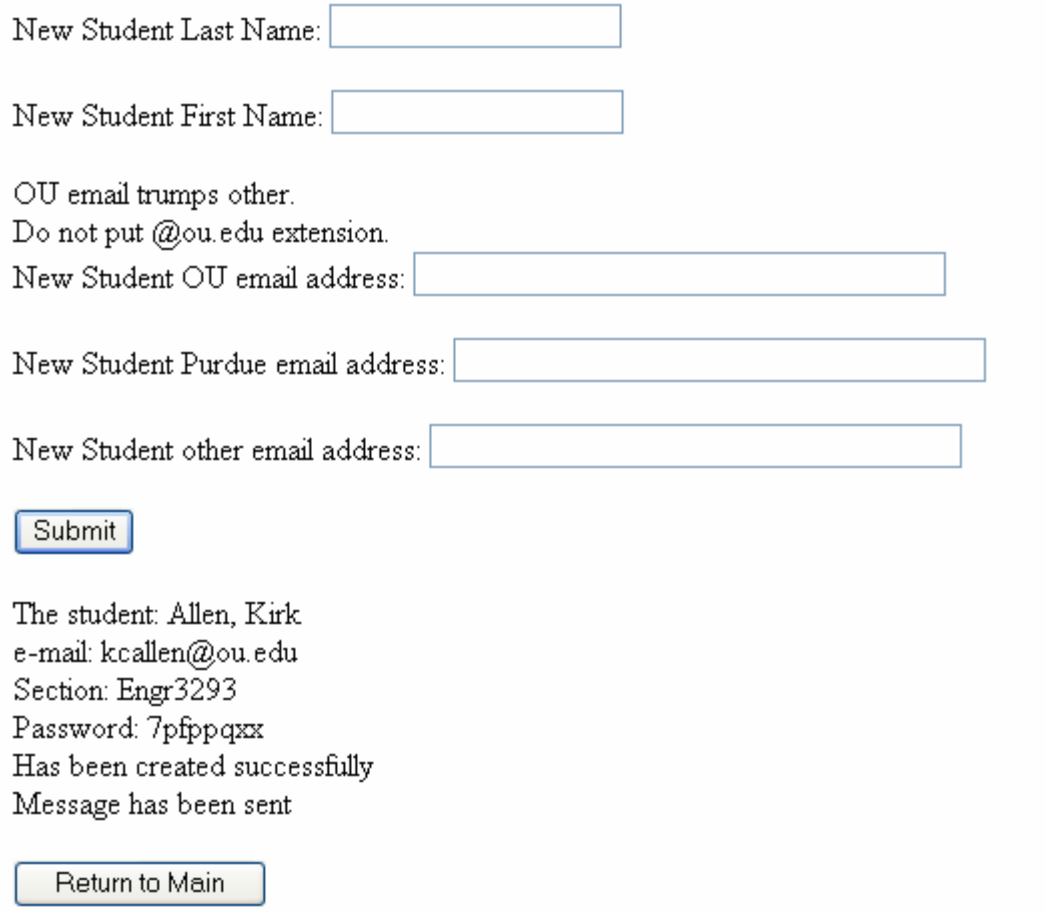


Figure 10b: Adding a student, after clicking "Submit"

Selecting topic areas

By default, each section will take all four SCI topic areas. The administrator page allows these to be toggled (Figure 11).

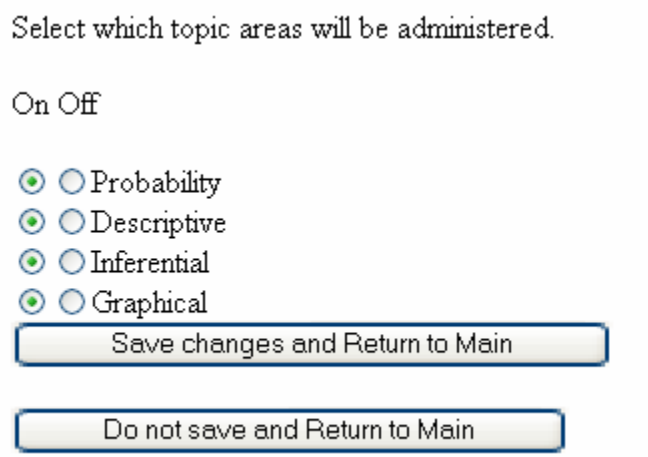


Figure 11: Topic area selection

Generating Excel data

This function outputs an HTML table which can be copied and pasted directly into Microsoft® Excel. The table contains five sections: 1) student demographics (major, year in school, mathematics experience, etc.); 2) login information (start time, stop time, number of logins); 3) letter answers for each item, defaulting to "." if omitted; 4) binary correct/incorrect, calculated within the PHP code, for each item; 5) confidence ratings for each item, again defaulting to "." if omitted. The students are numbered sequentially, and their names and email addresses are listed. However, to comply with IRB regulations, the identifying columns are saved to a separate file when performing the analysis in order to preserve the anonymity of the subjects. Additionally, only students who agreed to allow their results to be used for research purposes are displayed in this section (i.e., only those whose value of the "IRB" column in the "Student" table is 1).

List participants

This function eases gathering of student participants for instructors who give credit for participation. The first portion of this page lists participants by email address,

last name, first name, and number of items answered. The second portion lists non-participants by the same information, excepting the number of items answered.

4.3 High-level administrator

This "Super Administrator" page is intended for use only by the author or possibly other members of the SCI team. It facilitates several functionalities which were previously done manually through mySQL, such as adding administrators and viewing recent student logins. The commonly-used functions are described below.

Listing questions

This section is intended for use of reviewing the SCI with a class (Figure 12). The question is displayed in the same way as a student would view it. An additional button allows the correct answer to be initially hidden and then viewed later, for example after a class discussion. Navigation allows sequential proceeding either forward or backward, as well as jumping to any question.

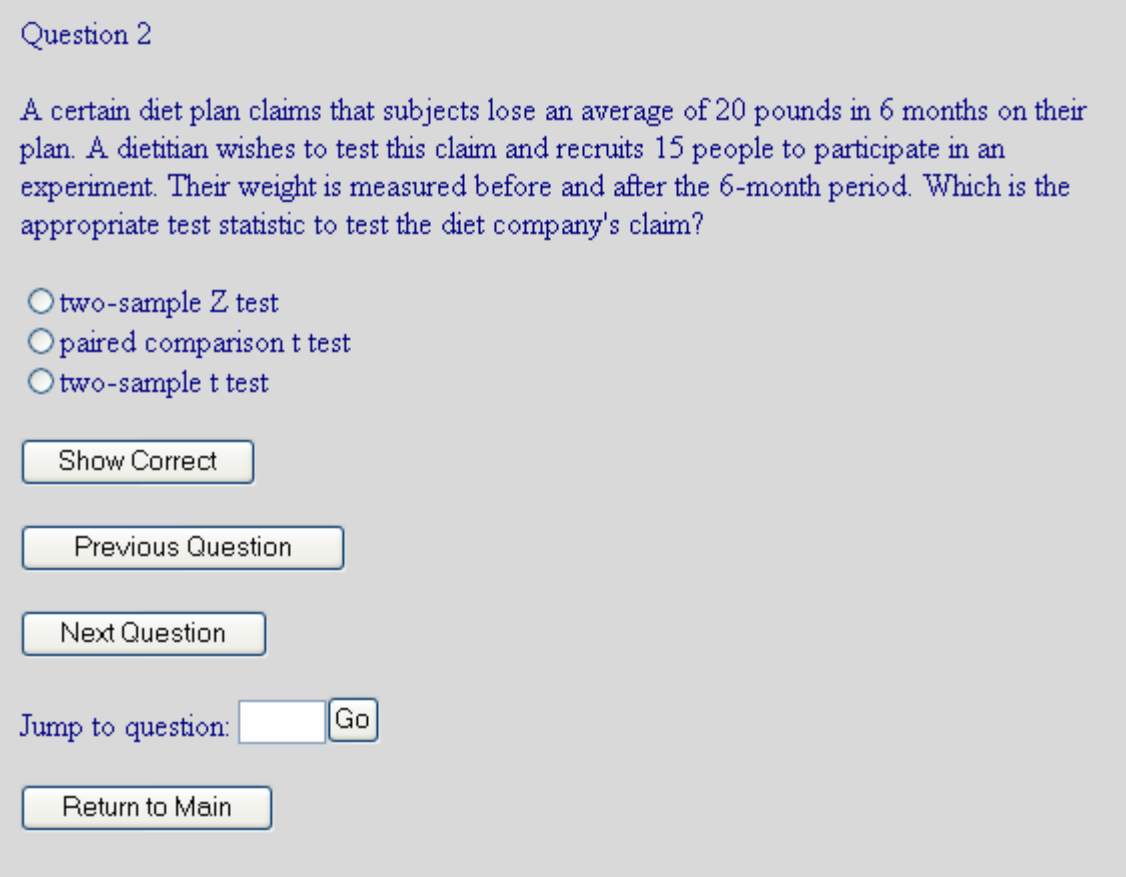


Figure 12: Listing questions

Display feedback

This allows easy gathering of the feedback left by students at the completion of the SCI. No identifying information is listed, and empty feedback is not displayed.

Manage administrator accounts

This section allows the user to add new administrators (login, password, section) or to manage existing accounts (Figure 13). For the existing accounts, two options are provided: 1) Delete, which erases that administrator from the database but leaves his students; 2) Reset, which erases the students for that section but leaves the administrator account (best to use when re-initializing a class from pre-test to post-test). A screen shot of a portion of this page is shown below.

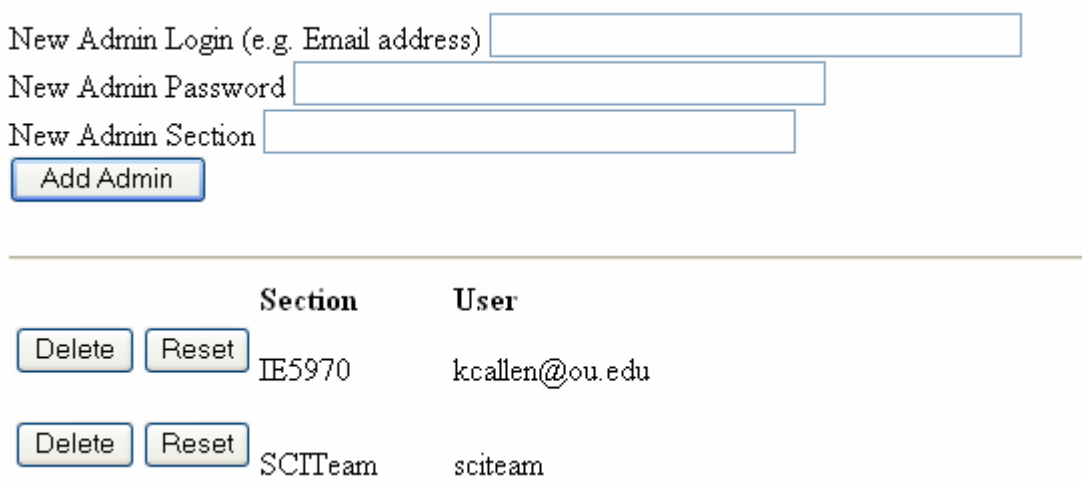


Figure 13: Administrator management page (snippet)

Logins today

This section gives the participation rate for the current day. The most common use is to see if any students are currently logged in, which is useful to know in the event that the server needs to be restarted.

4.4 Guest Login

To allow prospective professorial participants to review the instrument, a simple guest login was created. The login is similar to those for students and administrators, except that the password is hard-coded into the PHP code rather than being stored in the database. Viewers can choose to browse the full instrument or can filter by sub-topic. The main menu is shown in Figure 14. The questions are displayed as a group rather than individually, but the format is the same as students see (Figure 15).

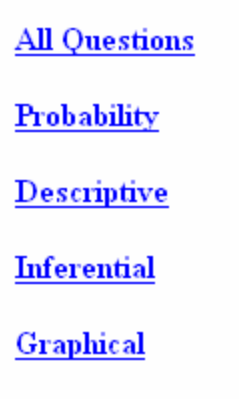


Figure 14: Guest account main menu

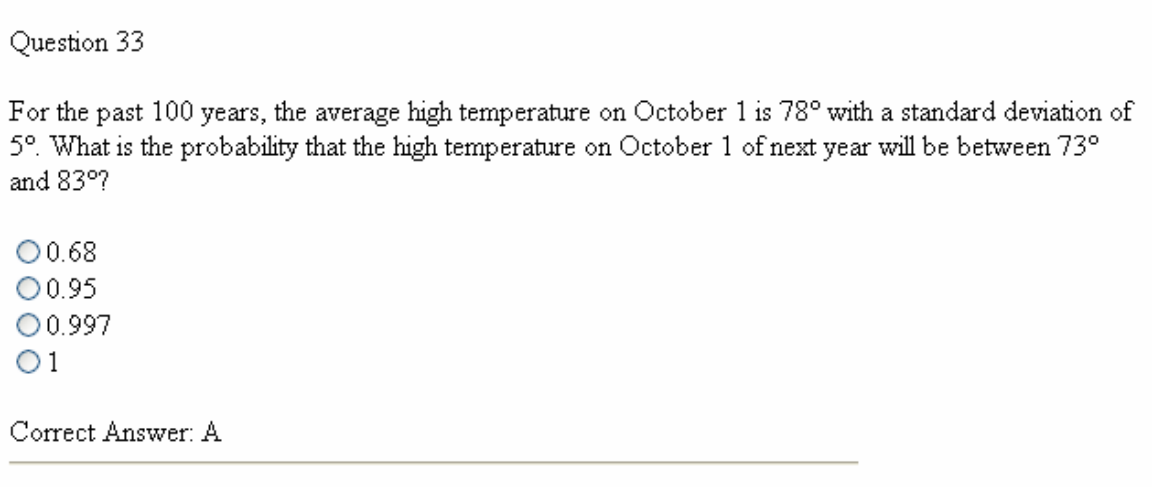


Figure 15: Sample of how questions are displayed

5. Online Analysis

This section analyzes the results which are specifically relevant to the online version of the test. These data are from the Fall 2005 post-test.

5.1 Participation rate

Table 6 summarizes the participation rate for the Fall 2005 online post-test. The participation varies widely by course. The highest rates were found in Math #1. This instructor reserved a computer lab so that the entire class could take the SCI at one time, even though she did not give extra credit. Only one student from this course did not participate. This contrasts with Engr, Math #2a, and Math #2b. To the best of the author's knowledge, these instructors did not provide any incentive or make announcements regarding potential SCI participation. On the positive side, nearly all classes had 100% completion of those students who logged in.

Table 6: Summary of online participation rates

Course	**Total students**	**Logins (%)**	**Completion (%)**	**Logins complete (%)**
Engr	82	13%	10%	73%
DOE	54	65%	63%	97%
Psych #1	74	30%	30%	100%
Psych #2	44	59%	59%	100%
Psych #3	38	55%	55%	100%
Math #1	34	97%	97%	100%
Math #2a	29	17%	17%	100%
Math #2b	32	34%	34%	100%
Metr	22	64%	64%	100%
External #5	216	68%	62%	91%
Total	*642*	*52%*	*49%*	*95%*
Column median	*41*	*57%*	*57%*	*100%*

How to read this table: "Total students" is the number of potential participants provided by the instructor; "Logins" is the percentage of those students who logged in to start the SCI; "Completion" is the percentage of total students who completed the SCI; "Logins Complete" is the percentage of students who completed the test out of those who logged in. The *Total* row provides aggregate numbers, while the *Column median* row estimates what might be considered "typical" participation rates for a course.

The online participation can be contrasted with the in-class paper administration at one university, depicted in Table 7. For all three sections, the participation rate is higher than all online courses except the aforementioned Math #1.

Table 7: Participation rates for paper administration at one university

Course	**Total students**	**Participants (%)**
External #2a	51	71%
External #2b	58	91%
External #2c	29	86%
Total	*138*	*83%*

5.2 Demographics

The students in the External #2 courses are nearly all engineers (111 of 114, 97%), while the online courses include three from the Psychology Department (zero engineers). Therefore, some differences between the online and paper versions could possibly be due to the inherent differences in these student populations. Table 8 summarizes student demographics. The rates are calculated using only students who supplied the corresponding information; however, nearly all students provide full information.

Table 8: Paper vs. online demographics, for all students and only engineering majors

Group	*Medium*	*Students*	*Gender (% male)*	*Ethnicity (% white)*	*Year in school*	*Major*
All students	Paper	121	75%	75%	63% Soph 28% Jun	98% Engr
	Online	308	62%	92%	43% Fresh 18% Jun 31% Sen	63% Engr 12% Soc 8% Geo
Engr only	Paper	116	76%	77%	66% Soph 29% Jun	--
	Online	194	79%	94%	68% Fresh 16% Jun 11% Sen	--

Because the paper administrations were nearly all engineers, the demographics are essentially the same across all students or only engineers. For the online version,

more than one-third of the students were non-engineers. When non-engineers are removed, the online courses are comparable to the paper courses in terms of gender. However, noticeable differences persist for ethnicity and year in school. The largest online course (referred to as External #5) contained all freshman and was predominantly white (96%).

5.3 Reliability

Table 9 summarizes the overall and sub-scale reliability for the online and paper versions of the SCI; the online version has varying *n* for the sub-scales because the large external course did not use the Inferential items. The total reliability is not appreciably different for the two versions (0.70 online vs. 0.75 paper). However, the paper version is more reliable for the Probability (0.26 vs. 0.38), Inferential (0.30 vs. 0.51), and Graphical (0.23 vs. 0.32) sub-scales. The online version is more reliable for the Descriptive sub-scale (0.57 vs. 0.47). Only the Inferential difference might be considered large. Due to this fact and the similarity of the total reliability, there is no strong evidence that either version is more reliable than the other.

Table 9: Online vs. paper reliability (α), Fall 2005 post-test, all students

	Total	*Probability*	*Descriptive*	*Inferential*	*Graphical*
Online	0.7041 (n = 174)	0.2555 (n = 308)	0.5662 (n = 308)	0.3048 (n = 174)	0.2295 (n = 308)
Paper (all n = 121)	0.7449	0.3844	0.4743	0.5138	0.3180

When the reliability is calculated for only engineering students (Table 10), the difference between online and paper is minimal and similar to that for all students (Table 9). The Probability and Inferential sub-scale reliabilities are more pronounced in favor of the paper version, while the Descriptive and Graphical sub-scales have nearly equal reliabilities across medium.

Table 10: Online vs. paper reliability (α), Fall 2005 post-test, Engineering majors only

	Total	*Probability*	*Descriptive*	*Inferential*	*Graphical*
Online	0.6805 (n = 62)	0.1212 (n = 194)	0.4703 (n = 194)	0.1711 (n = 62)	0.2624 (n = 194)
Paper (all n = 116)	0.7338	0.3616	0.4870	0.5043	0.2916
Test ($H_0 : \alpha_{o=} \alpha_p$) ($W$, one-sided p)	1.200 (p = 0.22)	1.377 * (p = 0.03)	1.033 (p = 0.42)	1.672 * (p = 0.02)	1.041 (p = 0.41)

* significantly different at 0.05

Feldt (1969) provides a statistical test for comparing reliability from two tests. The test statistic, W (equation 1), is used to test the null hypothesis $H_0 : \rho_1 = \rho_2$, where the indices 1 and 2 refer to the different tests. W is distributed as F; the degrees of freedom adjustment is based on the number of subjects and items and is too tedious to re-produce in this space. The probability and inferential sub-tests have significantly different reliability across version based on this test.

$$W = \frac{1 - r_2}{1 - r_1} \tag{1}$$

where: W is the test statistic, distributed as F with adjusted *d.f.*
r_1 is the reliability of test 1 (paper)
r_2 is the reliability of test 2 (online)

Reliabilities of the sub-tests were scaled-up by the Spearman-Brown prophecy formula (Equation 2, below) to account for the varying lengths; a test length of 38 was used to further allow comparison with the total SCI.

$$r' = \frac{Kr}{1 + (K - 1)r} \tag{2}$$

where: r' is the adjusted reliability estimate
K is the scale-up factor (e.g., 2 for doubling the length)
r is the original reliability estimate

The adjusted reliabilities are shown in Tables 11 (all students) and 12 (engineers only); values of n are omitted as they are the same as found in Tables 9 and 10 above,

respectively. With this adjustment, the paper version has remarkably similar reliability across the sub-tests. The online test looks marginally acceptable across the four sub-scales for all students, but the Probability and Inferential sub-scales are still quite low when including only engineering majors.

Table 11: Scaled-up sub-test reliabilities, all students

	Total	*Probability*	*Descriptive*	*Inferential*	*Graphical*
Online	0.7041	0.5917	0.8185	0.6023	0.6989
Paper	0.7449	0.7250	0.7571	0.7850	0.7168

Table 12: Scaled-up sub-test reliabilities, engineering majors only

	Total	*Probability*	*Descriptive*	*Inferential*	*Graphical*
Online	0.6805	0.3680	0.7541	0.4163	0.6588
Paper	0.7338	0.7051	0.7663	0.7785	0.6908

5.4 Overall Scores

Tables 13 and 14 show the scores for the online and paper versions of the SCI for all students and only engineering majors, respectively. One concern of the online test is that graphics may not display at sufficient resolution. However, there is not a large difference in the average scores between the paper and online versions for the Graphical sub-test. The Inferential sub-test is the only portion where the difference between paper and online is large enough to merit concern. Even when non-engineers are removed, the large difference persists. Causality is difficult due to confounding with the paper administration coming predominantly from one university. The largest of three sections (n = 53) had the highest average Inferential score (63%), while the other two sections (52%, 56%) were somewhat more in line with other courses (e.g., Math #1 and Metr both 47%).

Table 13: Online vs. paper scores, Fall 2005 post-test, all students

	Total	*Probability*	*Descriptive*	*Inferential*	*Graphical*
Online	49% (n = 174)	41% (n = 308)	61% (n = 308)	44% (n = 174)	42% (n = 308)
Paper (all n = 121)	57%	47%	71%	57%	46%

Table 14: Online vs. paper scores, Fall 2005 post-test, Engineering majors only

	Total	*Probability*	*Descriptive*	*Inferential*	*Graphical*
Online	52% (n = 62)	46% (n = 194)	64% (n = 194)	43% (n = 62)	42% (n = 194)
Paper (all n = 116)	57%	48%	71%	57%	47%

The variability of scores across medium is another important consideration. Tables 15 and 16 display the overall and sub-scale standard deviations for all students and engineering majors only (n values same as previous tables). The total scores have nearly identical variance across medium, regardless of major. The differences between media are generally smaller for the engineering majors, except the Inferential sub-scale, which has the largest. Given that this sub-test also has the largest difference in average scores, it is likely that the online sample had a higher degree of guessing, which is known to be a variance-reducing influence.

Table 15: Online vs. paper standard deviation, Fall 2005 post-test, all students

	Total	*Probability*	*Descriptive*	*Inferential*	*Graphical*
Online	14.1% †	17.0%	21.9%	16.5%	19.1%
Paper	14.1%	18.1%	16.7%	19.7%	21.1%

† Even though one group did not take Inferential, this makes little difference in the standard deviation.

Table 16: Online vs. paper standard deviation, Fall 2005 post-test, Engineering majors

	Total	*Probability*	*Descriptive*	*Inferential*	*Graphical*
Online	13.5% †	16.6%	18.3%	15.7%	19.7%
Paper	13.6%	17.5%	16.5%	19.6%	20.8%

Tables 17 and 18 summarize the significance tests for differences between means and variances across medium. Normality was tested first via Anderson-Darling (H_0: *data are normal*); only the total score of engineering majors attained normality. Robust test

statistics were sought to test due to lack of normality. Levene's test was selected for comparing variance, and Wilcoxon's Signed Rank test was used to compare means.

There are no significant differences in variance across mode for engineering majors, while only the descriptive sub-test shows a significant difference for all students ($s^2_{web} > s^2_{paper}$). Difference in means are present at nearly every point, however; only the probability sub-test among engineering majors does not show a significant difference. While these results are *statistically* significant, the *practical* significance is minimal: among engineering majors, only the inferential sub-test has a difference of greater than one item (1.49), while the total score shows a difference of 1.83 items.

Table 17: Online vs. paper statistical tests, Fall 2005 post-test, all students

	Total	*Probability*	*Descriptive*	*Inferential*	*Graphical*
Normality (A-Sq, *p*)	1.213 * *< 0.005*	4.928 * *< 0.005*	3.107 * *< 0.005*	1.954 * *< 0.005*	5.936 * *< 0.005*
Variance (Levene's F, *p*)	0.383 *0.537*	0.286 *0.593*	9.333 * *0.002*	2.653 *0.104*	1.707 *0.192*
Mean (Wilcoxon Z, 2-sided *p*)	4.908 * *< 0.0001*	3.135 * *0.0017*	4.883 * *< 0.0001*	2.260 * *0.0238*	5.769 * *< 0.0001*

* significant at 0.05

Table 18: Online vs. paper statistical tests, Fall 2005 post-test, engineering majors only

	Total	*Probability*	*Descriptive*	*Inferential*	*Graphical*
Normality (A-Sq, *p*)	0.492 *0.223*	2.878 * *< 0.005*	2.058 * *< 0.005*	1.104 * *0.007*	3.814 * *< 0.005*
Variance (Levene's F, *p*)	0.002 *0.966*	3.186 *0.075*	1.317 *0.252*	0.747 *0.388*	2.428 *0.121*
Mean (Wilcoxon Z, 2-sided *p*)	3.028 * *0.0025*	1.158 *0.2467*	3.397 * *0.0007*	4.497 * *< 0.0001*	1.961 * *0.0499*

5.5 Item Scores

There is a strong relationship between percent correct at the item level between online and paper administrations (r = 0.92, all students; r = 0.90, Engr only). The summary statistics for the difference between paper and online versions (paper <minus>

online) are shown in Table 19. When adjusted for major, the differences become less pronounced but are still sizable.

Table 19: Summary statistics for difference in item difficulty (paper <minus> online)

	Minimum	*1st Q*	*Median*	*3rd Q*	*Max*	*Mean*	*St. Dev.*
All	-4.4%	+2.9%	+8.6%	+15.5%	+28.7%	+9.0%	+8.5%
Engr	-10.0%	-0.3%	+6.9%	+14.1%	+27.0%	+7.2%	+9.3%

Figures 16a (all) and 16b (Engr) further analyze the distribution of item difficulty differences. For engineers, nearly half of the items (18 of 38) favor the paper administration by less than 10%. This is still cause for concern, but potential differences due to teaching method, student demographics, and motivation are sufficiently confounding to make further interpretation impossible at this juncture.

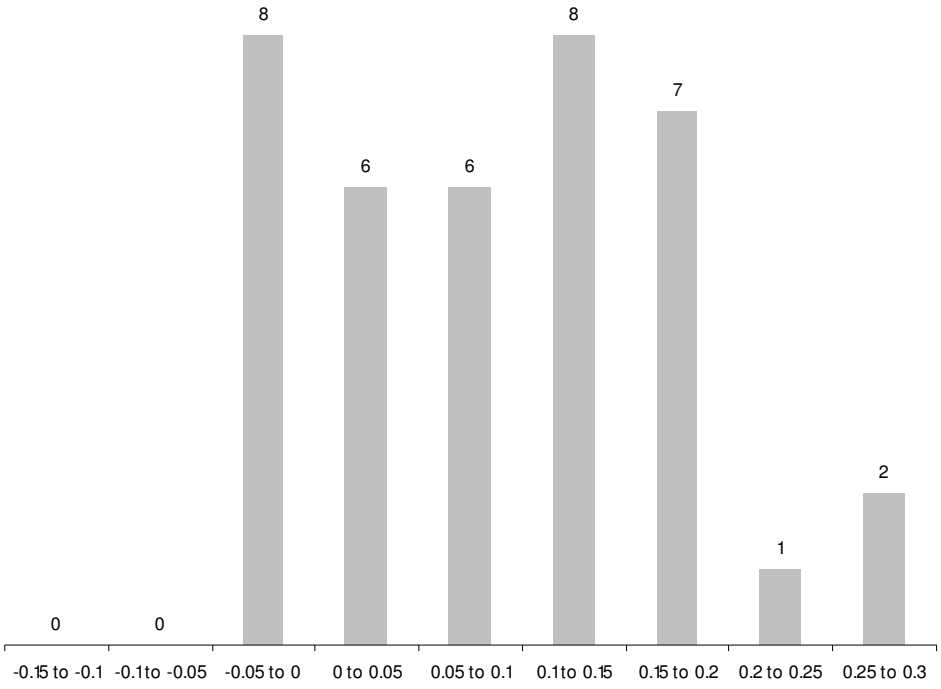


Figure 16a: Histogram of item difficulty differences (all students)

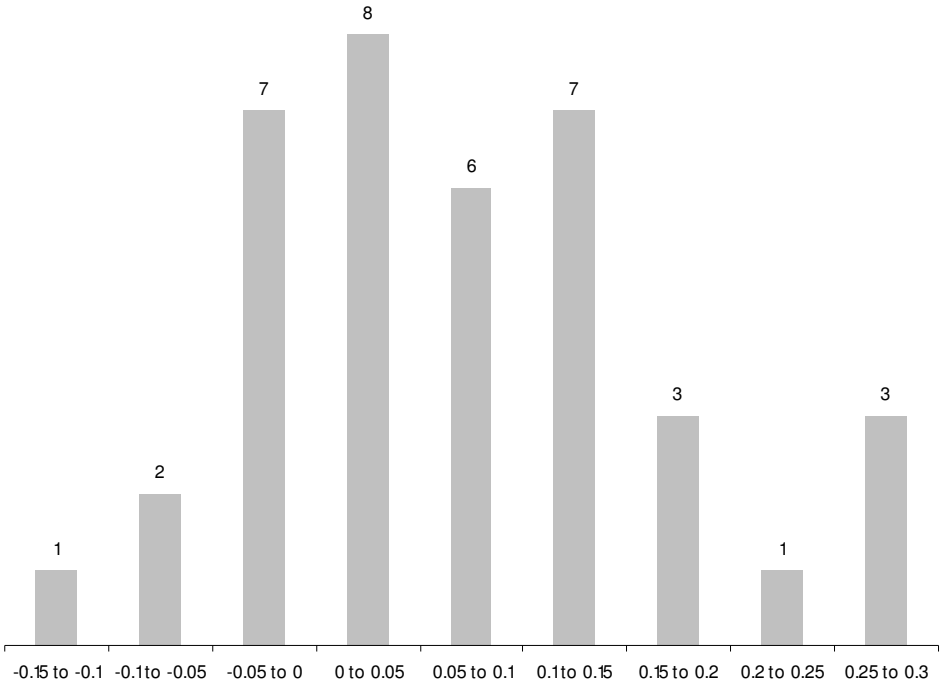


Figure 16b: Histogram of item difficulty differences (engineers only)

Figures 17a and 17b show the relationship between fraction correct on paper and web versions of the SCI, for all students (a) and engineering majors (b). The dashed line corresponds to equality (i.e, it is not a regression line). Fisher's Exact Test was conducted on a 2×2 table for each item (web/paper × right/wrong), and a significance level of 0.01 was assumed. The points marked with light gray are those which did not show a significant difference, while the larger black points correspond to items where a difference was found between the paper and web versions. Both graphs show strong and highly-significant correlations (All: $r = 0.92$, $p < 0.001$; Engr: 0.90, < 0.001).

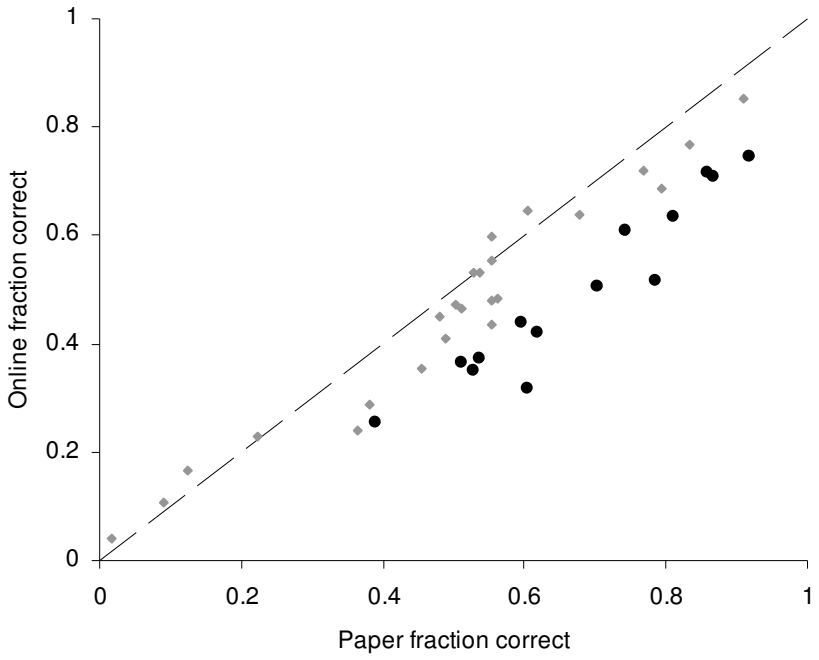


Figure 17a: Online vs. Paper fraction correct, all students

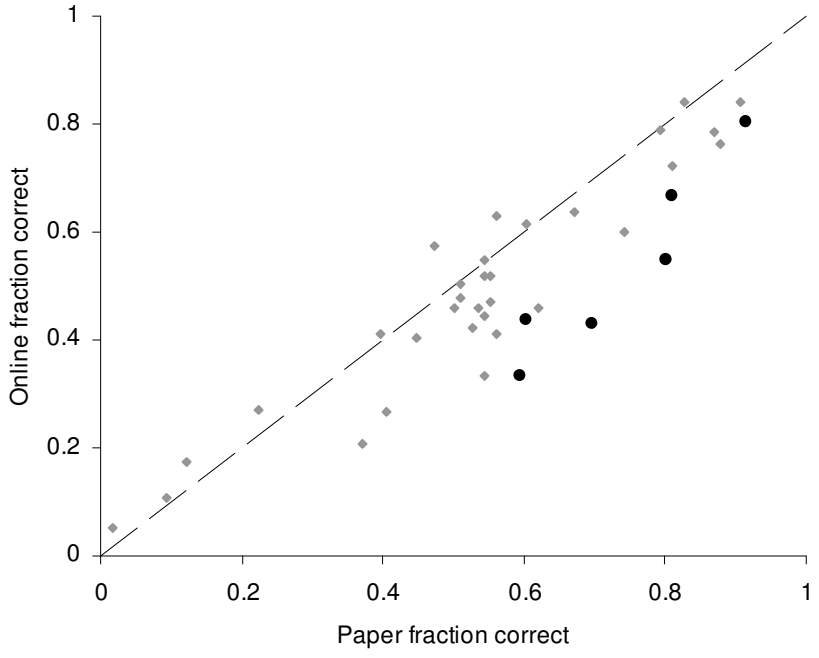


Figure 17b: Online vs. Paper fraction correct, engineering majors only

In addition to percent correct, correspondence of the response patterns is important. An example of a contingency table is shown in Table 20.

Table 20: Sample of a contingency table for response pattern comparisons

Item #3	*A*	*B*	*C*	*D*
Web	28	101	71	108
Paper	5	56	27	33

Association was tested using Fisher's Exact Test. A χ^2 test was conducted as well and results are nearly identical, but Fisher's Exact Test is more appropriate to deal with low cell counts (Agresti and Finlay, 1997), which occur on many SCI items. Table 21 (next page) shows the results of both the χ^2 and Fisher's Exact. Items marked with one star show a significant difference in response patterns across versions for all students, while items marked with two stars also show a difference among engineering majors.

Table 22 (following Table 21) summarizes the conclusions of these statistical tests. The differences are lessened when major is controlled. The most glaring inconsistency is in the overall answer patterns for engineering majors on the Inferential sub-test, with 5 of 11 items showing different response patterns across version.

Table 21: Results of degree of association tests

		All Students			Engineers		
Item	Topic	χ^2	p: χ^2	p: Fisher	χ^2	p: χ^2	p: Fisher
1	P	6.90	(0.141)	(0.111)	5.24	(0.264)	(0.262)
2 * *	I	12.16	(0.002)	(0.002)	13.17	(0.001)	(0.001)
3	D	8.78	(0.032)	(0.035)	11.38	(0.010)	(0.011)
4	P	10.50	(0.015)	(0.016)	5.39	(0.146)	(0.159)
5	P	5.56	(0.135)	(0.125)	3.50	(0.321)	(0.322)
6	D	5.86	(0.119)	(0.121)	1.81	(0.613)	(0.655)
7 * *	G	23.85	(< 0.0001)	(< 0.0001)	22.68	(< 0.0001)	(< 0.0001)
8	D	4.95	(0.176)	(0.152)	1.43	(0.699)	(0.732)
9 * *	D	18.38	(< 0.001)	(< 0.0001)	10.04	(< 0.0001)	(0.006)
10	I	8.11	(0.044)	(0.045)	7.56	(0.056)	(0.052)
11*	D	15.68	(0.001)	(0.001)	8.57	(0.036)	(0.033)
12*	D	12.34	(0.006)	(0.006)	6.66	(0.084)	(0.092)
13	P	5.34	(0.148)	(0.161)	3.59	(0.309)	(0.337)
14	G	4.40	(0.222)	(0.251)	4.12	(0.249)	(0.27)
15* *	D	11.48	(0.009)	(0.005)	12.30	(0.006)	(0.004)
16	P	4.92	(0.178)	(0.192)	4.36	(0.225)	(0.26)
17*	I	17.59	(0.001)	(< 0.001)	1.62	(0.652)	(0.657)
18* *	I	13.69	(0.003)	(0.003)	13.97	(0.003)	(0.003)
19*	I	13.51	(0.009)	(0.009)	2.35	(0.671)	(0.680)
20*	I	18.32	(< 0.001)	(< 0.001)	11.54	(0.009)	(0.011)
21* *	P	32.44	(< 0.0001)	(< 0.0001)	28.05	(< 0.0001)	(< 0.0001)
22* *	I	18.71	(< 0.001)	(< 0.001)	7.30	(0.063)	(0.065)
23	D	4.36	(0.225)	(0.262)	3.01	(0.390)	(0.423)
24*	G	13.77	(0.003)	(0.009)	12.11	(0.007)	(0.014)
25	G	11.03	(0.026)	(0.011)	10.69	(0.030)	(0.013)
26	D	6.12	(0.106)	(0.101)	3.55	(0.314)	(0.336)
27	I	3.04	(0.386)	(0.401)	2.98	(0.394)	(0.383)
28	G	6.88	(0.076)	(0.094)	3.58	(0.310)	(0.314)
29*	D	12.23	(0.007)	(0.005)	5.41	(0.144)	(0.147)
30	G	7.19	(0.066)	(0.057)	1.97	(0.579)	(0.572)
31	P	0.15	(0.014)	(0.015)	2.19	(0.533)	(0.517)
32	I	1.49	(0.685)	(0.695)	2.34	(0.505)	(0.522)
33	P	2.42	(0.490)	(0.503)	4.23	(0.238)	(0.235)
34	P	4.16	(0.245)	(0.234)	3.74	(0.291)	(0.281)
35* *	I	26.28	(< 0.0001)	(< 0.0001)	14.30	(0.003)	(0.002)
36* *	I	12.35	(0.006)	(0.006)	16.97	(0.001)	(0.001)
37	G	0.95	(0.187)	(0.864)	1.12	(0.773)	(0.805)
38	D	5.19	(0.158)	(0.151)	8.18	(0.042)	(0.045)

χ^2 d.f. is 3 for items with 4 choices [most items], except #2 [d.f.=2] and #1,19,24 [d.f.=4]

Table 22: Number of significant differences in tests of association (correct only, and all answers) for all majors and engineers only, grouped by sub-test

	Correct		*Answers*	
	All	Engr	All	Engr
Probability	2	1	1	1
Descriptive	5	2	5	2
Inferential	5	2	8	5
Graphical	2	1	2	1
Total	*14*	*6*	*16*	*9*

5.6 Completion time

Table 23 provides summary statistics for the completion time. Only students who completed the SCI during one login are included (n = 294). Completion time is positively skewed, as can be seen in the histogram of completion times (Figure 18). Nearly one-half of the students complete the SCI in 15 to 25 minutes.

Table 23: Summary statistics for online completion time (minutes)

Minimum	*1st Q*	*Median*	*3rd Q*	*Max*	*Mean*	*St. Dev.*
3.37	16.78	21.47	30.25	84.00	24.15	11.28

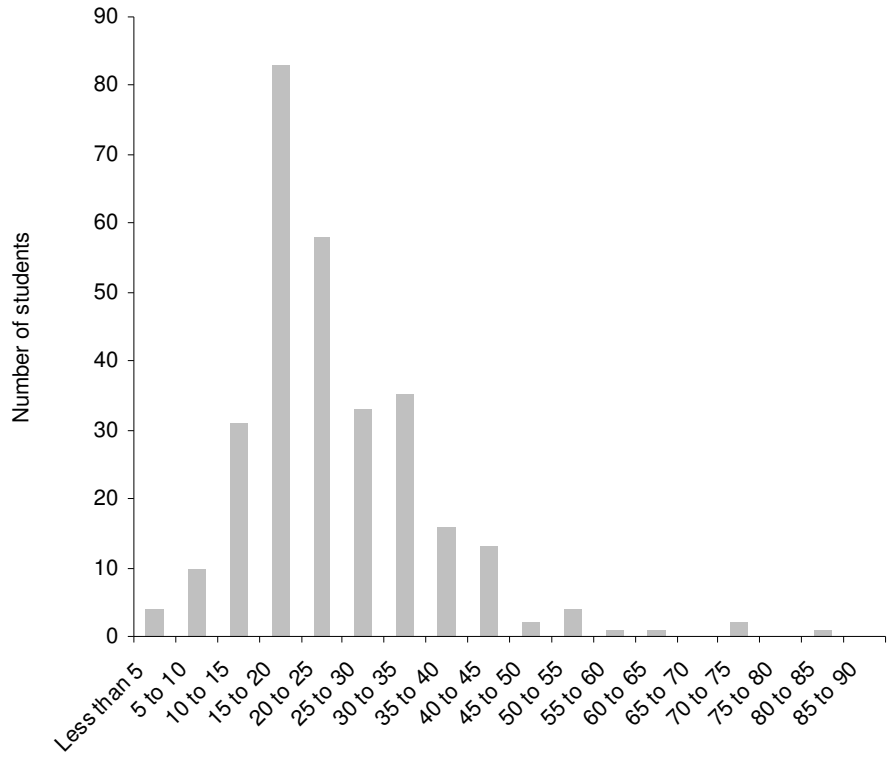


Figure 18: Histogram of completion time (minutes)

Figures 19a and 19b show number correct vs. completion time on two scales. Number correct is used rather than percent to account for the external section answering fewer questions (i.e., it is possible to achieve a higher percent correct during a lesser time).

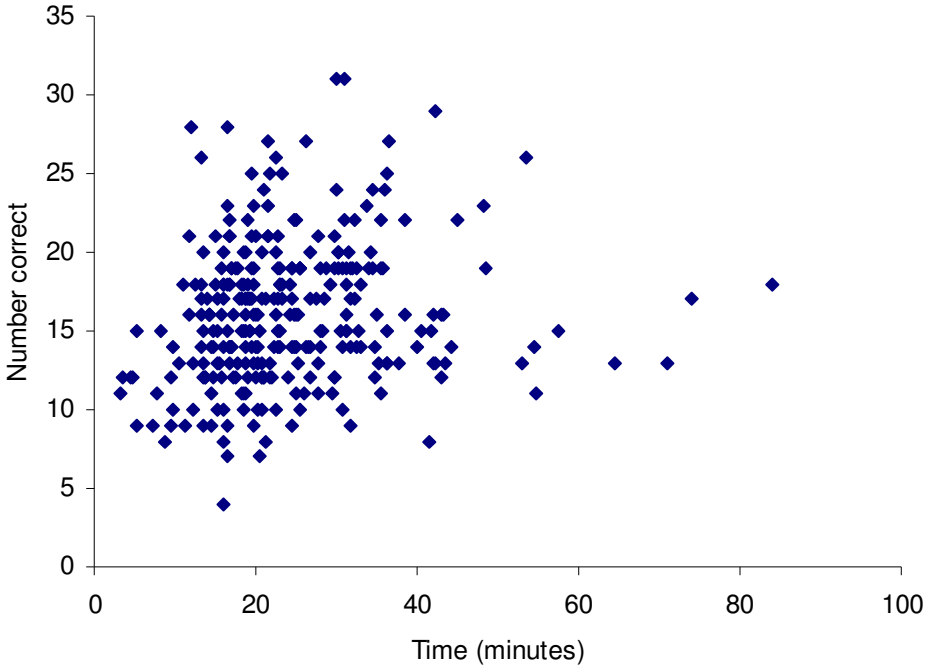


Figure 19a: Number correct vs. completion time, full scale

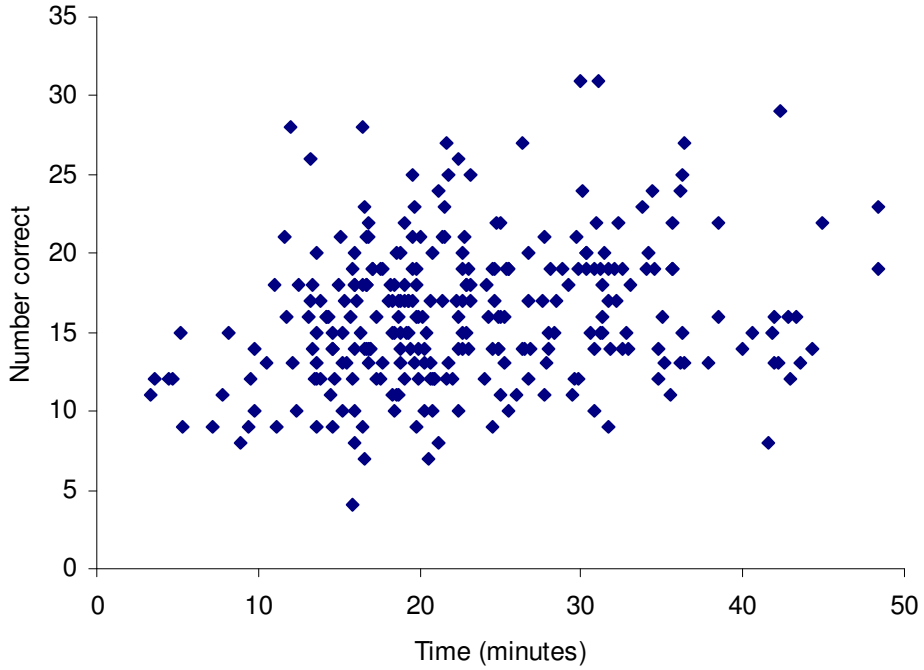


Figure 19b: Number correct vs. completion time, restricted Time scale

A slight positive trend is found between number correct and completion time (r = 0.160, p = 0.003). The correlation is highly significantly, in part due to the large sample size. Number correct suffers from restriction of range due to a hypothetical guessing percentage of around 25% correct. The pattern most closely resembles a funnel.

5.7 Order effects

Question order

Questions are randomly ordered for each student. As such, a question should have an equal probability of occurring at any position in the order (i.e., uniform distribution). Additionally, each order position should contain an equal proportion of each question (within sampling error). These two situations are depicted graphically in Figures 20 and 21, for item number 1 and order position 1, respectively.

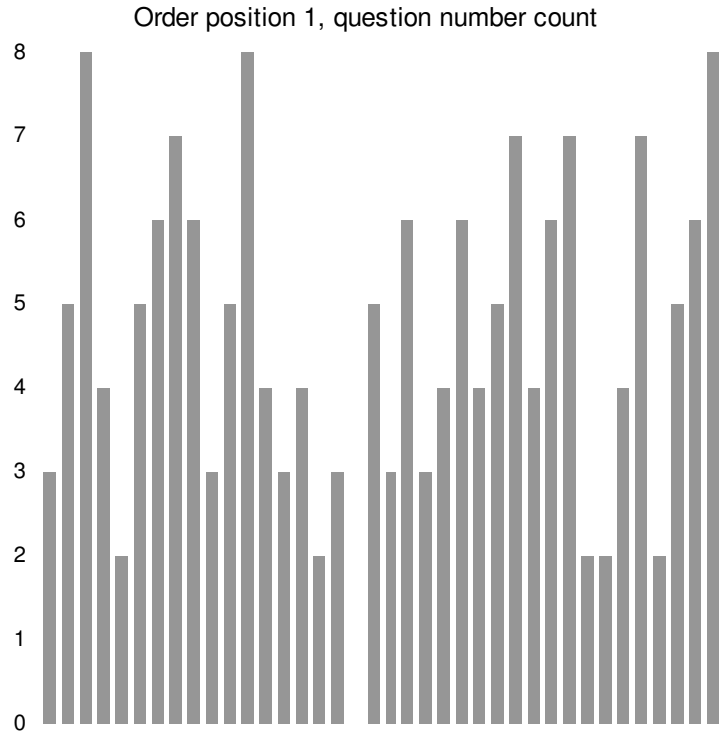


Figure 20: Sample histogram of question counts at position 1
(e.g., position 1 had question #1 three times, #2 five times, etc.)

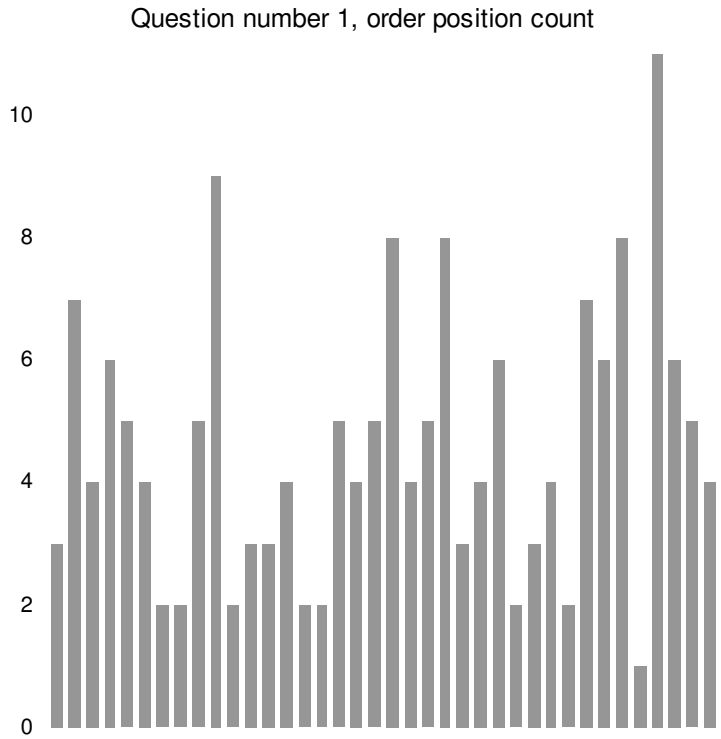


Figure 21: Sample histogram of order counts for question #1
(e.g., question #1 had position one 3 times, position two 7 times, etc.)

The hypothesis of a uniform distribution was tested by a χ^2 goodness-of-fit using `PROC FREQ` in SAS. Only the students who took the entire SCI were included (n = 174), but there is no reason to believe these results are not generalizable. Seventy-six statistical tests were conducted (38 items, 38 positions). Only three of these yielded a p-value to reject the hypothesis of a uniform distribution at a significance-level of 0.05: question 27 ($\chi^2_{37} = 52.7$, $p = 0.0454$), question 38 ($\chi^2_{37} = 55.3$, $p = 0.0269$), position 23 ($\chi^2_{34} = 51.3$, $p = 0.0289$). Given the large number of tests, these results are sufficient to conclude that the random selection process did not bias the order in any way.

Difficulty

Figure 22 shows the mean item difficulty as a function of order position. The dark points are for students who took the entire SCI (n = 174), while the lighter gray

corresponds to the group who did not take the inferential sub-test (n = 134). The dashed lines represent the median difficulty for each group. In neither case is a trend present. Statistically, the effect is essentially zero (r^2 = 0.0006, full; r^2 = 0.0003, partial), nor is any pattern apparent.

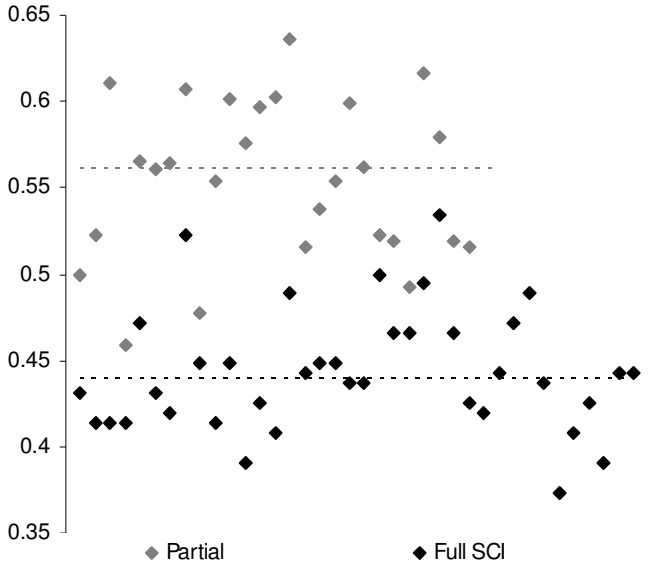


Figure 22: Fraction correct versus order position
[dashed lines are group medians]

Confidence

Figure 23 depicts mean item confidence data in a format analogous to the difficulty shown previously. A downward trend is apparent for both groups. This is detected statistically as well (r^2 = 0.523, full; r^2 = 0.405, partial; both $p < 0.0001$).

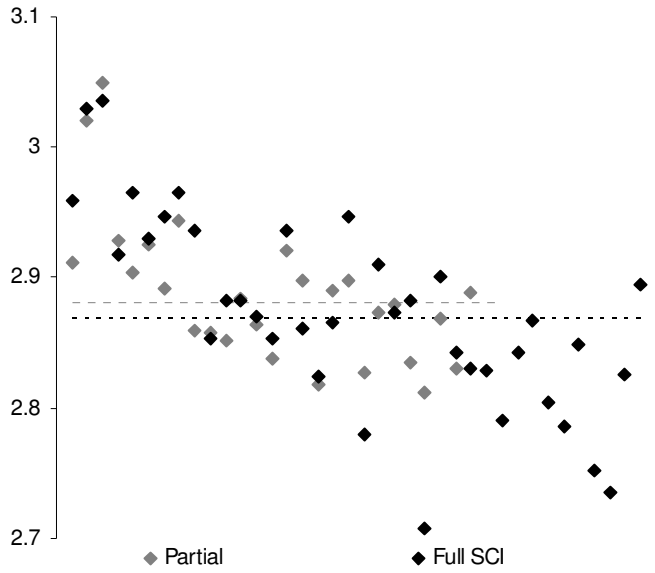


Figure 23: Mean item confidence versus order position

Considering the magnitude of the decrease, the results are less ominous: the slope of -0.0048 (full) or -0.0044 (partial) corresponds to a decrease in confidence of 0.1824 (full, 38 items) or 0.1188 (partial, 27 items) from beginning to end. With the confidence scale ranging from 1 to 4, these values represent 6% and 4% of the total scale, respectively. The previous graphic was oriented to show maximal spread of the confidence values. When the means are viewed across the range of possible values (Figure 24), the decline is minimal. Further, these graphics present the *mean* confidence at each position; when individual points are used, the model parameters are unchanged, but the correlation is meaningless ($r^2 < 0.0005$). Therefore, the decline in confidence is not so over-whelming that it should cause concern, but it should be monitored.

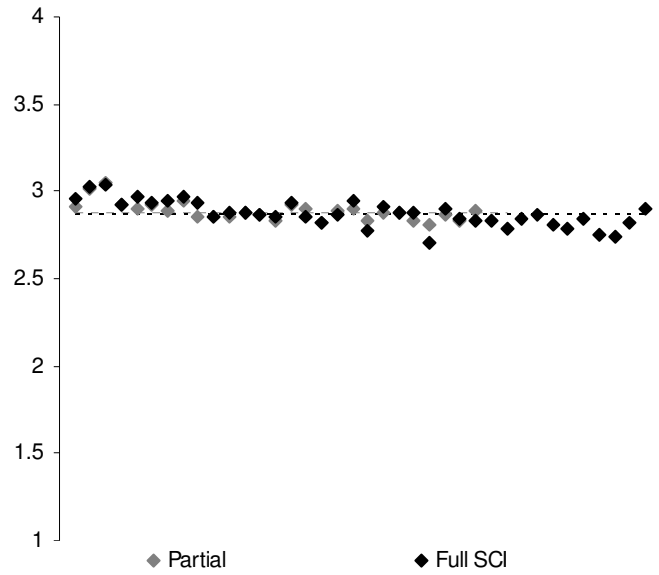


Figure 24: Mean item confidence versus order position (full confidence scale)

5.8 Student feedback

A small percentage of students (71 of 308, 23%) elect to use the feedback box at the end of the SCI. Feedback is generally positive. A few representative comments are presented below.

Positive

- "The online way is much better then [*sic*] doing it on paper."
- "much better than written"
- "I encountered no errors."
- "No errors, well layed [*sic*] out."
- "I thought that asking for a confidence level was a very useful tool for your survey."
- "This assesment [*sic*] does a good job of testing to see what knowledge I actually do posess [*sic*] about statistics. I thought I understood it fairly well, but this made me think twice on most of the questions"
- "I really liked being able to take this test online."

Negative

- "It was very long and more tedious then [*sic*] i [*sic*] expected."
- "I DONT LIKE THE FACT THAT YOU HAVE TO SUBMIT, THEN RATE YOUR CONFIDENCE AFTER THE PAGE REFRESHES. THIS TAKES TIME, AND SHOULD ALL BE DONE ON ONE PAGE." [*sic*: capitalization]

Other

- "Questions were confusing even if you know basic statistics. So if you were trying to be tricky, you succeeded magnificantly [*sic*]."
- "The test would have been easier had I had a more focused, organized stats teacher."

5.9 Benefits of the online test

A benefit of the online test is the reduction of processing time. With paper, it takes approximately one minute per student to input results. This includes manual coding, as well as data entry. The only data entry for the online test are the student names and email addresses, which takes approximately 15 seconds per student. Data processing is as simple as copying-and-pasting the results into Excel. Additional data is collected in the form of answer confidence and testing time, which allows an extra level of analysis. On a personal note, programming in PHP and learning mySQL has proven to be a valuable and oftentimes enjoyable challenge.

6. Preliminary Conclusions

Cole, *et al.* (2001) found nearly exact correspondence between web and paper versions of the FCI. However, that study was explicitly designed to compare the two versions: it took place at one university and balancing efforts were possible. The SCI online test is designed to increase the number of universities who can easily participate and to ease the data processing. It may even be viewed positively that less than one-fourth of the SCI items demonstrated divergent response patterns across versions when major was controlled, in spite of the still-obvious differences in participation rates and demographics; differences in motivation (student *and* instructor) are possible as well, but the evidence along these lines is anecdotal.

7. A Controlled Study

The results from Fall 2005 suggest the online version is not vastly different from the paper version. However, it was not possible to draw rigorous conclusions due to different populations. For Spring 2006, an opportunity was available to perform a more

controlled comparison: the same professor teaching two sections of the same course, with one taking the online SCI and one taking the paper SCI. Combined with the larger sample size of Fall 2005 which provides more powerful statistical comparisons, the combined analysis of both studies should allow more generalizable conclusions.

The data for Spring 2006 are for the pre-test, which is an acknowledged difference from the Fall 2005 data. The online version had the confidence portion removed because it is unclear if affects students' reasoning skills. The data was collected during the last half of the class periods on the first day of the second week of classes; this time was preferred by the instructor due to enrollment flux typical of the first week.

7.1 Participation Rate

Table 24 shows the participation rates for the two courses, with total students determined from the class roster obtained from the instructor. Although such low participation is not ideal, it is typical of the rates found for Fall 2005 (average completion 49%; median participation 57%; Table 6). All participants included in this analysis completed the entire SCI; two students who took the online version at a later time outside of class are not included.

Table 24: Participation, Spring 2006

Course	**Total students**	**Participants**
Online	31	14
Paper	32	16

7.2 Demographics

Tables 25 and 26 summarize the participant demographics from the paper and online versions; the numbers are provided as counts rather than percents due to the small sample size. There are no differences of practical significance among class characteristics (Table 25). The online participants showed a higher level of statistics experience (Table

26), as fewer students indicated no level of prior experience; slightly more of these students also have completed at least one course beyond calculus.

Table 25: Participant demographics, part 1

	Students	*Gender (males)*	*Ethnicity (whites)*	*Year in school (seniors)*	*Major (engineers)*
Online	14	10	9	9	8
Paper	16	11	10	9	10

Table 26: Participant demographics, part 2

	Students	*Statistics (no experience)*	*Mathematics (all calculus)*	*Mathematics (at least one course beyond calculus)*
Online	14	4	13	13
Paper	16	8	13	11

7.3 Reliability

Tables 27 and 28 display the reliability based on the raw scores and also a scaled-up version to allow comparison across sub-tests. Two sub-tests had reliabilities which calculated as negative. Because this is not theoretically meaningful, they are reported as zero; a value of zero was used in the significance test as well.

Table 27: Reliability of online and paper versions of the SCI, with significance test

	Total	*Probability*	*Descriptive*	*Inferential*	*Graphical*
Online (α)	0.7093	0.1002	0.4020	0.2350	0 †
Paper (α)	0.5439	0 †	0.3379	0.3723	0.0260
Test	0.637	0.900	0.903	1.219	1.027
(W, one-sided p)	*p = 0.22*	*p = 0.44*	*p = 0.44*	*p = 0.38*	*p = 0.49*

† calculated as negative but set to zero

Table 28: Reliability of online and paper versions of the SCI, scaled-up to $k = 38$

	Total	*Probability*	*Descriptive*	*Inferential*	*Graphical*
Online	0.7093	0.3198	0.6990	0.5148	0 †
Paper	0.5439	0 †	0.6381	0.6720	0.1266

† calculated as negative but set to zero

The statistical tests on α reveal no significant differences. With such small sample sizes, this is not surprising. For example, the reliability values for the full SCI would require sample sizes of approximately 65 each for a significant difference at 0.05, all else

being constant. The scaled-up reliabilities show that the probability and graphical sub-scales are unreliable for both versions, while the descriptive and inferential sub-scales are not appreciably different from the total reliability and at more acceptable levels, excepting perhaps the online inferential.

7.4 Overall Scores

Tables 29, 30, and 31 summarize the total and sub-test scores for the two SCI versions. The total, probability, and descriptive scores appear equivalent, although probability has a significantly smaller variance for the paper version. The inferential and graphical scores, interestingly, are nearly reversed across versions. The difference is statistically significant for the inferential scores and marginally so for graphical.

Table 29: Mean percent correct, across version

	Total	*Probability*	*Descriptive*	*Inferential*	*Graphical*
Online	44.4%	41.3%	57.1%	32.5%	46.9%
Paper	43.4%	37.5%	56.8%	43.4%	31.3%

Table 30: Standard deviation of percent correct, across version

	Total	*Probability*	*Descriptive*	*Inferential*	*Graphical*
Online	13.4%	16.6%	17.6%	14.6%	16.3%
Paper	10.9%	8.9%	18.0%	17.0%	17.4%

Table 31: Significance tests for differences between versions

	Total	*Probability*	*Descriptive*	*Inferential*	*Graphical*
Normality (A-Sq, *p*)	0.304 *p* > 0.250	1.017 * *p = 0.010*	0.496 *p = 0.207*	0.713 *p = 0.058*	1.245 * *p < 0.005*
Variance (Levene's F, *p*)	0.279 *p = 0.601*	4.653 * *p = 0.040*	0.043 *p = 0.837*	0.098 *p = 0.756*	0.219 *p = 0.643*
Mean (Wilcoxon Z, 2-sided *p*)	0.0626 *p = 0.950*	0.7355 *p = 0.462*	0.3588 *p = 0.7198*	2.3611 * *p = 0.0182*	1.9120 *p = 0.0559*

7.5 Item Scores

Figure 25 compares online and paper fraction correct. The items which show a significant difference in response patterns are marked in black, while the smaller gray points represent items which did not show a significant difference; a liberal significance

level of 0.10 was assumed due to the small samples sizes. The correlation is of lower magnitude than those from Fall 2005 ($r > 0.90$) but still highly-significant ($r = 0.673$, $p < 0.001$).

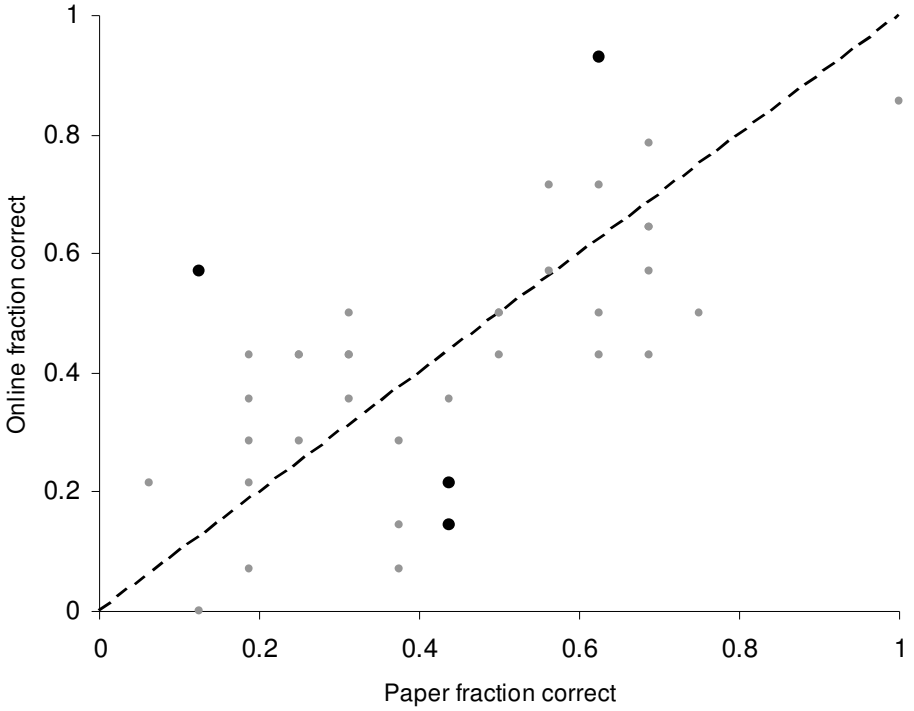


Figure 25: Online vs. Paper fraction correct

Table 32 (next page) summarizes the contingency table tests on the response patterns (letters). Items which showed significant differences at 0.10 are marked with *; items marked † showed a significant difference in correct/incorrect patterns. In contrast to Fall 2005, these additional markings allow as many differences as possible to be considered significant, to account for the low power associated with this study.

Table 32: Results of degree of association tests

Item	Topic	χ^2	p: χ^2	p: Fisher
1 *	P	11.85	(0.018)	(0.007)
2	I	3.55	(0.170)	(0.211)
3 *	D	7.01	(0.072)	(0.064)
4	P	3.08	(0.379)	(0.431)
5	P	2.88	(0.411)	(0.193)
6	D	3.02	(0.388)	(0.443)
7	G	1.25	(0.741)	(0.783)
8	D	3.88	(0.274)	(0.358)
9 †	D	4.08	(0.253)	(0.140)
10	I	5.17	(0.160)	(0.187)
11	D	0.60	(0.104)	(0.924)
12	D	0.54	(0.089)	(0.822)
13	P	4.39	(0.223)	(0.338)
14	G	1.41	(0.704)	(0.924)
15	D	1.60	(0.658)	(0.746)
16	P	0.02	(0.001)	(1.000)
17	I	7.95	(0.047)	(0.049)
18 *	I	8.27	(0.041)	(0.050)
19 †	I	5.02	(0.285)	(0.255)
20	I	2.74	(0.434)	(0.452)
21	P	2.59	(0.459)	(0.257)
22	I	6.26	(0.100)	(0.112)
23	D	0.27	(0.034)	(1.000)
24	G	3.01	(0.556)	(0.682)
25	G	4.64	(0.200)	(0.245)
26	D	2.93	(0.402)	(0.483)
27	I	2.45	(0.485)	(0.209)
28 *	G	3.85	(0.279)	(0.086)
29	D	1.21	(0.75)	(0.614)
30	G	2.08	(0.556)	(0.418)
31	P	6.03	(0.110)	(0.133)
32	I	1.59	(0.661)	(0.761)
33	P	4.58	(0.205)	(0.261)
34	P	3.88	(0.274)	(0.300)
35	I	1.44	(0.696)	(0.848)
36	I	2.78	(0.427)	(0.486)
37	G	2.50	(0.475)	(0.505)
38	D	0.70	(0.127)	(1.000)

χ^2 d.f. is 3 for items with 4 choices [most items], except #2 [d.f.=2] and #1,19,24 [d.f.=4]

7.6 Completion time

The completion time was not explicitly recorded for the paper test. However, the small class size allowed the tests to be collected in the order that students finished. This

allows a rank-order correlation to be performed, with comparison to the online version. Figure 26 shows this relationship.

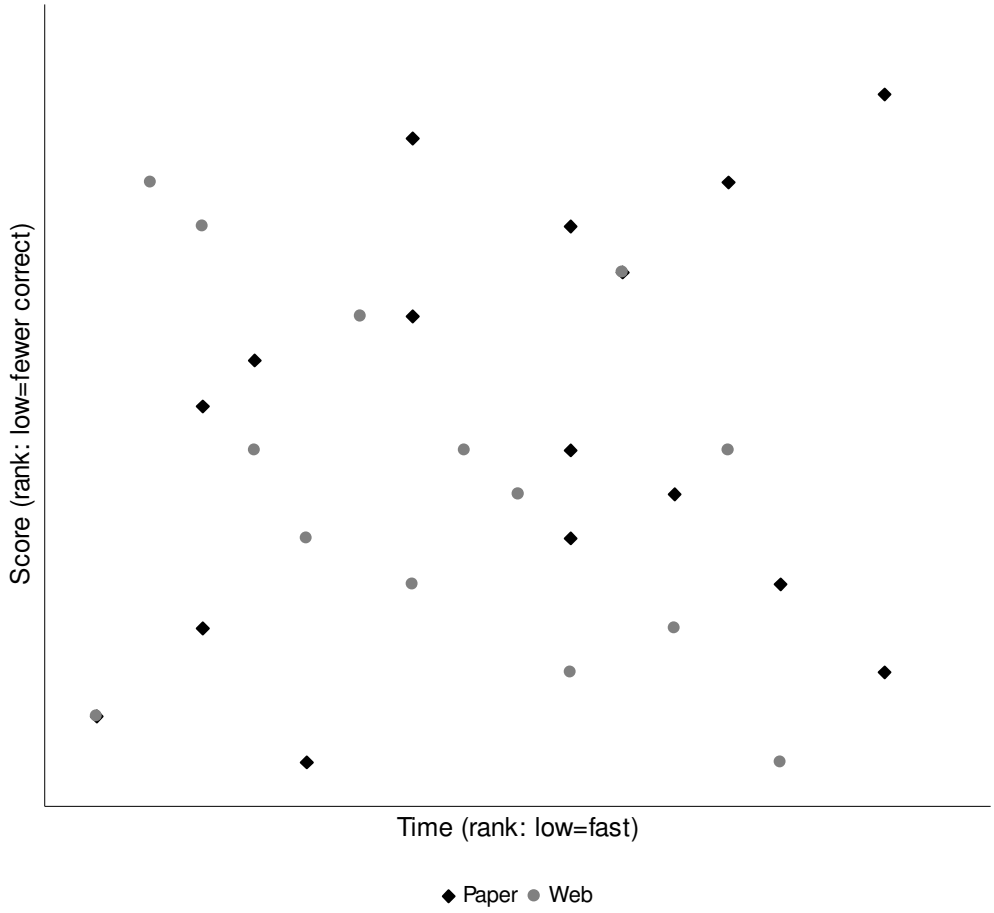


Figure 26: Score vs. Time, rank-order

While no strong trend is evident, it is generally observable that the paper version slopes upward, while the online version slopes downward. These correlations are nearly equal in magnitude as well (paper: $r = +0.30$, $p = 0.13$; online: -0.34, 0.12). This suggests possibly differing strategies across versions: "taking your time" seems to indicate thoughtful responses on the paper version, whereas a more instinctual speeded response pattern is observed on the online version. When the correlation is performed on the raw data for the online test, the correlation is slightly stronger ($r = -0.40$, $p = 0.08$). However,

owing in part to the small sample sizes, none of the correlations are significantly different from zero. The pattern differs from Fall 2005, which saw a slight positive trend between percent correct and completion time ($r = 0.16$).

8. Synthesis and Comparison

Table 33 summarizes the significant differences for the online and paper versions from the Fall 2005 (engineers only) and Spring 2006 studies. Only in three places do significant differences occur in both studies: mean scores from the inferential sub-test; and response patterns for items #9 and #18.

Table 33: Summary of significant differences across semesters

Measure		Fall 2005	Spring 2006
Reliability	Probability	×	
	Inferential	×	
Mean	Total	×	
	Descriptive	×	
	Inferential	×	×
	Graphical	×	
Variance	Probability		×
Items	#1 (Probability)		×
	#2 (Inferential)	×	
	#3 (Descriptive)		×
	#7 (Graphical)	×	
	#9 (Descriptive)	×	×
	#15 (Descriptive)	×	
	#18 (Inferential)	×	×
	#19 (Inferential)		×
	#21 (Probability)	×	
	#22 (Inferential)	×	
	#28 (Graphical)		×
	#35 (Inferential)	×	
	#36 (Inferential)	×	

Clearly, the data from both semesters are not ideal: Fall 2005 had confounding with university, while Spring 2006 had a small sample size and was from a pre-test. Given these limitations, there is still no reason to believe the online and paper versions

are different. The best way to draw firm conclusions would be to test two large sections (e.g. 50 students each), taught by one instructor, as a post-test. Finding such an opportunity may prove exceedingly difficult. To put this work in perspective, the online FCI study by Cole, *et al.* (2001) was published 16 years after the first FCI article (Halloun and Hestenes, 1985). Meanwhile, this SCI comparison was conducted less four years after the beginning of the project (Fall 2002 → Spring 2006). Introductory physics is also often taught in large sections ($n > 100$), whereas the SCI is typically administered to classes with around 30 students and rarely approaching 100.

8.1 Corollary: The Problem with Educational Research

This chapter highlights a predicament in educational research, which is pictorially approximated below with a linearity assumption (Figure 27). This, of course, is a simplification. Given adequate time and resources, a large-scale, controlled study could be conducted. The connectedness and marketing capabilities of the researchers is likely to enhance the research as well. Time constraints often stem from deadlines imposed by research grants and the associated pressure to produce publishable results.

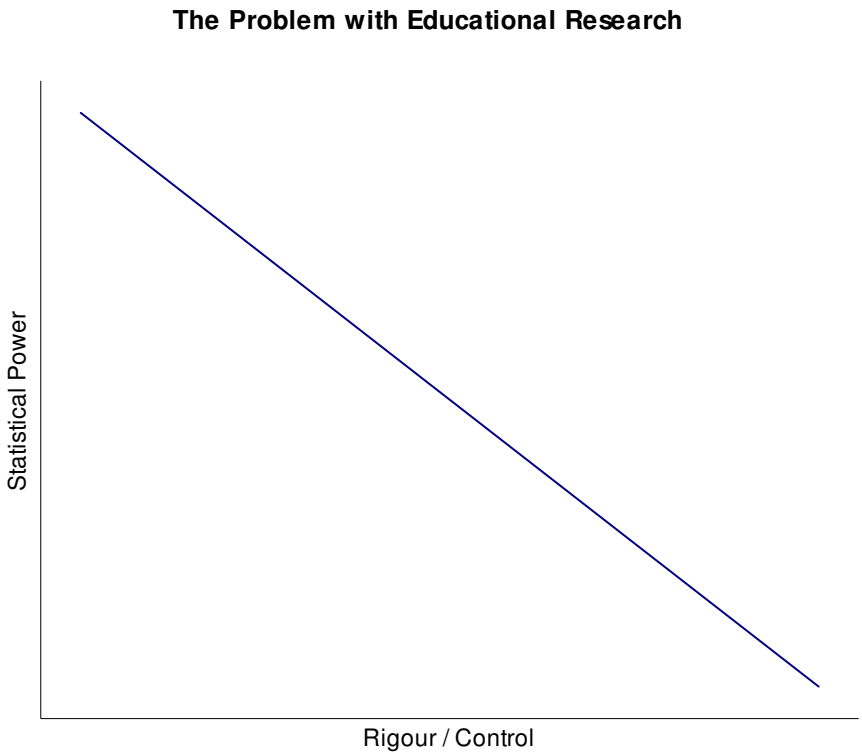


Figure 27: The Problem with Educational Research

References

Agresti, A., and B. Finlay. 1997. Statistical Methods for the Social Sciences. 3rd ed. Prentice-Hall, Inc.: Upper Saddle River, NJ.

Clariana, R., and P. Wallace. 2002. Paper-based versus computer-based assessment: key factors associated with the test mode effect. *British Journal of Educational Technology*. 33 (5): 593-602.

Cole, R.P., D. MacIsaac, and D.M. Cole. 2001. A Comparison of Paper-Based and Web-Based Testing. *Annual Meeting of the American Educational Research Association*. ERIC Document ED453224.

Feldt, L. S. 1969. A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. *Psychometrika*. 34: 363-373.

Goldberg, A.J., and J.J. Pedulla. 2002. Performance Differences According to Test Mode and Computer Familiarity on a Practice Graduate Record Exam. *Educational and Psychological Measurement*. 62 (6, December): 1053-1067.

Halloun, I. and D. Hestenes. 1985. The initial knowledge state of college physics students. *American Journal of Physics*. 53 (11): 1043-1055.

Mead, A.D., and F. Drasgow. 1993. Equivalence of Computerized and Paper-and-Pencil Cognitive Ability Tests: A Meta-Analysis. *Psychological Bulletin*. 114 (3): 449-458.

Neuman, G., and R. Baydoun. 1998. Computerization of Paper-and-Pencil Tests: When Are They Equivalent? *Applied Psychological Measurement*. 22 (1, March): 71-83.

Spray, J.A., T.A. Ackerman, M.D. Recakse, and J.E. Carlson. 1989. Effect of the Medium of Item Presentation on Examinee Performance and Item Characteristics. *Journal of Educational Measurement*. 26 (3, Autumn): 261-271.

CHAPTER VIII

Self Efficacy of Statistical Reasoning Skills

Abstract

This chapter contains two portions. The first is a literature review of statistics and probability reasoning skills, both affective and cognitive, with discussion of teaching strategies that may alleviate incorrect heuristics. The second portion utilizes the SCI to identify topics where students have conceptual difficulties. This is accomplished by a confidence rating scale for the online SCI. A general positive trend was found between percent correct and answer confidence, but many items were identified which point to either over- or under-confidence, perhaps pointing to misconceptions (over) and correct but incomplete heuristics (under).

A. Literature Review

1. Difficulties

Watts (1991) highlighted several reasons why students have difficulties in introductory statistics. Statistics often lacks a visual representation of its abstract concepts. In Calculus, for example, a derivative or integral has a graphical representation, whereas a random variable, for instance, might best be described as "the value of the *next* observation in an experiment" (p. 290). Additionally, the inherent randomness of statistical processes contrasts with a deterministic world-view: there is often not a "correct" answer and results are open to interpretation.

Murtonen and Lehtinen (2003) surveyed students to assess the perceived difficulty of statistics relative to other subjects. Education and sociology graduate

students enrolled in a statistical methodology course participated in the study; all students had previously completed an introductory statistics course. Subjects were asked to rate 11 topics from -5 to +5 along two dimensions: easy-difficult and concrete-abstract. The results are summarized in the Table 1 (next page).

Figure 1 (reproduced from Murtonen and Lehtinen, 2003) demonstrates the relationship between difficulty and abstractness. Mathematics (topic #1) and statistics (#2) are viewed as the most abstract, along with statistical significance tests (#9). The quantitative topics (#1, 2, 7, 8, 9, 10) are viewed as the most difficult. Foreign languages (#3) and major subjects (#5, 6) were included to provide a comparison point. As anticipated, these topics were rated as less difficult and less abstract than the quantitative topics.

Table 1: Difficulty and Abstractness of statistics and other educational topics

Topic	Difficulty	Abstractness
1. Mathematics in general	+0.2	+1.1
2. Statistics in general	+1.8	+1.4
3. Foreign languages	-1.0	-1.3
4. Research and methodology in general	-0.1	-0.1
5. Introductory course of the student's major subject	-2.6	-0.2
6. Student's major subject without methodology studies	-1.9	-0.4
7. Use of statistical programmes with the computer	+1.0	+0.0
8. Statistical parameters (e.g. mean and standard deviation)	+0.1	-0.3
9. The statistical test for significance (e.g. t-test)	+2.0	+1.1
10. Quantitative research methods	+0.6	-0.5
11. Qualitative research methods	-0.4	-0.7

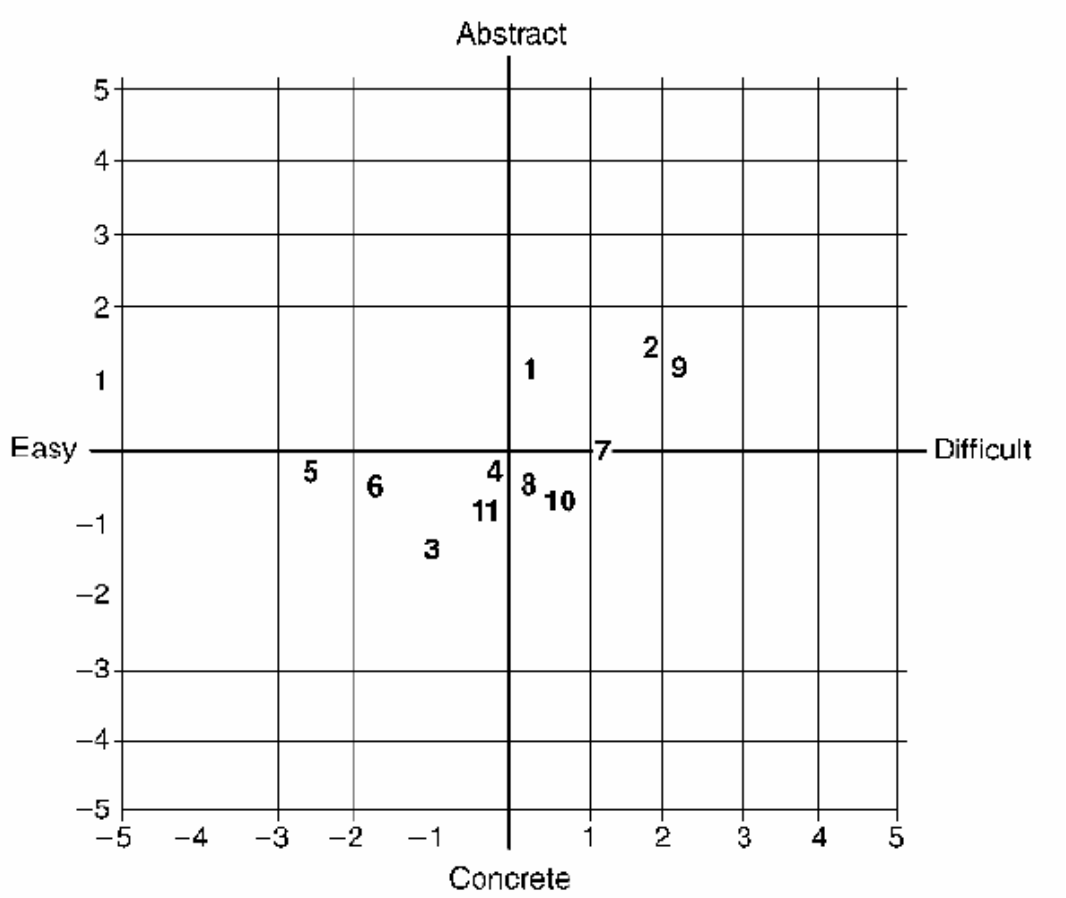


Figure 1: Relationship between Difficulty and Abstractness (keyed to Table 1)

The majority of students also documented their struggles in a journal. The researchers identified five common causes of difficulties (listed in order of prominence): superficial teaching, failure to link theory and practice, unfamiliarity with and difficulty of the content, trouble integrating various aspects of scientific research, and negative attitudes towards the subject.

2. Attitudes

A number of studies have examined affective aspects of statistics education. Baloğlu (2003) investigated differences in terms of statistics anxiety as a function of gender, age (divided into young-middle-old), and previous mathematics experience (PME). Anxiety was measured using the Statistical Anxiety Rating Scale (STARS). This

is a 5-point Likert-style instrument containing 51 items grouped along six dimensions: Worth of Statistics, Interpretation Anxiety, Test and Class Anxiety, Computational Self-concept, Fear of Asking for Help, and Fear of Statistics Teachers. The participants were 246 college students (183 women), ranging in age from 18 to 57 (mean 27.15), most of whom were juniors (74) or seniors (94), and primarily enrolled as social science majors (71.1%). A between-subjects MANCOVA was used to analyze the results.

PME was found to have an overall significant effect on statistics anxiety. Specifically, there were significant effects ($p < 0.01$) for three STARS sub-scales: Worth of Statistics, Interpretation Anxiety, and Computational Self-concept. The gender main effect and gender × age interaction were not significant. Age group had a significant main effect: Worth of Statistics was significantly higher for young students, whereas Test and Class Anxiety was higher for old students.

Rhoads and Hubele (2000) investigated differences in student attitudes before and after a computer-integrated introductory statistics course. The course differed from the lecture blueprint through its emphasis on collaborative problem solving including design, analysis, and synthesis of data from in-class activities. The Attitudes Towards Statistics (ATS) instrument was used to gather data, which has sub-scales for course and field aspects; sixty-one students provided usable data on both the pre- and post-survey.

The broad hypothesis that the course would enhance students' attitudes was not confirmed, but several interesting contrasts were found when within demographics. On the pre-survey, students who owned computers (73% of $n = 61$) had more positive attitudes towards the course, while males (71%) were more apt to hold positive attitudes towards the field of statistics. On the post-survey, students majoring in industrial,

manufacturing, and civil engineering (30%) held more-positive attitudes towards both the course and the field compared to other majors, perhaps because these majors traditionally require statistics coursework and perceive more value. Students with previous statistics exposure (38%) viewed the course more positively at the end.

Schacht and Stewart (1990) conducted a small study on the effectiveness of humor in relieving students' statistics anxiety. To avoid offending or embarrassing students and to provide a measure of control, cartoons were utilized and data were often created which loosely related to the drawing. To ensure familiarity with the assessment task, students in two sections rated the effectiveness of the cartoon technique on a 0 to 4 scale which is the standard format for rating courses at that university. The technique received high ratings for anxiety reduction (mean 3.78 and 3.67 for two sections) and creating a positive learning environment (3.74, 3.60). However, the perceived enhancing understanding (2.86, 3.12) and fostering retention (2.83, 2.88) were less encouraging. On a mathematics anxiety scale administered pre-post, a significant decrease in anxiety was found ($p < 0.005$). Student comments indicate that the instructor created a relaxed, comfortable learning environment. The ability to foster such a situation, through cartoons or other means, seems to be the greatest lesson of this less-than-rigorous research.

3. Cognitive abilities: Probability

Garfield and Alhgren (1988) reviewed a number of articles relating to students' difficulties learning basic statistics and probability concepts. Instructors have long recognized that most college students fail to attain basic understanding and consequently fall into a rote "number-crunching" mode; these abilities quickly atrophy upon completing a course.

Probability concepts prove difficult to learn for three reasons: inadequate skills to deal with rational number concepts and proportional reasoning; teaching which is too formal and abstract; and concepts which conflict with how students view the world. The latter point is most relevant.

The renowned cognitive psychologist Jean Piaget (1951, with Inhelder) performed numerous studies of probabilistic reasoning in pre-adolescent children. One experiment consisted of allowing two sets of marbles, each a different color and initially separate, to roll from one end, bounce back, and recollect into two partitions; the device is re-produced in Figure 2. As the rolls were repeated, the colors became mixed in approximately equal proportions.

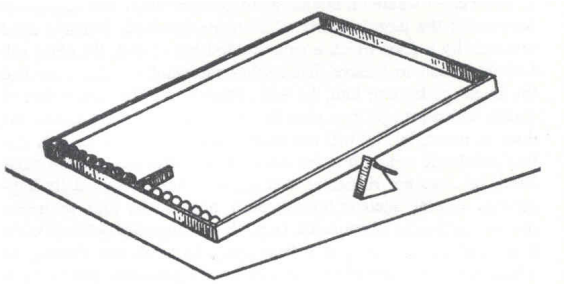


Figure 2: Marble rolling apparatus (from Piaget and Inhelder, 1951, p. 2)

The youngest children (age 4 to 7) often held "quasi-animistic" views of the process, such as "They know where to roll because they roll all by themselves" and "they will all go back where they belong" (pp. 7 and 9). Students in the next age group (7 to 11) often foretell a general crossing and mixing of the colors, and the beginnings of probabilistic reasoning emerge ("because it's chance, and nobody ever knows how that will turn out," p. 18). The oldest children (11 to 12) do not need convincing of the inherent randomness ("I'm sure that it is going to get mixed up," p. 23). They also begin to demonstrate an understanding of the law of large numbers: anything is *possible* though not necessarily *probable* if enough rolls are conducted (p. 24), atop next page.

Interviewer: “And if there were only four balls of each color, could they separate out?”
Child: “Yes, that could happen.”
I: “And with five balls?”
C: “After a long time.”
I: “With six?”
C: “I don’t think so”
I: “If I tip it two thousand times?”
C: “Maybe, but it will be difficult.”

Tversky and Kahneman (1974) identified three heuristics which form the basis for misconceptions of probability:

- *Representativeness*: When asked to identify a person’s profession based on a written description, subjects tended to estimate probabilities based on stereotypes (i.e., the description is *representative* of a person typically considered to be of that profession). For example, subjects were told the relative proportions of engineers and lawyers in the population were 0.3 and 0.7, respectively. Given a relatively ambiguous description of a person (e.g., “30 year old man”, “high ability and high motivation”, “well liked”), subjects judged the description to be equally likely to be an engineer (0.5) or lawyer (0.5). When the relative proportions were reversed, there was no change in the likelihood, indicating that subjects did not consider the prior (population) probabilities and rather based judgments solely on the description.
- *Availability*: Probability estimates tend to be based on cognitive abilities such as the ease of recalling or the ease of searching for an event. For example, subjects were asked if a randomly selected word is more likely to begin with an *r* or to have an *r* in the third position. Because it is easier to recall words that begin with *r*, subjects estimated this event to be more likely, although the reverse is actually correct.

- *Anchoring*: When asked to estimate the probabilities of compound events, subjects tend to bias their estimates towards a perceived initial value, such as that of a related simple event. For example, three types of outcomes were offered in a hypothetical betting scenario: 1) a simple event (probability 0.50); 2) a conjunctive event (simple probability 0.90, total probability 0.48); and 3) a disjunctive event (simple probability 0.10, total probability 0.52). Subjects generally preferred to bet on the conjunctive event, despite having the lowest probability, due to the high anchoring of the simple probability.

Tversky and Kahneman (1982) dissected representativeness and identified a "conjunction fallacy" which is present in many subjects. The now-famous "Linda problem" was introduced in this research, and it is re-printed below. Based on the description, subjects ranked eight probability statements; the numbers in front of the responses are the mean ratings.

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.
>
> Please rank the following statements by their probability, using 1 for the most probable and 8 for the least probable
>
> (5.2) Linda is a teacher in elementary school.
> (3.3) Linda works in a bookstore and takes Yoga classes.
> (2.1) Linda is active in the feminist movement. (F)
> (3.1) Linda is a psychiatric social worker.
> (5.4) Linda is a member of the League of Women Voters.
> (6.2) Linda is a bank teller. (T)
> (4.1) Linda is a bank teller and is active in the feminist movement. (T&F)

A similar "Bill problem" was also constructed. An experiment was conducted in two manners: 1) within-subjects design, in which subjects responded to full versions of the Bill and Linda problems; 2) between-subjects design, in which subjects either saw both options *T* and *F* or only option *T&F*. In design (1), as many as 92% of subjects

demonstrated a conjunction effect: the relative ranking was $F > T\&F > T$. Statistical experience made little difference, as at best 83% of "sophisticated" subjects (graduate students who had taken several advanced statistics and probability courses) made the this mistake. The mean rank of *T&F* was approximately 2.0 higher than *T* across all experience levels and problem versions. Design (2) produced similar results except that the different between mean ranks on *T&F* and *T* alone were closer in value.

Kahneman and Tversky (1972) identified another misuse of representativeness in the estimation of sequence likelihood. Subjects were told that 72 families in a city had six children with the exact birth order GBGBBG; they were asked to estimate the number of families with the birth order BGBBBB. Analytically, any sequence is as likely as any other. However, 82% of subjects judged the latter sequence to be less likely ($p < 0.01$), with a median estimate of 30 families. To test the whether subjects ignored the order aspect and interpreted the question as asking whether it is more likely to have 3 boys or 5 boys out of 6, these subjects were also asked to assess the relative probabilities of the sequences BBBGGG and GBBGBG. Again, the sequence which apparently lacks randomness (BBBGGG) was viewed as less likely ($p < 0.01$).

Building on this idea of perceived randomness, another experiment asked subjects to rate the more probable distribution of 20 marbles among five children. The results of the previous experiment were confirmed: subjects viewed the sequence with apparent randomness (4-4-5-4-3) as more probable than the sequence lacking randomness (4 marbles per child); 69% of subjects made this error ($p < 0.01$).

Konold (1989) identified a rational, non-normative form of probabilistic reasoning, which he dubbed the *outcome approach*. This incorrect logical schema occurs

when a probability is interpreted to be a prediction of the outcome of a single event. In interview sessions, subjects were presented vignettes of realistic situations designed to probe probabilistic reasoning. One of these stated that a meteorologist had predicted a 70% chance of rain on a given day, while the actual outcome was no rain. An example of the erroneous application of the outcome approach is a student quoted as "…that he [the meteorologist] maybe just fouled up," while a student possessing correct probabilistic reasoning stated "…on the basis of just the sample, I think an unrational response would be that the prediction is wrong" (p. 66).

Albert (2003) investigated the frequency and application of the classical, frequency (empirical), and subjective viewpoints in making probabilistic estimates. In problems where a classical sample-space evaluation could be made, such as drawing a ball randomly from a box of known composition, nearly all students provided the correct answer. When a problem afforded a frequentist estimation, responses showed greater variability but it was evident that population characteristics were considered: when asked the probability of randomly selecting a female student from among all university students, the median estimate (58%) was close to the true proportion and 55% of students gave an estimate above 50%, although 50% was the modal response.

Subjective probability estimates elicited a great deal of variability in response strategies. One question asked the student to rate his probability of graduating within four years. Nearly one-third estimated either 0% or 100%, which implies an "either/or" view of probability. Of those who gave an explanation, 73% used subjective reasoning. A small number (19%) attempted to make an objective estimate, such as assuming equal likelihood because there are two options or calculating 25% (1 divided 4 years). Another

question asked students to estimate the likelihood that they will get married before age 25. Fewer students (60%) used subjective methods in this case, although it appears that those who did gave a much greater variety of responses than the previous question. The attempted objective estimates of some students are fascinating: e.g., "There are 24 years before number 25 so you have a 1 in 24 chance to get married" and "I took my age and divided by 25. This then gave me a 72% chance that I will marry before 25" (p. 43).

In summary, the author recommends spending less time on classical probability because students clearly grasp it. The frequentist viewpoint should be given more attention to help students identify relevant population characteristics. Students should also be helped to identify situations which require a subjective judgment, especially in breaking some students of the urge to make meaningless calculations.

4. Cognitive ability: Statistics

Pollatsek, *et al.* (1981) assessed understanding of the mean at a very basic level. A preliminary study required calculating the cumulative GPA for a student who had attended one university for two semesters and another for three semesters. Thirty-seven undergraduates were surveyed at the beginning of a statistics course, and only 14 (38%) responded correctly, despite the apparent simplicity of calculating a weighted mean. To follow up on these results, interviews were conducted with 17 undergraduates (3 of whom had taken the initial quiz; the others had no statistics coursework). Of 15 subjects who worked out a GPA problem, only two (13%) responded correctly. All 13 incorrect students mentioned the unweighted solution at some point during the interviews.

Mevarech (1983) assessed students' misunderstandings of mean and variance in terms of the closure, commutative, associative, and distributive properties. A test was

constructed with problem statements and solutions, and students were asked to identify incorrect solution procedures. All solutions were computationally correct, allowing students to focus on conceptual mistakes. The subjects were 56 freshmen who had completed an introductory statistics course and 47 sophomores who had completed two statistics courses; all were education majors.

Incorrect solutions involving weighting were the most difficult for students to recognize. Weighting requires understanding of the associative property and closure. Two specific oversights were the computation of a simple mean when a weighted mean was appropriate and failure to recognize the order of operations when averaging three numbers. These properties were also the most difficult in variance calculations, but further comments are not possible because the question-solution pairs are not provided. The results between freshman and sophomores were not statistically different on 14 of the 15 items, implying that statistical experience has no effect on the conceptual understanding of mean and variance.

Bar-Hillel (1974) used graphical displays to probe students' intuitive grasp of sampling distribution when presented three trinomial distributions, such as Figure 3.

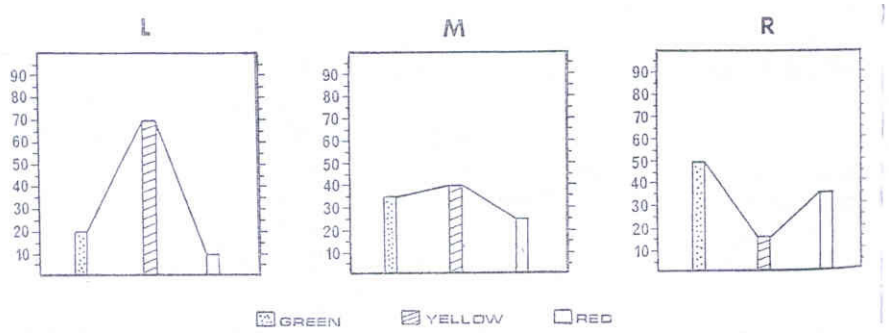


Figure 3: A typical family ("triples") of trinomial distributions (from Bar-Hillel, 1974, p. 278)

For each set of graphs, questions were posed in one of three ways:

1) similarity (S) – which graph (L or R) is more similar to the middle graph (M);
2) likelihood of populations (LP) – is L or R a more likely population distribution to yield sample M;
3) likelihood of samples (LS) – with M as the population, is L or R a more likely sample.

The distributions were constructed to test the hypothesis that similarity (as a proxy for representativeness) would serve as the decision heuristic. Several steps were taken to ensure that the true correct answer would be in contradiction to this similarity heuristic: each bar in M was half-way between its respective values for L and R; the rank-order of M was the same as either L or R but not both; based on formal probability calculations, all trinomial sets for LP and LS were contradictory to the corresponding rank-order (e.g., in Figure 3, M has the same rank-order as L, but the true more likely distribution is R).

With three groups of subjects (S, LP, LS; 25 or 26 subjects per group), two blocks of 15 triples were displayed for a duration of 10 seconds per triple. Responses were grouped by majority response (e.g., 20 subjects said L was more similar to M, while 5 said R was more similar to M, therefore L was taken as the response). The results confirmed the hypothesis: "there are no systematic differences between judgments of likelihood and similarity" (p. 281), that is, the S group responses displayed the same pattern as the LP and LS groups at a highly significant level (Fisher's exact probability test, $p < 0.001$).

Pollatsek, *et al.* (1984) sought to distinguish between students use of the representativeness heuristic and active balancing (i.e., gambler's fallacy) in the understanding of random sampling. One of the study questions is shown below.

> The average SAT for all the high school students in a large school district is known to be 400. You have randomly picked 10 students for a study in educational achievement. The first student you picked had an SAT of 250.
>
> What do you expect the average SAT to be for the entire sample of 10?
> What do you expect the average SAT to be for the next 9 students, including the 250?

The correct solution to the first part is 385 (weighted mean of 9 scores of 400 with the one score of 250), while the correct solution to the second part is 400 (equal to the population mean). The item was administered to 205 psychology undergraduates in questionnaire form. Only 21% of these students responded correctly. The most common incorrect answer was that both the 10-student and 9-student means will be 400, despite the mathematical impossibility; this is taken as a form of the representativeness misconception because samples are assumed to be representative of the population. The balanced solution (10-person mean of 400, balanced by a 9-person mean of over 400) was found in only 12% of respondents, while 9% expected the trend of selecting low scores to continue in the larger sample; 24% of respondents made some other error that was unclassifiable, likely due to misreading the problem.

To verify these findings, interviews were conducted with 31 students of similar background (21 answered the SAT item, while 10 answered a similar item about IQ). The results are not appreciably different from the larger questionnaire sample, except that fewer interviewees gave unclassifiable responses. When further probing allowed subjects to change their responses, the representativeness response proved to be deeply-held. These results contrast with many statistics textbooks which assume the gambler's fallacy to be the under-lying misconception.

Well, *et al.* (1990) performed a series of experiments to identify which aspects of the law of large numbers students fail to comprehend. A sample question is shown below (p. 293, from Experiment 1):

> *Common first paragraph*
> When they turn 18, American males must register for the draft at the local post office. In addition to other information, the height of each male is recorded. The national average height of 18-year-old males is 5 feet 9 inches.
>
> *Accuracy version*
> Yesterday, 25 men registered at post office A and 100 men registered at post office B. At the end of the day, a clerk at each post office computed and recorded the average height of men who registered there that day.
> 1. Would you expect one of these recorded average heights to be closer than the other to the national average?
> 2. If so, which one?
>
> *Tail version*
> Each day 25 men register at post office A and 100 men register at post office B. At the end of every day, a clerk at each post office computes and records the average height of the men who registered there that day.
> 1. On what percentage of days would you estimate that the average height recorded for post office A is greater than 6 feet?
> 2. On what percentage of days would you estimate that the average height recorded for post office B is greater than 6 feet?

In a similar format as displayed above, four experiments were performed. These are summarized below.

1) The performance of "tail" and "accuracy" versions of the same problems showed that students are aware that larger samples are more likely to provide accurate parameter estimates (73% correct), whereas students have a difficult time recognizing that larger samples are less likely to deviate more from the population value (43% correct).
2) To discern if possibly elaborate wording in the "tail" version was a source of confusion, a third similarly complexly-worded problem was added with the context of being more likely to be centered. Students demonstrated nearly equal comprehension of the "accuracy" (56% correct) as the "center" (59%) version, although these results were lower than found in Experiment 1.

However, the "tail" version proved extremely difficult (8% correct), suggesting an incomplete understanding of sample distributions.

3) To further probe these directional difficulties, Experiment 3 utilized "two-tailed" and "one-tailed" versions along with "center." Performance on the "center" version (48%) was significantly better than either tailed version, which were not significantly different from each other (34% "two-tailed"; 25% "one-tailed"). Multiple choice rationales were also offered. Among students who selected a reverse answer, the modal response was that a large sample provided more opportunities for extremes, which is a failure to view the problem as asking about proportion of *means* rather than a proportion of *counts*. Another group of students did not perceive that sample size has an effect (e.g., "number of [samples] not a factor in average", p. 302).

4) A group of students "qualified" for Experiment 4 by first missing a problem in this vein. These students were then interviewed. Most striking, only 3 of the 21 subjects were able to properly re-state the problem. As in Experiment 3, it was most common for subjects to state that the problem was asking about the percentage of *individual datum* that were likely to deviate rather than the percentage of *sample means*. Even after interactive, computerized training, 16 of 21 subjects failed to recognize that a sample of size 10 is likely to deviate more from the population mean than a sample of size 100.

A common misconceptual heuristic was identified throughout these experiments: some subjects reasoned that larger-sample means were more likely to vary from the population mean because a large sample provides more opportunities for extreme values.

Further, correct answers were sometimes gleaned by use of a naïve heuristic that might be dubbed "bigger is better."

In four experiments, Fong, *et al.* (1986), explored the degree to which frequency of statistical reasoning and quality of responses were influenced by type of training, problem type, and statistics experience. The statistical principles of the law of large numbers and regression to the mean were the topics. The goal is to determine whether formal training enhances statistical reasoning across domains, as opposed to an empiricist view that learning is domain-specific.

Throughout, responses were rated on a 3-point scale: 1 = "an entirely deterministic response"; 2 = "a poor statistical response"; 3 = "a good statistical response." The frequency of statistical responses was the percentage of 2 or 3 ratings, while the quality of statistical responses was the percentage of 3's out of all 2's and 3's [i.e., p(3 | 2 or 3)].

In Experiment 1, statistically-naïve individuals (primarily high school students and female adults) were given training either through examples only, rules only, or both; one control group received no training, while another viewed only a single sentence about the law of large numbers before completing the exercises. Eighteen problems were answered by subjects in three general categories: 1) *probabilistic* – a sample was clearly drawn from a population containing random variation; 2) *objective* – variation was not explicit but was meant to be inferred; 3) *subjective* – the scenario involved a subjective decision where sample size effects should have been considered.

Collapsing across other variables, training had a significant effect over the control conditions. Further, those who received full training (both through rules and examples)

performed significantly better than those who received only one type of training. These results held for both frequency of statistical responses and the quality of those responses which were statistical, although the differences were less pronounced on the latter. Problem-type also had a significant effect on frequency: the pattern *probabilistic* > *objective* > *subjective* held across all training levels. Quality did not vary significantly across problem-type, although it is interesting that the *subjective* generally afforded the highest quality rather than the lowest. Moreover, training did not lead to the occurrence of "false alarms" (i.e., using the law of large numbers in inappropriate situations).

In Experiment 2, transfer effects of the training method were analyzed by giving subjects the full training but with only examples from one of the three problem types; a control group with no training was included as well. The testing materials were identical to Experiment 1, and subjects answered responded to all three problem types. All subjects were undergraduates in introductory psychology.

Confirming the previous experiment, the same pattern of problem-type was found in terms of frequency and the quality did not vary significantly across problem-type. Training was found to result in significant improvements over the control group in terms of both frequency and quality, but there were no significant differences across training methods for either criterion. This supports the formalist hypothesis that statistical training is transferable across content domain.

Experiment 3 was a within-subjects design in which subjects were asked to reason about a scenario either with *no randomness cue* or with a *randomness cue*. Statistical experience was included as a variable with four levels: 1) *no statistics* – college students who had not taken a statistics course; 2) *statistics* – from the same group as (1) but had

taken a statistics course; 3) *graduate* – psychology graduate students who had taken at least one statistics course and typically several; 4) *tech* – technical staff members from a research laboratory who primarily had a Ph.D. and had taken several statistics courses. Across all experience levels, the *cue* condition produced higher frequencies ($p < 0.001$), although quality was approximately constant as a function of cue presence. Statistical experience had the anticipated effect ($4 > 3 > 2 > 1$), except that group (2) performed better than group (3) on the *no cue* quality criterion.

Lastly, Experiment 4 sought to determine if the previous experiments had a placebo effect of sorts: these subjects may have reasoned statistically because they were aware of the nature of the studies. A deceptive study was designed wherein subjects were told that they were participating in a survey about their opinions on sports. All subjects were male undergraduates in an introductory statistics course; half were surveyed at the beginning of the semester and half at the end, with those who indicated little knowledge of sports excluded. One question involved explaining the common "sophomore slump" phenomenon experienced by winners of the Rookie of the Year award. Training had a significant effect on frequency (16% pre-test, 37% post-test, $p < 0.005$), although quality had only a marginally significant increase (12% pre-test, 38% post-test, $p < 0.10$). One other question showed a significant training effect on both frequency and quality, while two others showed no effect on either criterion.

In sum, these studies reinforce the formalist view held by the authors, in contrast to those who believe in domain-specificity. These findings hold even for subjective social situations where chance arises but is rarely acknowledged, such as "first impressions." The authors acquiesce to the possibility that the immediacy of training

may be the cause of these promising results, and they endeavored to examine this in their next study.

Fong and Nisbett (1991) investigated transfer effects in learning the law of large numbers (LLN). A 2×2 factorial design was employed with content domain (sports or ability testing) and timing (immediate or 2-week delay) as the independent variables; a control group who received no training was included as well. The criterion measure was a three-level ordinal scale [refer to previous study by Fong, *et al.* (1986)]; two raters attained a high reliability of 85% exact agreement. Training materials were a booklet which took approximately 15 minutes to read, and subjects (n = 231) were undergraduates enrolled in an introductory psychology course.

When tested immediately following training, statistical reasoning scores were context independent: regardless of the training context, subjects scored approximately equally well on sports and ability testing questions; further, these scores were much higher than the control group who received no training. However, context proved to have a significant effect when tested after a 2-week delay: subjects scored significantly higher on the domain which matched their training context.

These findings raised the question of whether students scored better due to memory of example problems in the training booklet or due to context effects of the training. In a second experiment, in addition to being tested as above, subjects were given a questionnaire to assess their memory of the example problems, while another group was quizzed only on their ability to recall examples but was not asked to solve new problems. The contextual training effects of the first experiment were similar in the second experiment. Students demonstrated a poor recall of the example problems, with

only 35.9% able to accurately recall even one of three problems. The control group who were only tested on memory of the examples actually had a better recall than those who were fully tested: solving the new problems did not enhance recall of the training examples and actually may have clouded memory. Understanding of the LLN persisted, however, as 78.5% of subjects could recall the concept sufficiently to warrant the highest code in a 4-level scale. The relationship between memory of the LLN and statistical reasoning scores ($r = +0.31$, $p < 0.05$) was much stronger than the relationship between example problem memory and statistical reasoning scores ($r = +0.06$).

A third experiment was conducted along similar lines, with students being asked to rate the extent that they used the example problems in solving the test problems. Only 15.6% agreed with the most explicit statement ("I definitely used this sample problem in solving the test problems."). There was no domain effect, and the mean number of example problems used by students was 0.22 out of 3.

In sum, these experiments suggest that inferential rule training is possible, and although context does affect transfer, some amount of learning is transferable (i.e., scores on the un-trained domain are significantly higher than the control group). Memory effects were shown to be minimal, further suggesting that the learning was conceptual and not merely rote.

Ploger and Wilson (1991) temper these findings by questioning the extent of transfer. They highlight the fact that recall of the LLN rule was high. And while trained subjects out-performed un-trained subjects, a large proportion of students still demonstrated deterministic reasoning (e.g., 44% in the ability testing group with no delay), implying that even very-near transfer was poor. The lower results on the un-

trained domain further illustrate a lack of near transfer to similarly constructed problems. In contrast to the optimism of the original article, this response concludes that "most college students did not apply the LLN to problems in everyday life" (p. 213).

5. Teaching Strategies

Austin (1974) compares the effectiveness of three teaching strategies in an introductory statistics course consisting primarily of freshman and sophomores who were not science or mathematics majors. All students received the same oral lectures in tape format. The written materials varied across the three treatments in an ordinal nature: symbolic (S) – written materials contained only words and symbols, essentially a "traditional" teaching strategy; pictorial (P) added graphs, diagrams, and figures; manipulative-pictorial (MP) required the students to also perform simple random experiments such as with dice or coins.

The experiment consisted of 12 lessons over a one-month period. The criterion was a final examination, further taxonomized into four components: comprehension, computation, application, and analysis. Seventy-one students in two sections completed the study (nine were lost to attrition). The modal demographics were sophomores (n = 31), business or humanities majors (57), with no previous statistics instruction (40). The examination was found to be reliable (α = 0.90, 0.92, 0.93 across treatments; k = 40), as were the four components (minimum α = 0.59 for k = 4, computation; maximum α = 0.90 for k = 16, application).

The analysis method was a 2×2 ANOVA (3 treatments, 2 sections, plus interaction). For each component, as well as the total exam score, the section and interaction were non-significant ($p > 0.50$ in all cases) and these terms were pooled for

further analyses. The computation component produced a non-significant effect, but the total test score and the other three components yielded significant effects across teaching methods. The following pattern was found in terms of average score: (P) > (MP) > (S), with (P) and (MP) not significantly different from each other but greater than (S), except for comprehension where (MP) and (S) were not different.

These results highlight the effectiveness of providing pictorial representations of statistical concepts to students beyond the strict use of symbols, although the additional task of performing experiments did not enhance learning. The author cautions extrapolating these findings to all students due to the demographics and the use of audio-taped lectures as opposed to classroom interaction. The latter may have been an especial hindrance to the symbolic (S) group.

Simon, *et al.* (1976) describe an active, constructivist approach to teaching statistics using a Monte Carlo approach. For example, consider the following vignette:

> "John tells you that with his old method of shooting foul shots in basketball his average (over a long period of time) was .6. Then he tried a new method of shooting and scored successes with nine of his first ten shots. Should he conclude that the new method is really better than the old method?" (p. 734)

From a traditional, analytical teaching perspective, this problem would be viewed as a hypothesis test on a proportion. With the Monte Carlo approach, students perform their own random experiments to arrive at conclusions. In this case, the authors describe using 20 playing cards (12 hearts to represent made free throws; 8 spades for misses). Each student draws, *with replacement*, ten cards from the set and records the number of hearts; this is repeated 15 times. One student is described as having drawn 9 or more hearts on only one of his fifteen trials. Because 1 out of 15 is an unlikely result, the student concludes that the shooter's new method must in fact be an improvement. In

traditional parlance, he rejects the null hypothesis that the proportion equals 0.6 (H_o: p = 0.6) in favor of the alternate that the proportion is greater than 0.6 (H_a: $p > 0.6$).

The Monte Carlo method was further tested in controlled experiments at two colleges. In both cases, multiple sections of the same course were offered with at least one teaching in the traditional way and another employing Monte Carlo techniques. The sections were approximately equal in previous mathematical experience and attitude towards mathematics at the first college, while the pre-course results actually favored the traditional section at the other college. At the first college, the students in one (non-computer) Monte Carlo section received average scores 62% higher than the traditional section (the results were slightly less impressive for students in a computer Monte Carlo section). The differences between the sections were not only *statistically* significant: the Monte Carlo sections' performance was high enough to be of *practical* significance as well. The Monte Carlo students also increased attitudes towards mathematics, whereas the conventional section had slight decreases in attitude. At the other college, the Monte Carlo sections maintained stronger performance, although not as pronounced, despite the fact that these students had lower average incoming mathematical ability and attitudes. The greatest lesson gathered from this study is that Monte Carlo methods engage students in the learning process through active experimentation and facilitate instructor-student interaction.

B. Confidence Analysis of the SCI

1. Method

1.1 Motivation

As previously documented, a number of studies have sought to analyze why students have difficulty in statistics courses and with statistics/probability concepts. However, none of these studies examined a broad range of topics. The purpose of this study is to determine students' conceptual difficulties across a wide sampling of topics commonly encountered in introductory statistics.

1.2 Data collection

After answering each item, students are presented with a rating scale to gauge answer confidence. The scale ranges from 1 ("Not confident at all") to 4 ("Very Confident"). The scale was designed with an even number of options so that students could not naturally gravitate to a neutral opinion. If the "Next Question" button is clicked without providing a confidence, a default value of zero is recorded but ignored for analysis. A sample screen shot of an item with the confidence scale is shown in Figure 4.

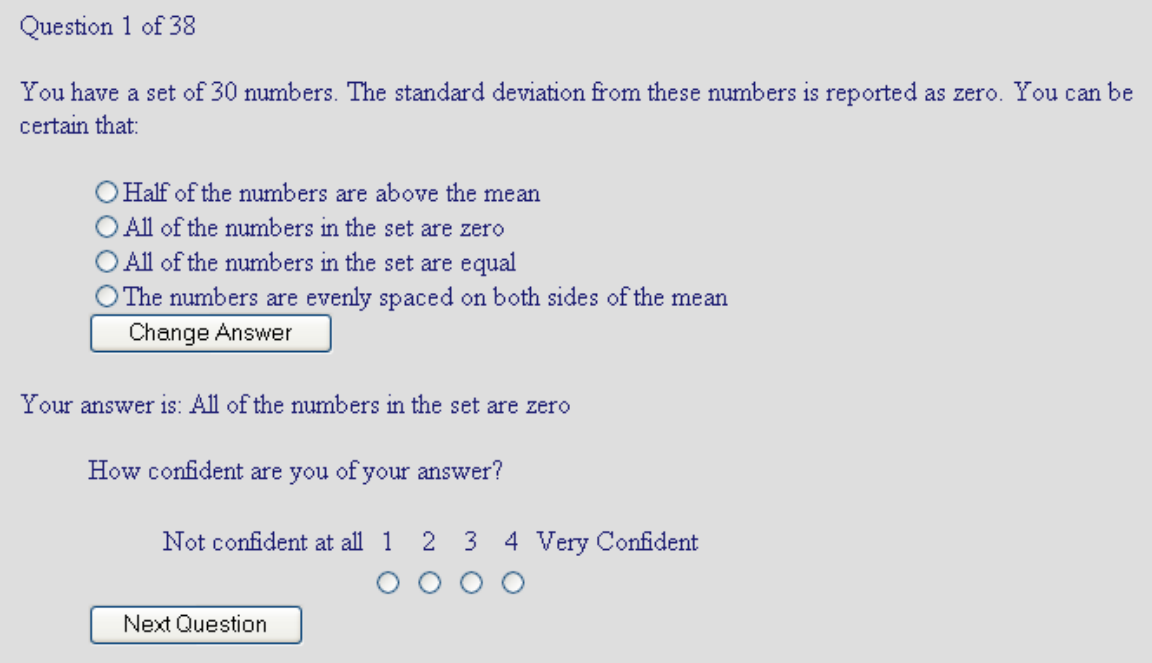


Figure 4: Sample SCI item with confidence rating scale

1.3 Results sample

For each item, two graphics are produced. The first of these groups student responses by confidence level, as shown in Figure 5a. The top of the graph lists the question number, sub-test, average percent correct and discrimination index (in parenthesis), and a more specific description of the topic. Each bar shows the percent correct and number of students at each confidence level. For example, a confidence of 1 was selected by 85 students, and 41% of these students were correct.

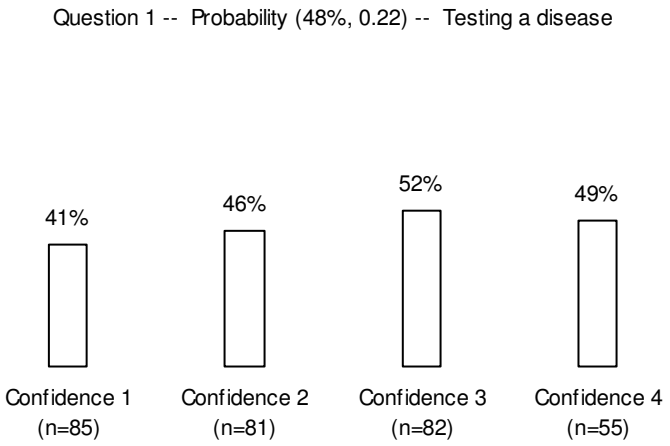


Figure 5a: Confidence graph sample, grouped by confidence level

The second form of graphical display is shown in Figure 5b. The top of the graph is the same as previously discussed. Each bar now represents the average confidence level for students who selected each letter option. For example, choice A was selected by 22 students with an average confidence rating of 1.73. The correct answer (C, in this case) is marked with "**".

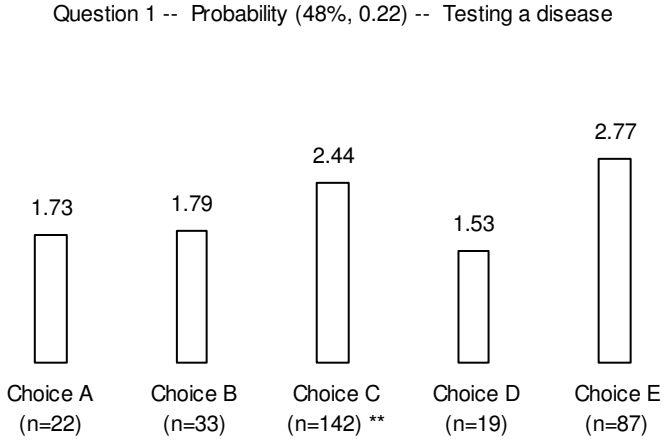


Figure 5b: Confidence graph sample, grouped by answer

1.4 Research question

It is hypothesized that a general positive relationship will exist between the average confidence and the percent correct; that is, questions which have high average confidence will generally be answered correctly, and vice versa. That said, the most interesting items will be those where the relationship does not hold: specifically, questions which are rated as highly confident but are missed by students. These represent possible misconceptions. If students lack confidence on an item and attain low scores, this should represent guessing (i.e., a "non-conception") rather than a misconception.

2. Results Summary

2.1 Reliability

Table 2 summarizes the reliability of the confidence scale for the overall test and each sub-scale. The number of responses (n) varies across sub-scales due students who skip the confidence rating; Cronbach's α can only be calculated for subjects who responded to all items (e.g., 296 students gave confidence rating on *all* Graphical items). The Spearman-Brown prophecy formula is used to adjust the given reliabilities to a common test length of 38 (the length of the complete SCI). The total reliability of the confidence scale is very high. The adjusted reliabilities are essentially constant across the sub-scales and equal to the overall reliability.

Table 2: Reliability of confidence scale

	Total	*Probability*	*Descriptive*	*Inferential*	*Graphical*
Reliability (α)	0.9324	0.7967	0.8131	0.8149	0.7062
Adjusted α	--	0.9430	0.9376	0.9383	0.9288
(n) of responses	(134)	(287)	(286)	(154)	(296)

2.2 Results – macro

Table 3 presents the summary statistics for the average confidence ratings across the 38 items on the SCI. Figure 6a displays these average confidence ratings graphically. The histogram appears slightly negatively skewed, but the summary statistics suggest an approximately symmetrical distribution (mean ≈ median; median nearly equidistance from Q1 and Q3).

Table 3: Summary statistics for average confidence ratings across items

Minimum	*1st Q*	*Median*	*3rd Q*	*Max*	*Mean*	*St. Dev.*
2.23	2.62	2.89	3.14	3.58	2.87	0.39

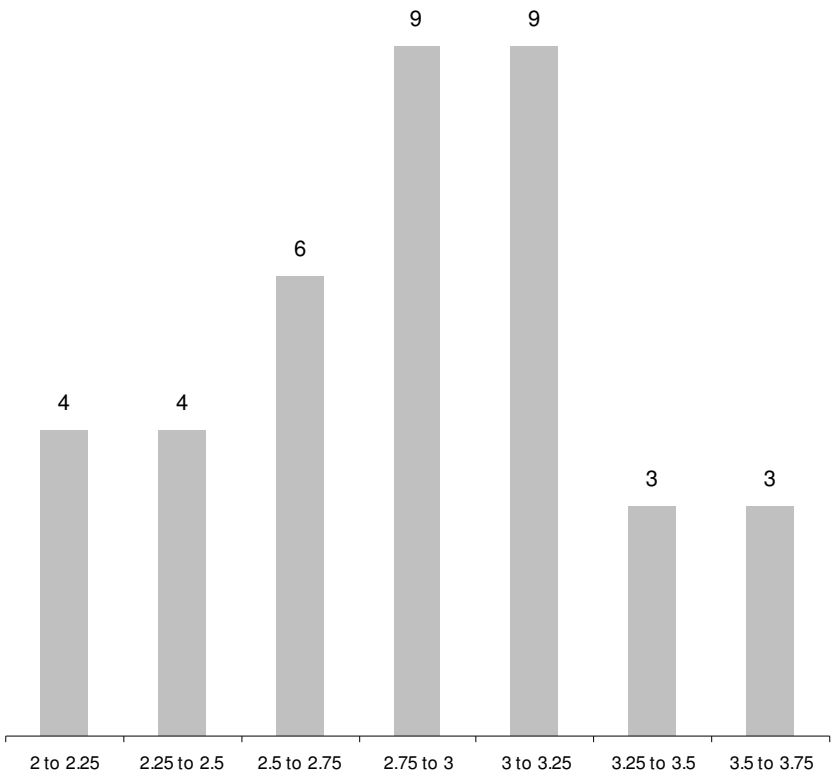


Figure 6a: Frequency distribution of average item confidence, with bin size 0.25

Table 4 summarizes the confidence for students who answered each item correctly. The mean is the simple mean, while the weighted mean favors items with

higher percent correct. The higher values of the weighted mean imply that students are more confident about items which they answer correctly.

Table 4: Confidence of correct answers

	Total	*Probability*	*Descriptive*	*Inferential*	*Graphical*
Mean	3.01	3.09	3.14	2.98	2.74
Weighted Mean	3.09	3.16	3.21	3.08	2.81

Table 5 displays the summary statistics grouped by correct and incorrect responses. Incorrect responses elicit lower confidence ratings across the board. The frequency distribution (Figure 6b) further illustrates the preponderance of low confidence for incorrect responses. The correct responses have an unanticipated pattern with only 3 items falling into the "2.75 to 3" bin, while the surrounding bins have 8 items. Further, the "2.25 to 2.5" bin has more correct items than incorrect (6 vs. 5).

Table 5: Summary statistics for average confidence ratings across items, partitioned by Correct vs. Incorrect

	Minimum	*1st Q*	*Median*	*3rd Q*	*Max*	*Mean*	*St. Dev.*
Correct	2.38	2.57	3.06	3.31	3.83	3.01	0.42
Incorrect	1.92	2.38	2.71	2.85	3.55	2.64	0.36

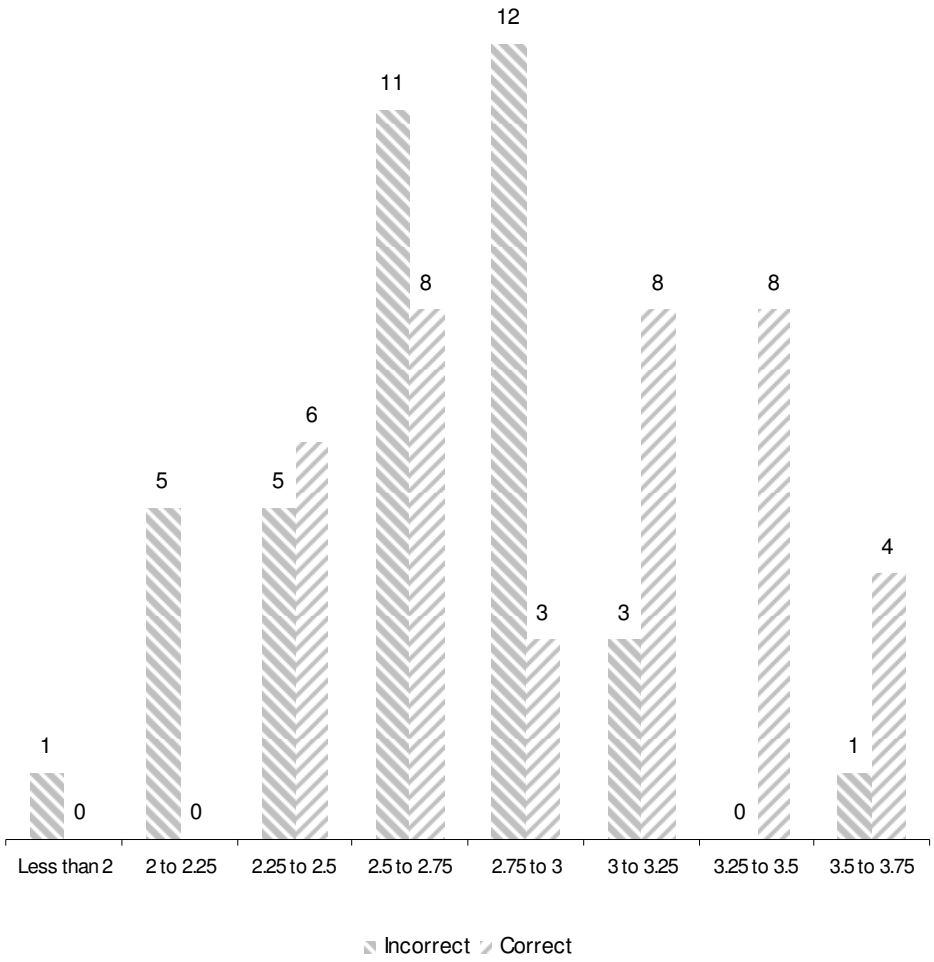


Figure 6b: Frequency distribution of average item confidence, with bin size 0.25, grouped by Incorrect (left) and Correct (right) responses

The relationship between percent correct and average confidence is of most interest. Figure 7a shows this relationship in numerical form ($r = 0.306$, $p = 0.031$), while Figure 7b uses the rank-orders to provide a wider spread of points ($r = 0.334$, $p = 0.020$). While both graphs provide moderately positive (and significant) correlations, the items of most interest are those which defy the pattern. In Figure 7b, the central dashed line corresponds to equal ranks. The lighter dashed lines above and below represent the points where the confidence and percent correct differ in rank by 10 (arbitrarily selected). Items

with over-confidence (confidence rank exceeds percent correct rank by 10 or more) are marked with a plus (+), while items with under-confidence are marked with a minus (–).

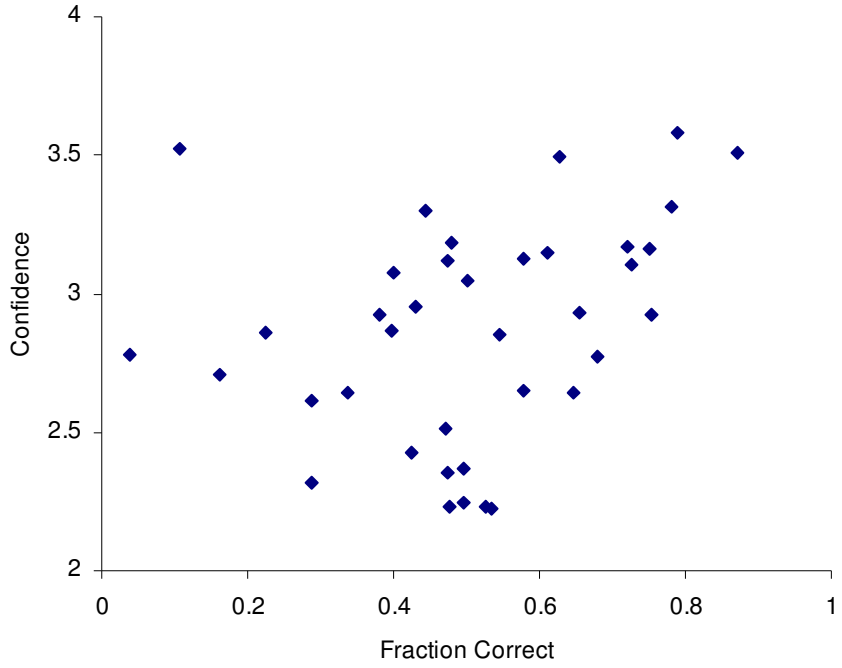


Figure 7a: Confidence vs. fraction correct, numerical values

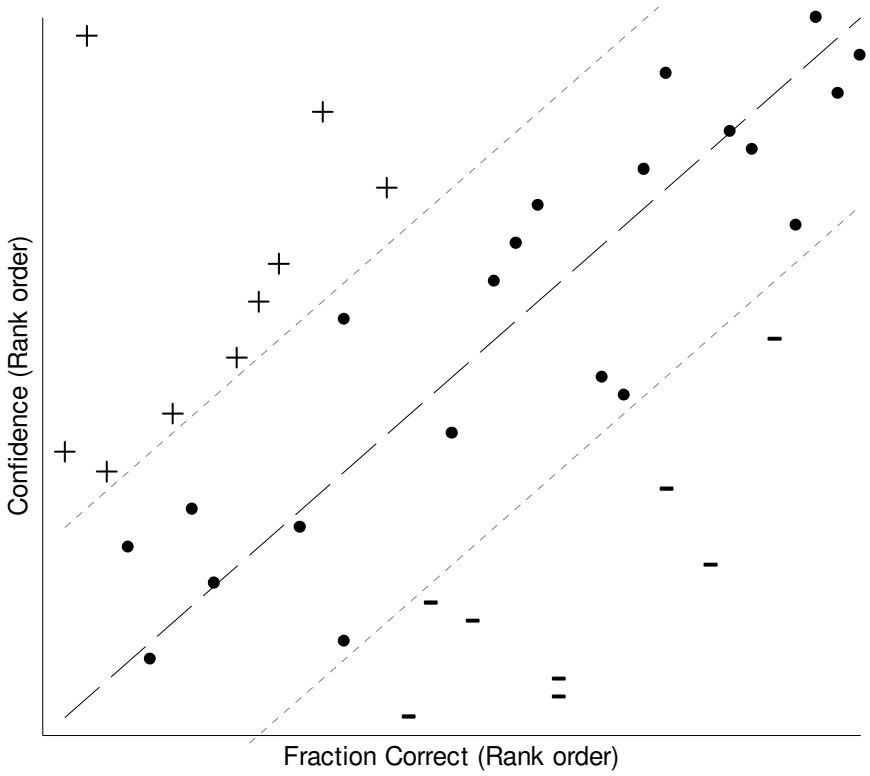


Figure 7b: Confidence vs. fraction correct, rank orders

The items which fall outside the central region are listed in Table 6 (next page). Table 7 shows the number of items from each topic area which fall into each region of the rank-order graph. Probability has nearly half of its items rated as over-confident, while the opposite is true of Descriptive. Inferential and Graphical are nearly equally distributed across the three regions.

Table 6: Items in the over-confidence and under-confidence regions

No.	Sub-test	Correct Rank	Confidence Rank	Difference
		over–confidence		
5	Probability	2	37	+35
10	Inferential	13	33	+20
4	Probability	10	25	+15
13	Probability	1	16	+15
3	Descriptive	4	18	+14
16	Probability	18	32	+14
30	Graphical	8	21	+13
33	Probability	15	27	+12
14	Graphical	3	14	+11
17	Inferential	12	23	+11
20	Inferential	9	19	+10
		under–confidence		
2	Inferential	25	13	-12
6	Descriptive	19	7	-12
25	Graphical	17	3	-14
29	Descriptive	35	20	-15
11	Descriptive	31	15	-16
22	Inferential	20	4	-16
38	Descriptive	29	11	-18
24	Graphical	22	2	-20
37	Graphical	23	1	-22

(low numbers represent low values, e.g., rank 1 is the lowest confidence)

Table 7: Item counts by topic area and confidence region

	Over-confident	*Under-confident*	*Neither*
Probability	5	0	6
Descriptive	1	4	6
Inferential	3	2	4
Graphical	2	3	2

Table 8 shows the probability of the three regions being so populated by the observed number of items. For example, the over-confident region has five Probability items. Based on 11 total items in this region and nine Probability items on the SCI, the expected value is taken to be $\frac{9}{38} * 11 = 2.6$; the over-confident region is therefore over-populated by Probability by 2.4 items. Based on a binomial distribution, the probability

of 5 or more out of 11 occurrences is 2.6%, assuming a base-rate of $\frac{9}{38}$. When regions are under-populated, the reported probability is the lower end (e.g., 1 or less for Descriptive in Over-confident).

Table 8: Probabilities of region dominance

	Over-confident			*Under-confident*			*Neither*		
	Obs.	Exp.	*p*	Obs.	Exp.	*p*	Obs.	Exp.	*p*
Probability	**5**	**2.6**	**0.03**	*0*	*2.1*	*0.09*	6	4.3	0.11
Descriptive	1	3.2	0.13	*4*	*2.6*	*0.09*	6	5.2	0.24
Inferential	3	3.2	0.60	2	2.6	0.49	4	5.2	0.63
Graphical	2	2.0	0.67	*3*	*1.7*	*0.07*	2	3.3	0.33

The abundance of the Probability sub-test in the over-confident region is verified by the probability calculation of Table 8; this is the only *p* less than 0.05. The under-confident region proves to be moderately over-populated by Descriptive and Graphical items and lacking Probability items (all $p < 0.10$).

2.3 Over- / Under-Confidence: A Statistical Basis

The determination of over- and under-confidence in the previous section was based on an arbitrary difference of ten between the confidence and correct ranks. The section searches for a statistical basis for making such claims. A confidence band for a regression line can be defined as the following (Neter, *et al.,* 1996). In SAS ™, this is obtained via the "clm" option of the *proc reg* "model" statement.

$$band = \hat{Y} \pm Ws\{\hat{Y}\} \qquad \text{with } W^2 = 2\,F(1-\alpha, 2, n-2) \tag{1}$$

where: $\hat{Y}$ is predicted from the regression fit
W is defined from the F distribution above
α is the confidence level
n is the number of observations
$s\{\hat{Y}\}$ is the standard error of the predicted value

Table 9 (next page) summarizes the results of regression fits, with a 99% confidence band used to classify over- and under-confidence. The results are grouped such that the original ±10 rule classifications are first, while the lower sections of the table display items which were classified as outside the fit region in subsequent models. It is evident that the over-confident region in the alternate formulations provides poor assessment of the easiest items, showing, for example, items ranked 26^{th} / 28^{th} and 27^{th} / 29^{th} considered over-confident on three of the four models. The under-confident region is inaccurate at lower values, for example 14^{th} / 9^{th} and 11^{th} / 8^{th}, although less so than the over-confident region(s).

The graphs, shown on subsequent pages (Figures 8a-b), make evident the reason for the poor classification: the regression fit is not steep enough. The ±10 rule for ranks constrained the regression line to a slope of 1 and an intercept of zero. The graphs displayed are for the rank-orders, both with and without the coin-flip question ("outlier"). Other fits yield similar results, but their presentation would not enhance the analysis.

Table 9: Confidence Band confidence classifications

Correct	Conf.	± 10 rank	Full data		Coin-flip removed	
			Ranks	Numbers	Ranks	Numbers
2	37	+	+		na	na
1	16	+				
3	14	+				
4	18	+				
9	19	+				
8	21	+			+	+
12	23	+	+		+	+
10	25	+	+	+	+	+
15	27	+	+	+	+	+
13	33	+	+	+		+
18	32	+	+	+		+
17	3	-	-	-	-	-
23	1	-	-	-	-	-
22	2	-	-	-	-	-
20	4	-	-	-	-	-
19	7	-		-	-	-
25	13	-	-	-	-	-
29	11	-	-	-	-	-
31	15	-	-			-
35	20	-				
26	28		+		+	+
27	29		+		+	+
32	31		+			
28	35		+	+		+
36	34		+			
38	36		+	+	+	
37	38		+	+	+	+
21	24			+	+	
14	9		-	-	-	-
11	8		-	-	-	-
15	6		-	-	-	-
5	5		-	-		-

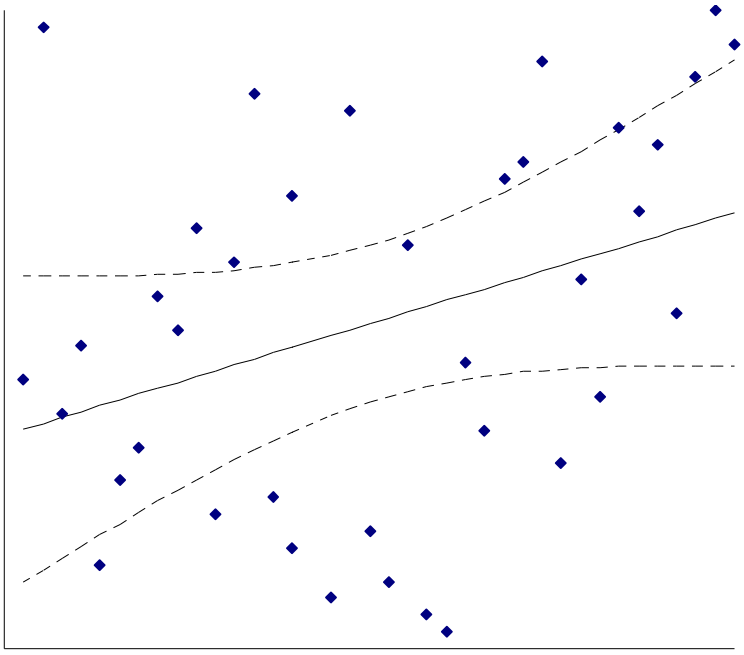


Figure 8a: Confidence bands, ranks, outlier included

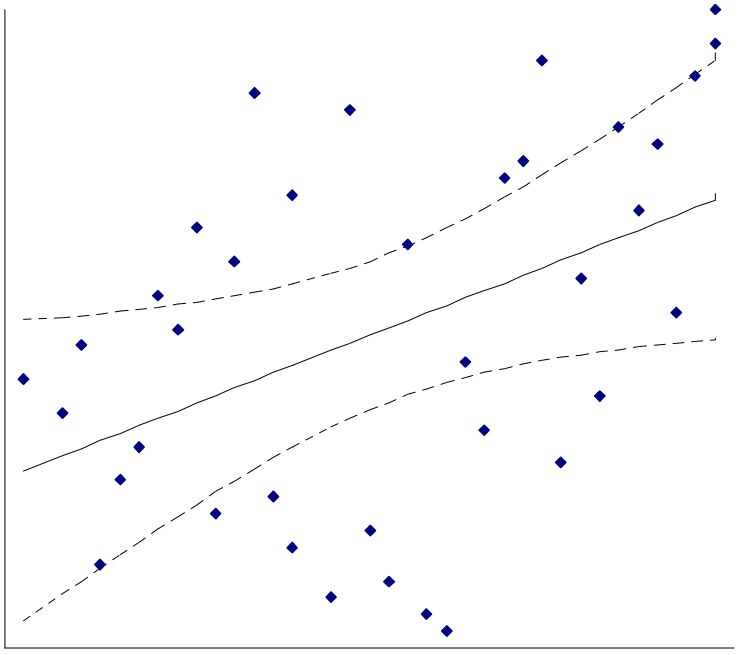


Figure 8b: Confidence bands, ranks, outlier removed

3. Results: Over-Confidence

The items of most interest are those which fall further from the equality line in the rank-order graph. These are discussed in the following paragraphs. The ranks given in parenthesis start from 1 being the lowest among the 38 items (e.g., rank 1 in correct is the lowest percent correct, or the most difficult).

Question 5: Probability (Correct 2nd, Confidence 37th)

> A coin of unknown origin is flipped twelve times in a row, each time landing with heads up. What is the most likely outcome if the coin is flipped a thirteenth time?
>
> a) Tails, because even though for each flip heads and tails are equally likely, since there have been twelve heads, tails is slightly more likely
> b) Heads, because this coin has a pattern of landing heads up
> c) Tails, because in any sequence of tosses, there should be about the same number of heads and tails
> d) Heads and tails are equally likely

Figures 9a and 9b depict the confidence profile for this item. This item is by far the most extreme is its difference between confidence and percent correct. Nearly all students seem trained that coins are "fair" in spite of the extreme unlikelihood of a 50/50 coin being flipped heads 12 consecutive times. Among students who selected the highest confidence, 90% selected choice D, compared with 70% of students of lesser confidence. The few students who select A or C are relatively unconfident, as these values fall slightly below the overall median confidence of all items (2.89, Table 3).

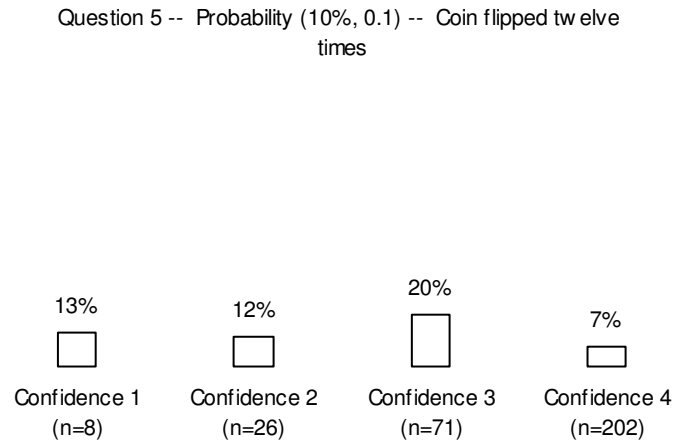


Figure 9a: Confidence profile, by confidence, coin flipping question

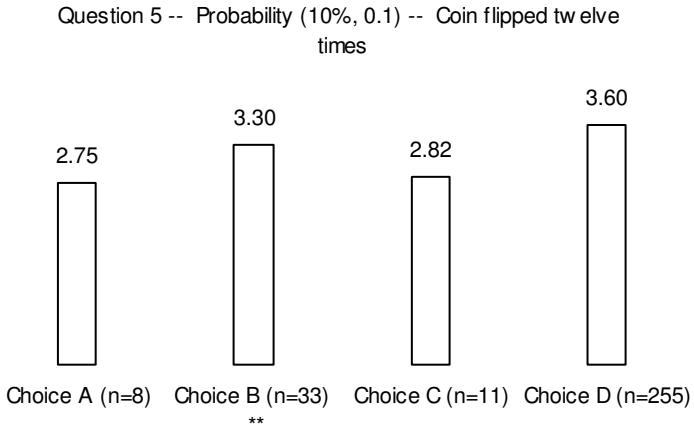


Figure 9b: Confidence profile, by answer, coin flipping question

Question 10: Inferential (Correct 13th, Confidence 33rd)

> A bottling company believes a machine is under-filling 20-ounce bottles. What will be the alternate hypothesis to test this belief?
>
> a) On average, the bottles are being filled to 20 ounces.
> b) On average, the bottles are not being filled to 20 ounces.
> c) On average, the bottles are being filled with more than 20 ounces.
> d) On average, the bottles are being filled with less than 20 ounces.

Figures 10a and 10b display the confidence profiles for this item. By confidence, the item has an anticipated pattern of people with higher confidence more likely to be correct (ignoring the small *n* of confidence 1). The problem with this item lies in the

difficulty, as only 46% even of the most confident students are correct. Option A attracts the highest confidence among incorrect answers, perhaps because some instructors prefer this notation for the null hypothesis, that is, it corresponds to something students have seen previously.

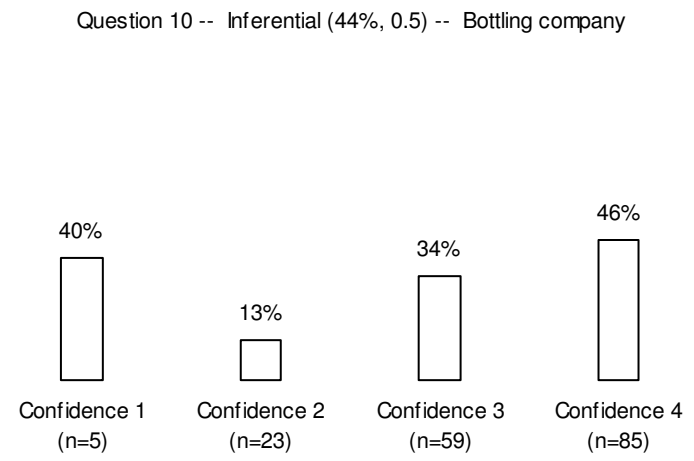


Figure 10a: Confidence profile, by confidence, alternative hypothesis question

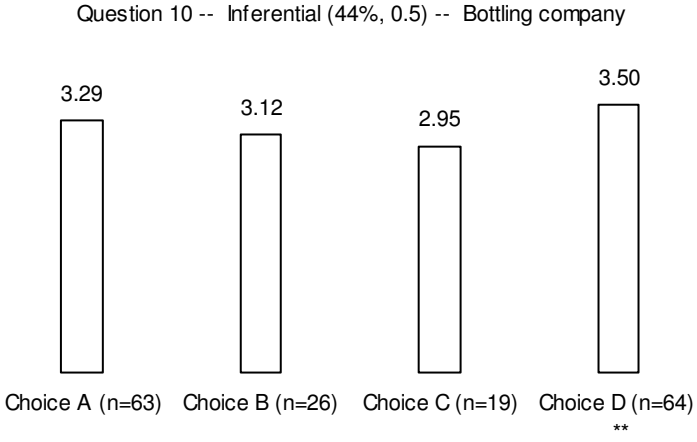


Figure 10b: Confidence profile, by answer, alternative hypothesis question

Question 4: Probability (Correct 10th, Confidence 25th)

> Which would be more likely to have 70% boys born on a given day: A small rural hospital or a large urban hospital?
>
> a) Rural
> b) Urban
> c) Equally likely
> d) Both are extremely unlikely

This item is diagnostically similar to the previous item: percent correct increases with increasing confidence; moderately low overall percent correct; correct answer chosen with highest average confidence; strong discrimination (Figures 11a and 11b). This item is a slight adaptation of one studied by Kahneman and Tversky (1972) and others. With a smaller study sample, they found approximately equal preference for their equivalents of options A (20%) and B (24%), with C (56%) the dominant selection. The SCI results are nearly equally split between A and C (41% and 43%, respectively in the full dataset; 37% and 45%, for online users only). Unlike Kahneman and Tversky, the SCI found scant support for the larger hospital (5% overall, 6% online).

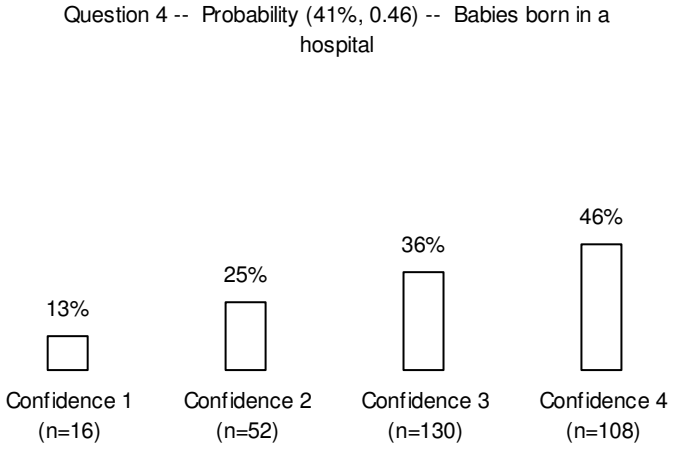


Figure 11a: Confidence profile, by confidence, hospital question

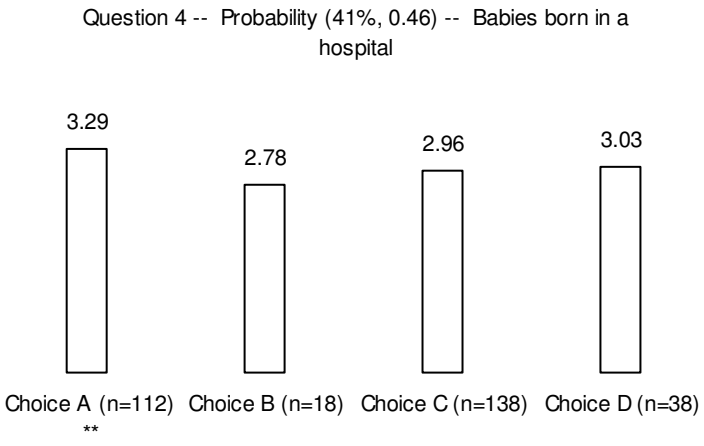


Figure 11b: Confidence profile, by answer, hospital question

Question 13: Probability (Correct 1st, Confidence 16th)

> You have called your cell phone provider to discuss a discrepancy on your billing statement. Your call was received and placed on hold to 'await the next available service representative.' You are told that the average waiting time is 6 minutes. You have been on hold for 4 minutes. How many more minutes do you anticipate you will have to wait before speaking to a service representative?
>
> a) 2
> b) 4
> c) 6
> d) there is no way to estimate

This item is missed by nearly all students (3.7% correct in this sample). Earlier versions of the SCI contained a comparable item with different content, and the results were nearly identical. The concept of the memoryless property is simply not grasped by students. Only one student out of 72 who responded with confidence 4 was correct. Choices A (55%) and D (33%) dominate (Figures 12a and 12b).

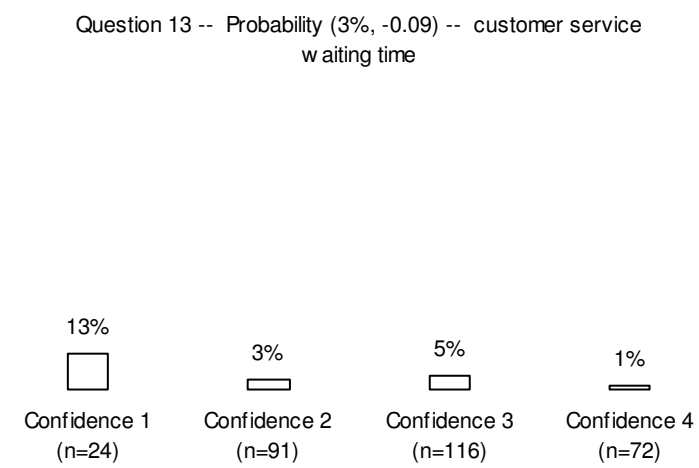


Figure 12a: Confidence profile, by confidence, waiting time question

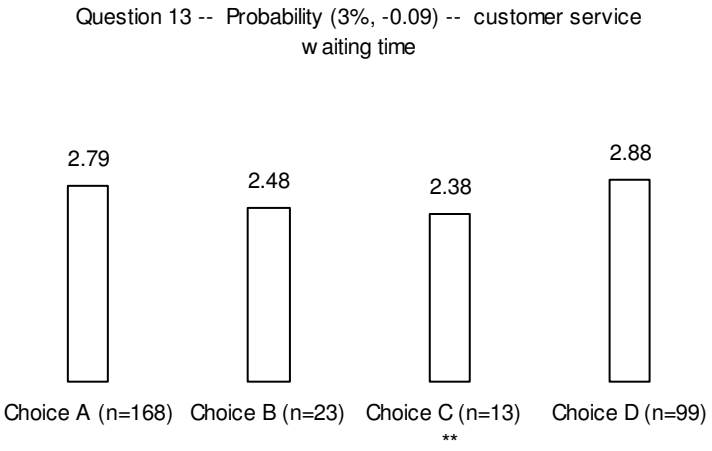


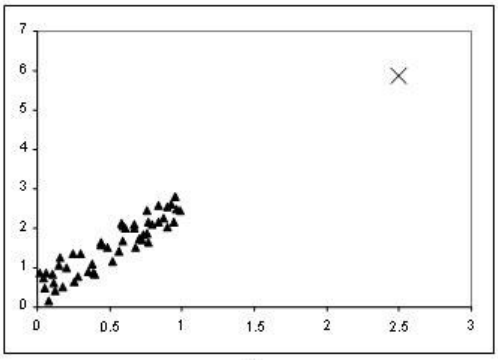
Figure 12b: Confidence profile, by answer, waiting time question

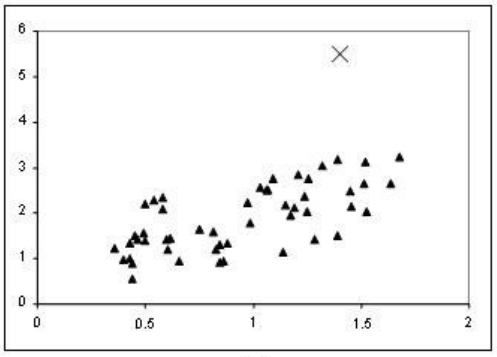
4. Results: Under-confidence

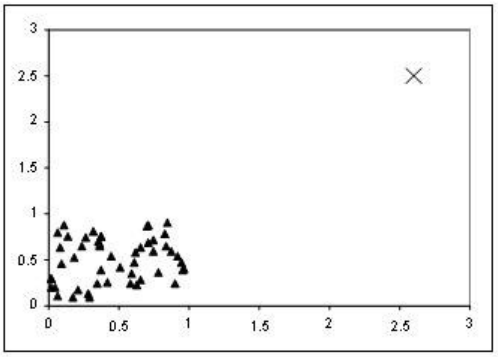
The first three items in this section concern correlation coefficients; as such, they are discussed as a group following the third item.

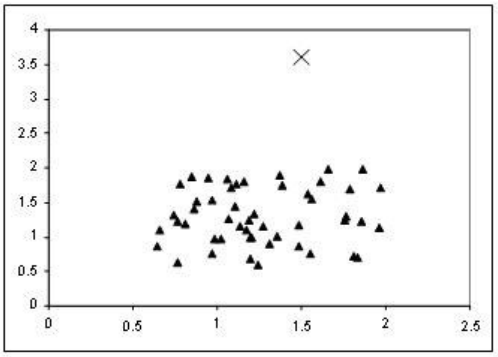
Question 37: Graphical (Correct 23rd, Confidence 1st)

Consider the correlation coefficients of the scatter plots below. If the data point that is marked by an X is removed, which of the following statements would be true?





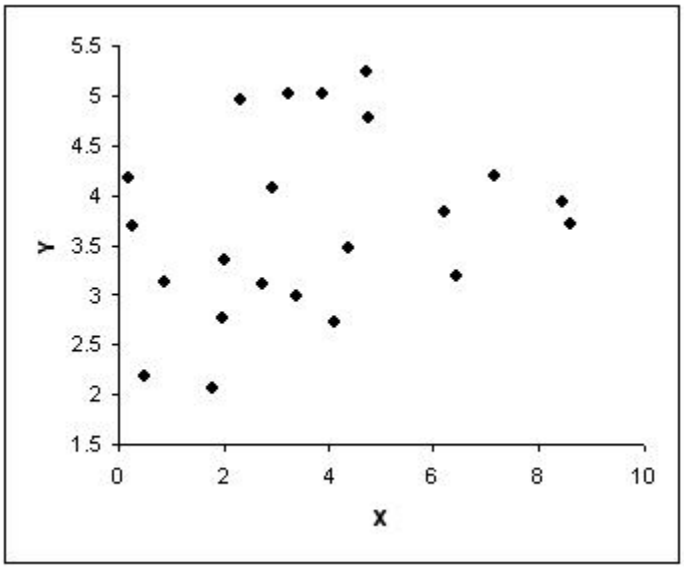




a) correlation of (I) decreases, correlation of (II) stays the same
b) correlation of (III) increases, correlation of (IV) increases
c) correlation of (I) stays the same, correlation of (III) decreases
d) correlation of (II) increases, correlation of (III) increases

Question 24: Graphical (Correct 22nd, Confidence 2nd)

Estimate the correlation coefficient for the two variables X and Y from the scatter plot below.



a) -0.3
b) 0
c) 0.3
d) 0.9
e) 1.6

Question 38: Descriptive (Correct 29th, Confidence 11th)

Information about different car models is routinely printed in public sources such as Consumer Reports and new car buying guides. Data was obtained from these sources on 1993 models of cars. For each car, engine size in liters was compared to the number of engine revolutions per mile. The correlation between the two was found to be -0.824. Which of the following statements would you most agree with?

a) A car with a large engine size would be predicted to have a high number of engine revolutions per mile.
b) A car with a large engine size would be predicted to have a low number of engine revolutions per mile.
c) Engine size is a poor predictor of engine revolutions per mile.
d) Engine size is independent of revolutions per mile.

It is interesting that the three items which rate the highest under-confidence pertain to correlation. Students clearly have an understanding of the topic, as illustrated by the moderate to easy difficulty, in spite of the low confidence. Students likely encounter the topic elsewhere, such as in a freshman chemistry lab. However, the topic is

typically not taught in an introductory statistics course. The confidence profiles (Figures 13a, 13b, 13c) generally have the appropriate shape. Only the first of these has an inconsistency with confidence 3 a tad higher than confidence 4; this item also has the lowest discrimination of the three. The middle item (#24) has the lowest confidence of incorrect answers for all SCI items (1.92), and the first item (#37) is next-lowest (2.00).

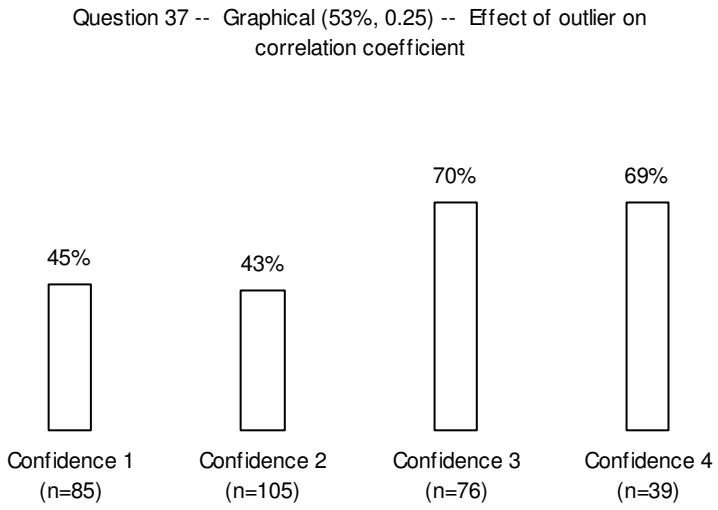


Figure 13a: Confidence profile, by confidence, least-confident correlation item

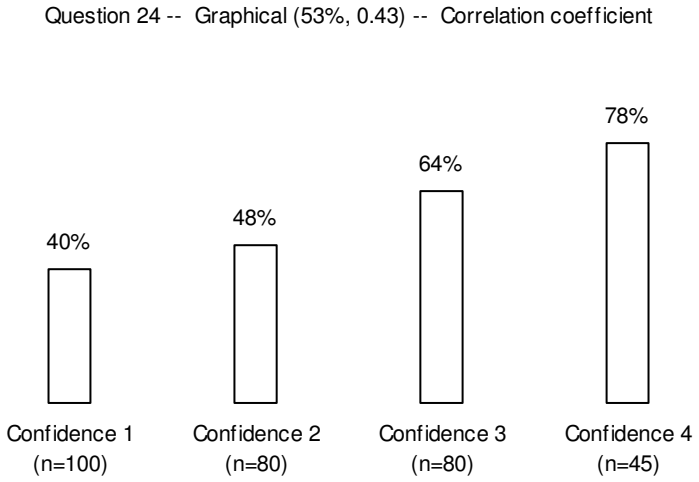


Figure 13b: Confidence profile, by confidence, middle-confident correlation item

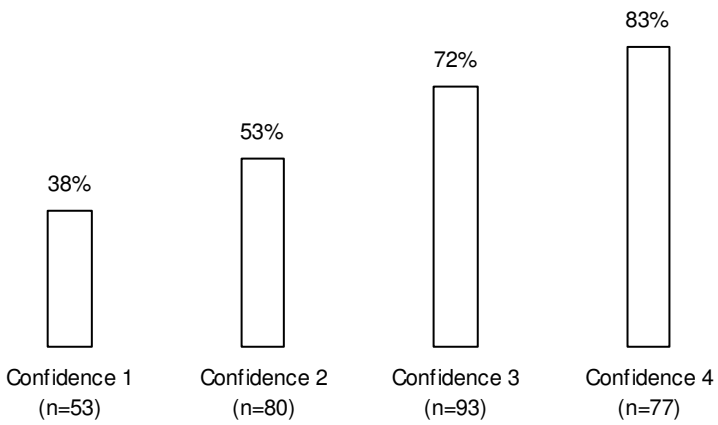


Figure 13c: Confidence profile, by confidence, higher-confident correlation item

Question 22: Inferential (Correct 20th, Confidence 4th)

> You perform the same two significance tests on large samples from the same population. The two samples have the same mean and the same standard deviation. The first test results in a p-value of 0.01; the second, a p-value of 0.02. The sample mean is equal for the 2 tests. Which test has a larger sample size?
>
> a) First test
> b) Second test
> c) Sample sizes equal
> d) Sample sizes are not equal but there is not enough information to determine which sample is larger

This item has a sharp contrast between the low confidence values (1 and 2) and the higher values (3 and 4), Figure 14a. The correct answer suffers from relatively low confidence (9th lowest of correct answers), while the low confidence of the incorrect answers points to guessing (3rd lowest confidence of incorrect answers).

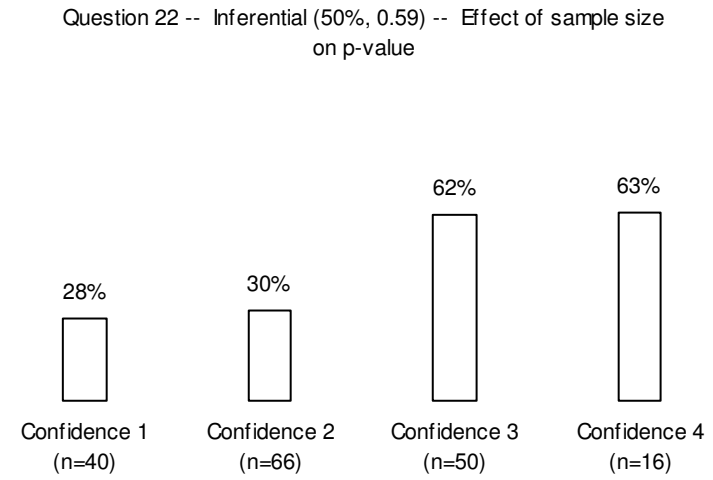


Figure 14a: Confidence profile, by confidence, p-value question

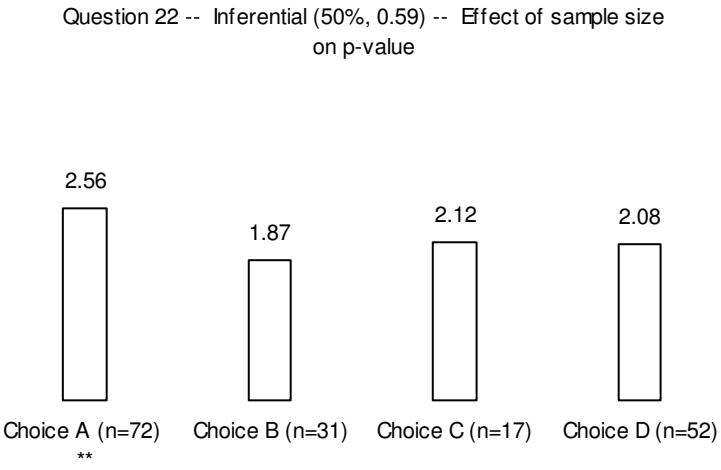


Figure 14b: Confidence profile, by answer, p-value question

5. Further comparisons to previous work

This section details items which are similar to some that have been described in the literature. This necessarily takes on a more comparative slant as opposed to the previous sections which were more exploratory in nature.

Question 12: Descriptive (Correct 34th, Confidence 30th)

> A student attended college A for two semesters and earned a 3.24 GPA (grade point average). The same student then attended college B for four semesters and earned a 3.80 GPA for his work there. How would you calculate the student's GPA for all of his college work? Assume that the student took the same number of hours each semester.
>
> a) $\frac{3.24 + 3.80}{2}$
>
> b) $\frac{3.24(2) + 3.80(4)}{2}$
>
> c) $\frac{3.24(2) + 3.80(4)}{6}$
>
> d) It is not possible to calculate the student's overall GPA without knowing his GPA for each individual semester.

This item was adopted from Pollatsek, *et al.* (1981) and was also used by Mevarech (1983). The SCI results are much more encouraging (75% correct; 42% in Mevarech; 38% and 13% in Pollatsek, *et al*). The Pollatsek numbers were obtained on a pre-test and in interviews with inexperienced statistics students, which possibly explain the low values, but the Mevarech subjects scored nearly the same despite having completed one or two statistics courses. The unweighted option (A), determined by Pollatsek to be a common error (87% in interviews), is unattractive to SCI participants (10%). The SCI item manages a high discriminatory index (0.52), despite attenuation due to a high percent correct, signifying that the higher-ability students are responding correctly at much higher rates than the low-ability students (Figures 15a and 15b).

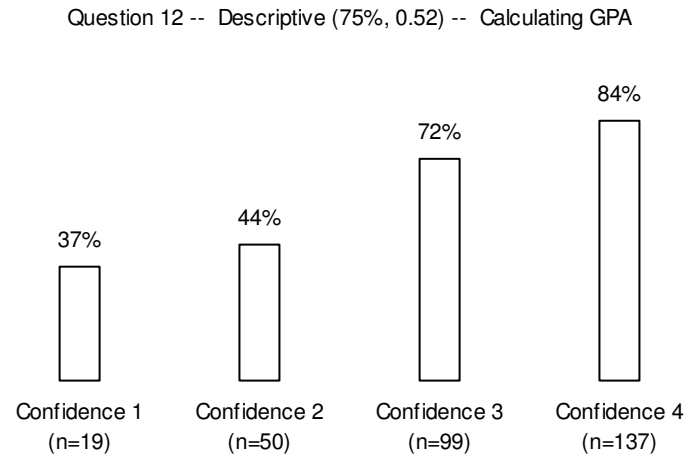


Figure 15a: Confidence profile, by confidence, GPA question

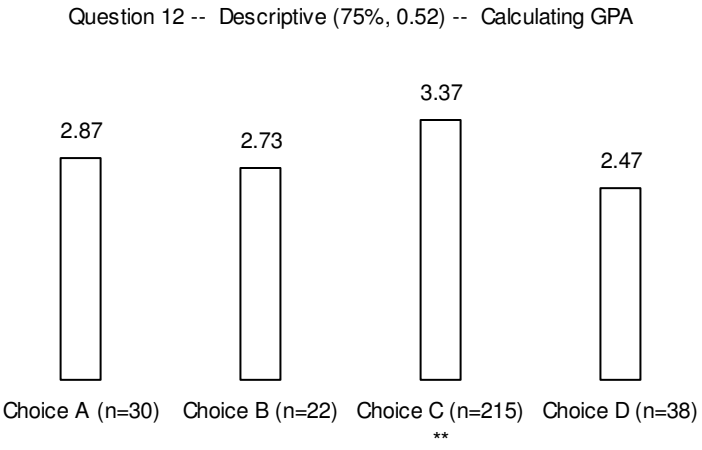


Figure 15b: Confidence profile, by answer, GPA question

Question 16: Probability (Correct 18th, Confidence 32nd)

> A standard deck of 52 cards consists of 13 cards in each of 4 suits: hearts (H), diamonds (D), clubs (C), and spades (S). Five separate, standard decks of cards are shuffled and the top card is drawn from each deck. Which of the following sequences is least likely
>
> a) HHHHH
> b) CDHSC
> c) SHSHS
> d) All three are equally likely.

This item is akin to the study by Kahneman and Tversky (1972) in which subjects were asked about the likelihood of various girl-boy birth orders. The results of the SCI

item are slightly in favor of the correct answer (48%) versus the use of the representativeness heuristic (option A, 40%). Kahneman and Tversky found the analog of option A to be judged as less likely. These differing results may be due to subject demographics: Kahneman and Tversky surveyed primarily high school students, whereas the SCI was administered to college students at the completion of a statistics course or at least a course which taught some aspects of probability and statistics. The presence of a rational thought process is perceptible in the similarity of confidence between option A (3.21) and option D (3.24), Figure 16b. This perhaps causes the low discriminatory index (0.22), as the difference between D (52%) and A (38%) was only slightly larger at confidence 4 than the overall difference across all confidence levels, Figure 16a.

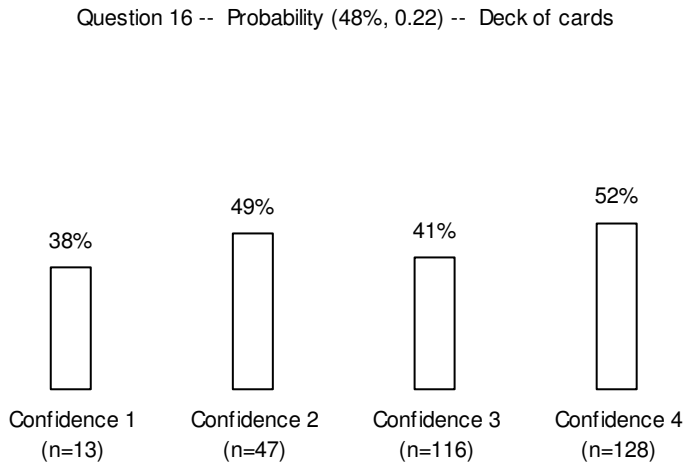


Figure 16a: Confidence profile, by confidence, card sequence question

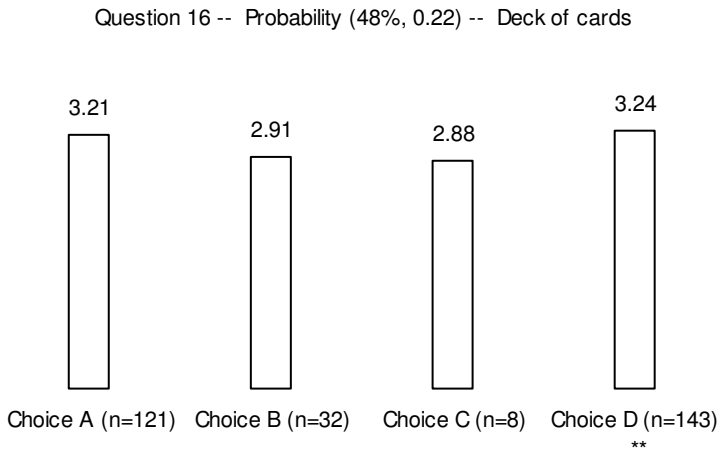


Figure 16b: Confidence profile, by answer, card sequence question

Question 20: Inferential (Correct 9th, Confidence 19th)

> The mean height of American college men is 70 inches, with standard deviation 3 inches. The mean height of American college women is 65 inches, with standard deviation 4 inches. You conduct an experiment at your university measuring the height of 100 American men and 100 American women. Which result would most surprise you?
>
> a) One man with height 79 inches
> b) One woman with height 74 inches
> c) The average height of women at your university is 68 inches
> d) The average height of men at your university is 73 inches

This problem is similar to the "Post Office Problem" studied by Well, *et al.* (1990), in which subjects were tested in their understanding of the law of large numbers. This SCI problem is more extreme in that it asks for a comparison between the smallest possible sample (n = 1) and a much larger one (n = 100). Students must distinguish between two options at each sample size. The nearest analogy with the earlier work is a one-tailed wording, which was found to be the most difficult type of problem (their Experiment 3), with only 25% correct responses.

The SCI lacks an analogous "equal" response, which renders detailed comparison futile. However, one noteworthy similarity was found: option A is, in fact, more unlikely than B, which equates A to the prior study's "reversed" categorization. Although the SCI has a marginally higher percent correct, the ratio of "correct" to "reversed" (25%; 20%)

is nearly identical to the ratio of option D to A (33%; 28%). The fact that option A has nearly equal confidence to option D (Figure 17b) suggests that incorrect students may be responding using some heuristic.

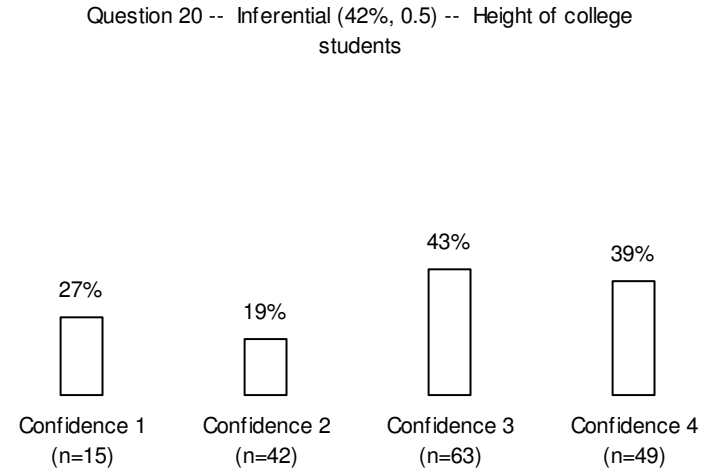


Figure 17a: Confidence profile, by confidence, height question

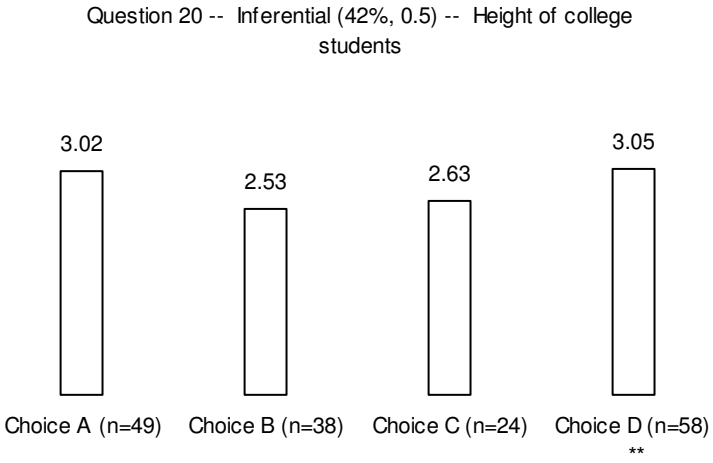


Figure 17b: Confidence profile, by answer, height question

Question 21: Probability (Correct 26th, Confidence 28th)

A meteorologist predicts a 40% chance of rain in London and a 70% chance in Chicago. What is the most likely outcome?

a) It rains only in London
b) It rains only in Chicago
c) It rains in London and Chicago
d) It rains in London or Chicago

This item may tap into two probability misunderstandings. First, the use of percentages calls to mind the *outcome approach* identified by Konold (1989): subjects may view probabilities as single-trial, either/or predictions. Under this rationale, the correct answer would be B because Chicago has a high probability while London is low; this is the most popular incorrect answer (35% overall, 71% of incorrect responses).

The conjunction fallacy, per Tversky and Kahneman (1982), is at first glance present in this item. However, this misconception should be suppressed by the word "only" in options A and B. (Note: Although the intent of option D is an "and/or" interpretation, an "exclusive or" interpretation [i.e., rain in only one] still attains the highest probability; this interpretation arises from the difference between the colloquial "or" and the statistical "union" operation.)

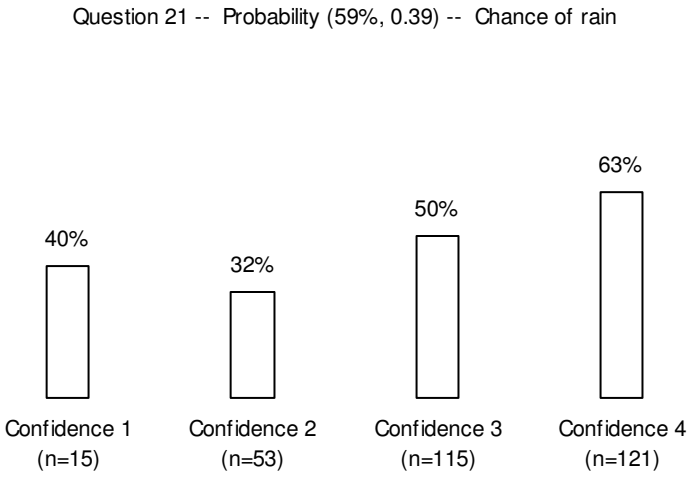


Figure 18a: Confidence profile, by confidence, rain question

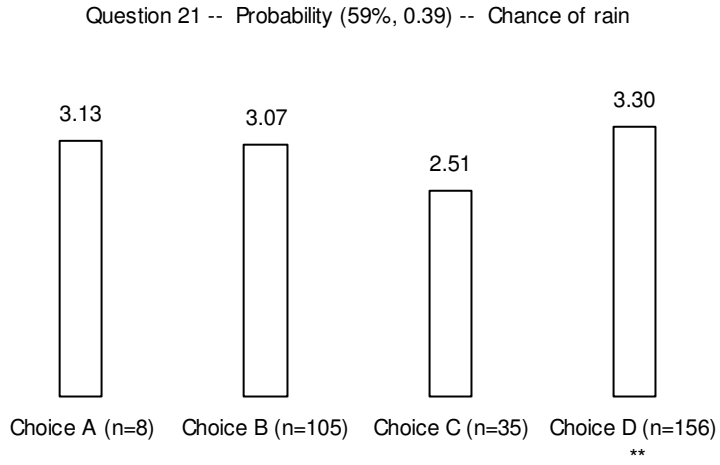


Figure 18b: Confidence profile, by answer, rain question

Question 34: Probability (Correct 33rd, Confidence 26th)

> You are rolling dice. You roll 2 dice and compute the mean of the number rolled, then 6 dice and compute the mean, then 10 dice and compute the mean. One of the rolls has an average of 1.5. Which trial would you be most surprised to find this result?
>
> a) Rolling 2 dice
> b) Rolling 6 dice
> c) Rolling 10 dice
> d) There is no way this can happen

This item probes recognition of the law of large numbers, which is at least implicit in nearly every study of statistical reasoning and oftentimes the explicit focus of the study. The results of the SCI item indicate at worst a rudimentary understanding for most students (73% correct); the higher percent correct at confidence 3 indicates perhaps an incomplete understanding. Option A, the most common and most confidence distracter, may attract students who view 10 rolls as yielding more low-value rolls (i.e., interpreting the item in terms of frequencies rather than means, *cf.* Well, *et al.,* 1990).

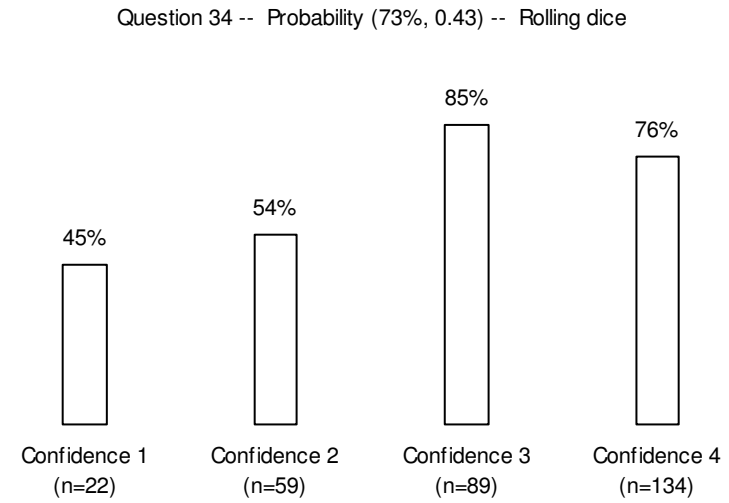


Figure 19a: Confidence profile, by confidence, dice rolling question

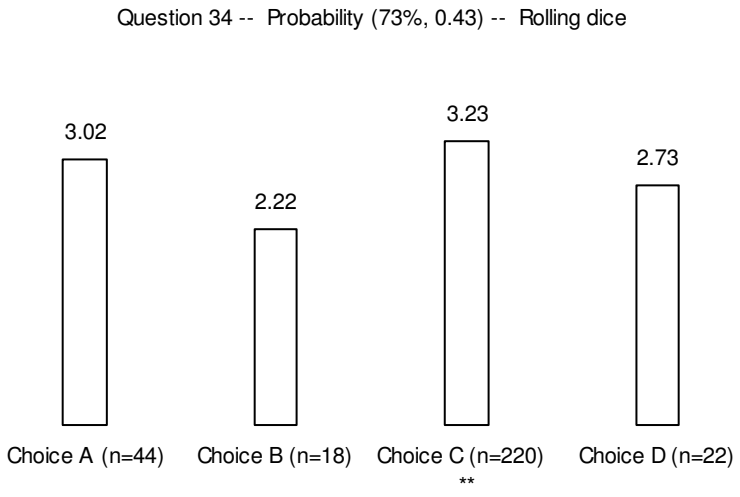


Figure 19b: Confidence profile, by answer, dice rolling question

6. Conclusions

Drawing on the extant literature of statistics education, this study utilized the Statistics Concept Inventory to assess students' confidence in introductory statistics topics. Probability concepts, which can have an intuitive basis, proved to be the area most susceptible to misconceptions. Descriptive statistics, often encountered in educational endeavors before a formal statistics course, yielded the greatest proportion of under-confident items.

These results essentially serve as "mini-interviews," helping to gauge understanding across the test. Probing the reasons why students are confident (or not), through interviews, is a logical next step in the process. This analysis focused on items falling into the over- and under-confident regions. Further analysis could identify guessing (low confidence – low correct) and mastery (high – high). It has been casually observed that items with strong discrimination generally have a positive trend in confidence vs. percent correct (e.g., Figures 11a and 13c). These methods therefore hold promise at relating psychometric properties to confidence. Users of other concept inventories have expressed interest in applying this scale to their instruments.

References

Albert, J.H. 2003. College Students' Conceptions of Probability. *The American Statistician*. 57 (1): 37-45.

Austin, J.D. 1974. An Experimental Study of the Effects of Three Instructional Methods in Basic Probability and Statistics. *Journal for Research in Mathematics Education*. 5 (3, May): 146-154.

Baloğlu, M. 2003. Individual differences in statistics anxiety among college students. *Personality and Individual Differences*. 34 (5) : 855-865.

Bar-Hillel, M. 1974. Similarity and Probability. *Organizational Behavior and Human Performance*. 11: 277-282.

Fong, G.T., Krantz, D.H., and Nisbett, R.E. 1986. The Effects of Statistical Training on Thinking about Everyday Problems. *Cognitive Psychology*. 18: 253-292.

Fong, G.T., and Nisbett, R.E. 1991. Immediate and Delayed Transfer of Training Effects in Statistical Reasoning. *Journal of Experimental Psychology: General*. 120 (1): 34-45.

Garfield, J., and Alghren, A. 1988. Difficulties in Learning Basic Concepts in Probability and Statistics: Implications for Research. *Journal for Research in Mathematics Education*. 19 (1): 44-63.

Kahneman, D., and Tversky, A. 1972. Subjective probability: A judgment of representativeness. *Cognitive Psychology*. 3: 430-454. (reprinted as Chapter 3 in Judgement Under Uncertainty, eds. Kahneman, Slovic, Tversky. 1983)

Kahneman, D., P. Slovic, and A. Tversky, eds. 1982. Judgement under uncertainty: Heuristics and biases. Cambridge University Press: Cambridge.

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

Mevarech, Z.R. 1983. A Deep Structure Model of Students' Statistical Misconceptions. *Educational Studies in Mathematics*. 14: 415-429.

Murtonen, M., and E. Lehtinen. 2003. Difficulties Experienced by Education and Sociology Students in Quantitative Methods Courses. *Studies in Higher Education*. 28 (2): 171-185.

Neter, J., M.H. Kutner, C.J. Nachtsheim, and W. Wasserman. 1996. Applied Linear Statistical Models. 4th Edition. McGraw-Hill: Boston.

Piaget, J., and B. Inhelder. 1951. La genèse de l'idée de hazard chez l'enfant (in English: The Origin of the Idea of Chance in Children). translated by Leake, Burrell, Fishbein and published by W.W. Norton & Company, Inc.: New York, 1975.

Ploger, D., and M. Wilson. 1991. Statistical Reasoning: What Is the Role of Inferential Rule Training? Comment on Fong and Nisbett. *Journal of Experimental Psychology: General*. 120 (2): 213-214.

Pollatsek, A., S. Lima, and A.D. Well. 1981. Concept or Computation: Students' Understanding of the Mean. *Educational Studies in Mathematics*. 12: 191-204.

Pollatsek, A., C.E. Konold, A.D. Well, and S.D. Lima. 1984. Beliefs underlying random sampling. *Memory & Cognition*. 12 (4): 395-401.

Schacht, S., and B.J. Stewart. 1990. What's Funny about Statistics? A Technique for Reducing Student Anxiety. *Teaching Sociology*. 18 (1, Jan.): 52-56.

Simon, J., D. Atkinson, and C. Shevokas. 1976. Probability and Statistics: Experimental Results of a radically different teaching methods. *American Mathematical Monthly*. 83: 733-739.

Rhoads, T.R., and N.F. Hubele. 2000. Student Attitudes Toward Statistics Before and After a Computer-Integrated Introductory Statistics Course. *IEEE Transactions on Education*. 43 (2, June): 182-187.

Tversky, A., and D. Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science*. 185: 1124-1131. (also available as Chapter 1 in same-titled book; eds. Kahneman, Slovic, Tversky. 1982.)

Tversky, A., and D. Kahneman. 1982. Judgments of and by representativeness. Chapter 1 in Judgment under uncertainty; eds. Kahneman, Slovic, Tversky.

Watts, D.G. 1991. Why Is Introductory Statistics Difficult to Learn? And What Can We Do to Make It Easier? *The American Statistician*. 45 (4, Nov.): 290-291.

Well, A.D., A. Pollatsek, and S.J. Boyce. 1990. Understanding the Effects of Sample Size on the Variability of the Mean. *Organizational Behavior and Human Decision Processes*. 47: 289-312.

Book Three

Table of Contents

List of Tables 334
List of Figures 335

IX Statistics as a multi-dimensional construct 336

Abstract 336

1. Introduction 337
    1.1 Factor Analysis 337
2. Reliability 338
    2.1 Background 338
    2.2 Reliability of the SCI 339
3. Exploratory Factor Analysis 340
    3.1 Background 340
    3.2 Results 345
    3.3 Judgment 353
4. Confirmatory Factor Analysis 360
    4.1 Background 360
    4.2 Proposed Models 364
    4.3 Methods 367
    4.4 Preliminary Results 368
    4.5 Revision 369
    4.6 Conclusions 371
    4.7 Extension 375
5. Reliability Revisited 377
    5.1 Enquiring Minds Want to Know 378
    5.2 Cross-validation 380
    5.3 *Cum Grano Salis* 381
6. Conclusion 383

References 384

X Content validity: focus groups and importance survey 386

1. Introduction 386
2. Methods 386
    2.1 Student Interviews 386
    2.2 Faculty Survey 388
3. Results 390
    3.1 Retained Items 392
    3.2 Deleted Items 398
    3.3 Topic Coverage 402
4. Conclusions and Proposals 412
    4.1 Proposals 412

References 414

Appendix 1: 25 retained items 414
Appendix 2: 13 deleted items 426

XI The Concept Inventory Cookbook: A Comparative Study of Methods 433

Abstract 433

1. Introduction 433
2. Force Concept Inventory 434
3. Other Concept Inventories 435
4. Best Practices 437
   4.1 General Considerations 437
   4.2 Teaching Implications 439
   4.3 Similar Concept Inventories 439
   4.4 Content Validity 440
   4.5 Construct Validity 441
   4.6 Predictive and Concurrent Validity 442
   4.7 Reliability 444
   4.8 Discrimination 444
5. Dissemination 446
6. Conclusions 449

References 451

List of Tables

Chapter IX

Table 1: Potential values of Ω....340
Table 2: Comparison between unrotated and rotated solutions....343
Table 3: Comparison between SCI data eigenvalues and random data eigenvalues....349
Table 4: Correlations between eigenvalues and factor numbers....349
Table 5: Highly similar items grouping in Varimax and Promax solutions....355
Table 6: Grouping from Table 5 on a 9-factor Varimax solution....355
Table 7a: Unrotated, 4-factor solution....357
Table 7b: Varimax rotation (orthogonal), 5-factor solution....358
Table 7c: Promax (κ = 3) rotation (oblique), 5-factor solution....359
Table 8: Fit function and test statistics for six structural models....368
Table 9: Fit indices for models, keyed to Table 8....369
Table 10: Fit function and test statistics for preferred structural models....370
Table 11: Fit indices for preferred structural models....370
Table 12: Fit summary for 1- to 4-factor exploratory solutions....371
Table 13: Fit summary comparison for uni-dimensional and G + 4 models....371
Table 14: Summary of model estimates for models (1) and (5)....374
Table 15: Deleted items for 25-item SCI....379
Table 16: Retained items for 25-item SCI....379
Table 17: Uni-dimensional model fit summary for 38-original and 25-cut SCI....380
Table 18: Summary statistics for 1000 replicates of a 15-item-removed SCI....380

Chapter X

Table 1: Item Summary Statistics for full 38-item SCI....391
Table 2: Demographics summary, with *n* = 24....403
Table 3: Summary statistics for New and Old surveys....404
Tables 4: Topics surveys results....408
Table 5: Coverage of Top 25 Important Topics, for 25-item SCI....411

Chapter XI

Table 1: List of Concept Inventories (CI) and similar instruments....436
Table 2: Concept Inventory scores....438
Table 3: Cronbach's alpha for concept inventories....444

List of Figures

Chapter IX

Figure 1: Reliability estimates vs. number of factors 340
Figure 2: Graphical representation of successful rotation 344
Figure 3: Scree plot, unrotated principal components, Eigenvalue vs. Factor number ...347
Figure 4: Scree plot, zoomed view of Figure 3.....348
Figure 5: Number of factor loadings per item vs. loading threshold.....350
Figure 6a: Number of factor loadings per item vs. rotation method, 4-factor.....351
Figure 6b: Number of factor loadings per item vs. rotation method, 5-factor.....352
Figure 7: Number of factor loadings per item vs. delta (δ) 352
Figure 8: Number of factor loadings per item vs. kappa (κ) 353
Figure 9: Difficulty vs. factor number, promax rotation 356
Figure 10: Structural models, (a) confirmatory and (b) exploratory 361
Figure 11: One-factor "G" model for SCI 365
Figure 12: Correlated errors for similar items 365
Figure 13: Specific factors modeled as external to "G" 366
Figure 14: Sub-domain structure, showing relationship between similar items.....367
Figure 15: Proposed sub-domains for Statistics discipline.....376
Figure 16: Proposed sub-domains for Physics discipline 376
Figure 17: Alpha vs. number of items removed 378
Figure 18: Cross-validation summary, count of items falling in Bottom 15 381

Chapter X

Figure 1: Sample of online survey format 389
Figure 2: Comparison of New and Old topic rankings.....404

Chapter XI

Figure 1: General model for concept inventory development (Beichner, 1994) 437
Figure 2: Discriminatory Indices from the Strength of Materials CI.....446
Figures 3: Cumulative examinees across project years for engineering concept 447

CHAPTER IX

Statistics as a multi-dimensional construct

sta·tis·tics
(stə-tĭs tĭks) *n.*

The mathematics of the collection, organization, and interpretation of numerical data, especially the analysis of population characteristics by inference from sampling.[dictionary.com]

Statistics is the science of gaining information from numerical data. (Moore, 1997)

Everything dealing with the collection, processing, analysis, and interpretation of numerical data belongs to the domain of statistics. (Johnson, 1994)

A quick glance at the newspaper yields statistics that deal with crime rates, birth rates, average income, average snowfall, and so on. By a common definition, therefore statistics consist of facts and figures. (Gravetter and Wallnau, 1988)

In short, statistics is the science of data. (Mendenhall and Sincich, 1995)

Abstract

This chapter analyzes the Statistics Concept Inventory from a multi-dimensional perspective. Based on 295 students who completed the full SCI at the end of Fall 2005, an exploratory factor analysis suggests a uni-dimensional structure for the instrument, although some small item groups of substantive meaning were identified. These results were subjected to a confirmatory factor analysis, verifying the potential to identify sub-domains of the Statistics discipline, while confirming that a uni-dimensional model is

most accurate for analyzing the SCI in its present form. Finally, a blueprint for a shorter instrument is proposed.

1. Introduction

There is no shortage of definitions for Statistics, but the field certainly incorporates the ability to analyze and interpret data. This breadth begs the question of what skill-sets are present within the discipline. The purpose of this chapter is to analyze data from the Statistics Concept Inventory to answer this question or perhaps to state that no answer exists.

The first section describes multi-dimensional test reliability. Before diving into the structure of the SCI, it is important to determine if the data are reliable from this multi-dimensional perspective.

The bulk of this chapter focuses on factor analysis as a tool to determine the relationship(s) between SCI questions and thus the under-lying structure of the test and hence statistics as a discipline. The background on factor analysis is intended to walk the reader through the solution procedure, although it is not intended to be exhaustive or mathematically rigorous.

1.1 Factor Analysis

Factor analysis was developed in the early 20th century by those concerned with the measurement of intelligence. The goal was to determine ability sub-domains separate from a larger general ("G") intelligence, by searching for groupings in the correlation matrix of item or test scores. If a group of items correlate highly with each other but not with other items, it is tenable that these items constitute a unique ability.

A factor analysis is not straight-forward, and many decisions must be made, often subjectively, to determine the best model for the data. These decisions include the estimation method, number of factors, and basic model class (exploratory vs. confirmatory). Details are provided as they arise in later sections.

Regardless of the methods, the basic goal of a factor analysis is to re-produce a correlation matrix in as few factors as possible. For a given model, a portion of each variable is predicted by the model (communality), while the remainder is error (uniqueness). A basic mathematical statement is given below for those so inclined (Johnson and Wichern, 2002); other representations are possible.

$$\underset{(p\times 1)}{X} = \underset{(p\times 1)}{\mu} + \underset{(p\times m)}{L}\ \underset{(m\times 1)}{F} + \underset{(p\times 1)}{\varepsilon} \qquad (1)$$

where: X is the observed data of *p* variables
μ is the mean vector
L is the matrix of factor loadings
F is the vector of *m* common factors
ε is the error

2. Reliability

2.1 Background

Coefficient alpha equals the true reliability only when the items are parallel or at least tau-equivalent; this essentially means the items measure the same thing, i.e., the test is uni-dimensional. When these conditions are not met, alpha serves as a lower bound to reliability. Two alternatives, theta and omega, provide more accurate estimates of reliability in these circumstances. Theta is based on a principal components analysis (PCA), with the formula in (2). Theta has been shown to be a maximized alpha with respect to a weighting vector applied to the items (Green and Carmines, 1979).

$$\theta = \frac{k}{k-1}\left(1 - \frac{1}{\lambda_1}\right) \tag{2}$$

where: k is the number of items
λ_1 is the maximum eigenvalue from PCA

Omega is based on the common factor model. Omega suffers from indeterminancy unless the number of factors is fixed, because the communalities depend on the number of extracted factors (Armor, 1974). The basic formula is below.

$$\Omega = 1 - \frac{\sum \sigma_i^2 - \sum \sigma_i^2 h_i^2}{\sum\sum \sigma_{x_i x_j}} \tag{3}$$

where: σ_i^2 is the covariance of the i^{th} item
h_i^2 is the communality of the i^{th} item
$\sum\sum \sigma_{x_i x_j}$ is the sum of the covariances among items

When using correlations, the formula reduces to the following.

$$\Omega = 1 - \frac{a - \sum h_i^2}{a + 2b} \tag{4}$$

where: a is the number of items
b is the sum of the correlations among the items

The three metrics (α, θ, Ω) have the following relationship. Equality holds when items are parallel.

$$\alpha \le \theta \le \Omega \tag{5}$$

2.2 Reliability of the SCI

Figure 1 shows the reliability of the SCI, as measured by α, θ, and Ω. Because omega depends on the number of factors, the abscissa is the number of retained factors, ranging from 1 to 38. The number of factors to retain determines the reported value of Ω. Table 1 shows some potential values.

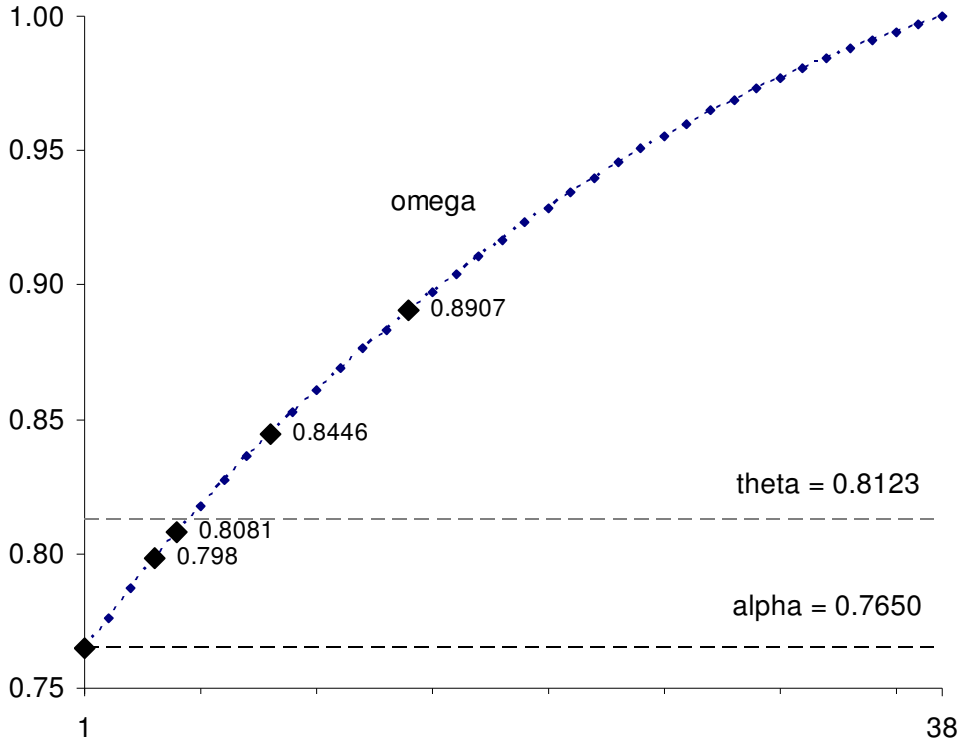


Figure 1: Reliability estimates vs. number of factors

Table 1: Potential values of Ω

Factors	Ω	Comment
1	0.7647	Equals α [within rounding error]
4	0.7980	Largest secondary elbow on scree plot
5	0.8081	G plus four-factor model?
9	0.8446	Tertiary elbow
15	0.8907	Eigenvalues > 1

3. Exploratory Factor Analysis

3.1 Background

An exploratory factor analysis involves many decisions to determine a best solution. The considerations ultimately boil down to methods at determining a *simple structure* for the factor solution, which is essentially a synonym for *parsimony*. Dating

back to Thurstone, various criteria exist for determining the simple structure. The basics can be stated in the following rules:

- Each variable should load on as few factors as possible, preferably one.
- The number of variables on each factor should be minimized.

Extraction method

Many techniques exist for finding a solution to a factor analysis. In EFA, the most common procedure is *principal components* (PC), based on the early work of Karl Pearson and later expanded by Hotelling. PC maximizes the variance extracted on the first factor, with maximal residual variance extracted orthogonally on the second factor, and so on (Harman, 1976).

Being based on variance, PC does a poor job of explaining the overall covariance structure of a dataset. *Maximum likelihood* (ML) and *least squares* (LS) are two methods able to explain the full covariance matrix. These methods are most commonly encountered in confirmatory factor analysis (structural equation modeling); descriptions will be presented in Section 4.

Number of factors

The number of factors retained is crucial to the interpretability of the factor solution. Many rules exist for determining this number. The techniques to be compared are described as follows:

- *Eigenvalues > 1* : This assessment states that factors having eigenvalues greater than one ($\lambda > 1$) are retained. While apparently arbitrary, this rule has been shown to be both theoretically and empirically sound (Rummel, 1970). One criticism of this method arises when eigenvalues are near one. For example, an eigenvalue of

1.01 would be retained, while one of 0.99 would be rejected, although they are essentially equal.

- *Scree plot* : A scree plot displays the eigenvalues vs. factor number. The number of retained factors is up to the point where an approximate discontinuous dropoff occurs. For instance, the sequence {1.72, 1.31, 1.04, 0.94, 0.68, 0.59} has a large drop from 0.94 to 0.68, with a less severe decrease to 0.59 on the next factor. Therefore, the first four factors would be retained. (Rummel, 1970)
- *Parallel analysis* : Also based on eigenvalues, this technique compares the obtained solution to that of a random dataset. The factors are retained which are greater than those from the random data (Loehlin, 2004).
- *Meaningfulness / Interpretability* : These subjective criteria are similar. They are based on the researcher's ability to assign meaning to a solution. The smaller factors (lower λ) are only retained if there is an interpretation applicable to the data (Rummel, 1970).

Assigning items to factors

In practice, every variable will load on every factor. The interpretability of the solution depends on how large a loading need be to retain as something other than random error. The authors of the Concept Inventory for Natural Selection (Anderson, *et al.,* 2002) used a threshold of 0.40. Harman (1976) provides a formula for the standard error of factor coefficients, shown below.

$$\sigma_a = \frac{1}{2}\sqrt{\left(3/r - 2 - 5r + 4r^2\right)/N} \qquad (6)$$

where: σ_a is the standard error
r is the average value in the correlation matrix
N is the sample size

Rotation

Rotation is a technique to enhance the interpretability of a solution. In principal components, for example, the goal is to maximize the variance extracted on the first factor and the residual variances on subsequent factors. However, the solution may prove difficult to reconcile with a theoretical model of the proposed construct.

Rummel (1970, pp. 373-378) illustrates the power of rotation for a case with eight variables loading on two factors. Table 2 shows the factor loadings for the unrotated and rotated solutions. The unrotated solution has high loadings for each variable on both factors, but the second-factor loadings cluster into positive and negative groups. The rotated solution, meanwhile, individuates the variables into distinct factors, with very high loadings on one and near-zero loadings on the other.

Table 2: Comparison between unrotated and rotated solutions

	Unrotated		*Rotated*	
	Factor 1	Factor 2	Factor 1 *	Factor 2 *
1	0.76	0.45	0.25	0.85
2	0.83	0.53	0.27	0.95
3	0.59	0.73	-0.05	0.94
4	0.63	0.66	0.02	0.91
5	0.77	-0.60	0.98	0.08
6	0.64	-0.71	0.97	-0.08
7	0.72	-0.53	0.91	0.09
8	0.81	-0.52	0.96	0.15

With two factors, a graphical explanation is possible (Figure 2). The unrotated solution (axes S_1, S_2) show the approximate equal loadings on factor 1, along with the opposite-signed clustering along factor 2. The rotated solution (axes S_1^*, S_2^*) successfully captures the clustering of variables 1-4 on the positive side of S_2, with the unrotated negative S_2 variables now clustering along the rotated S_1^*.

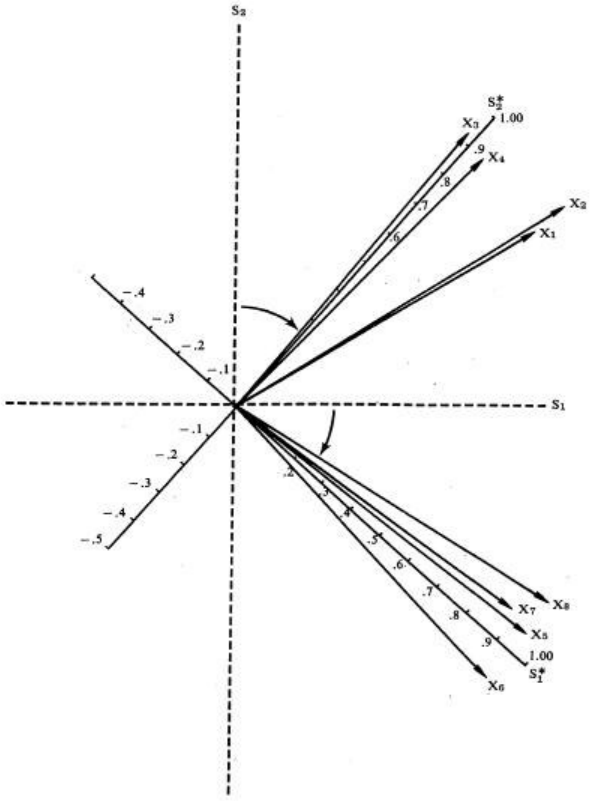


Figure 2: Graphical representation of successful rotation (Rummel, 1970, p. 377)

While a graphical rotation is helpful for illustrative purposes, it is not possible for solutions with more than two factors. Two classes of analytical techniques exist. The first of these, orthogonal rotations, maintains perpendicular (i.e., uncorrelated) axes. Three common criteria for orthogonal rotations are described briefly below (Harman, 1976):

- *Quartimax* seeks to maximize the variance of the squared loadings, which equates to maximizing the fourth power of the loadings (hence the name). The simplification is thus along variables (rows of the factor matrix).
- *Varimax*, conversely, seeks simplicity among the factors (columns of the factor matrix). This is accomplished by seeking the maximized variance across the retained factors.

- *Orthomax* combines the above criteria in a linear combination. The optimal balance between the two methods is referred to as *equamax*.

An oblique rotation, on the other hand, allows the factors to be correlated. Graphically, this amounts to non-perpendicular axes. Factor loadings can be assessed either by parallel (pattern) or perpendicular (structure) projections onto the oblique axes. The pattern matrix is considered best for determining clusters of variables (Rummel, 1970), which is the purpose of this analysis. SPSS ™ allow two oblique rotation algorithms. These are briefly described below:

- *Direct Oblimin* seeks "a simple structure solution by minimizing a function of the primary-factor-pattern coefficients" (Harman, 1976, p. 321). A parameter delta (δ) is included as a measure of the correlation between primary factors; δ varies from -∞ (uncorrelated; less oblique) to 1 (most oblique). Typical values of δ are in the range 0 to –10.
- *Promax* is essentially a tweaking of the orthogonal varimax rotation. After normalizing the pattern rows and columns, loadings are raised to the exponent kappa (κ) to find the best solution. Typically, κ of 4 is considered best, but lower values may perform better for neatly-structured data (Rummel, 1970).

3.2 Results

This section documents the considerations that go into choosing the optimal model. The decision will follow in section 3.3 because the decision factors interact and the accumulation of evidence must be considered.

Extraction Method

Principal components is the selected extraction method for this exploratory analysis. It is acknowledged that this is not ideal for this SCI data, where explaining the overall covariance is ideal. There are two reasons for this decision: 1) PC is the most common method for EFA in the literature; 2) maximum likelihood will be used in the confirmatory analysis, which can allow direct comparison to these results by choosing the appropriate model.

As a quick comparison of PC and ML, the first-factor loadings from a four-factor solution were compared. High correspondence was found, with a mean absolute difference of 0.030 between loadings from the two methods, across the 38 items. Acknowledging that principal components is not ideal, it is retained as the method for this exploratory analysis to maintain a common language with concept inventory literature, for example, as a way of walking the reader through the considerations before leading the way to a confirmatory paradigm.

Number of factors

> *For illustrative purposes, these results are based on the unrotated principal components solution.*

The scree plot is shown in Figure 3. The first factor accounts for three times as much variance as the second factor (12.6%, 4.2%). Eigenvalues are greater than one ($\lambda > 1$) up to the 15th factor.

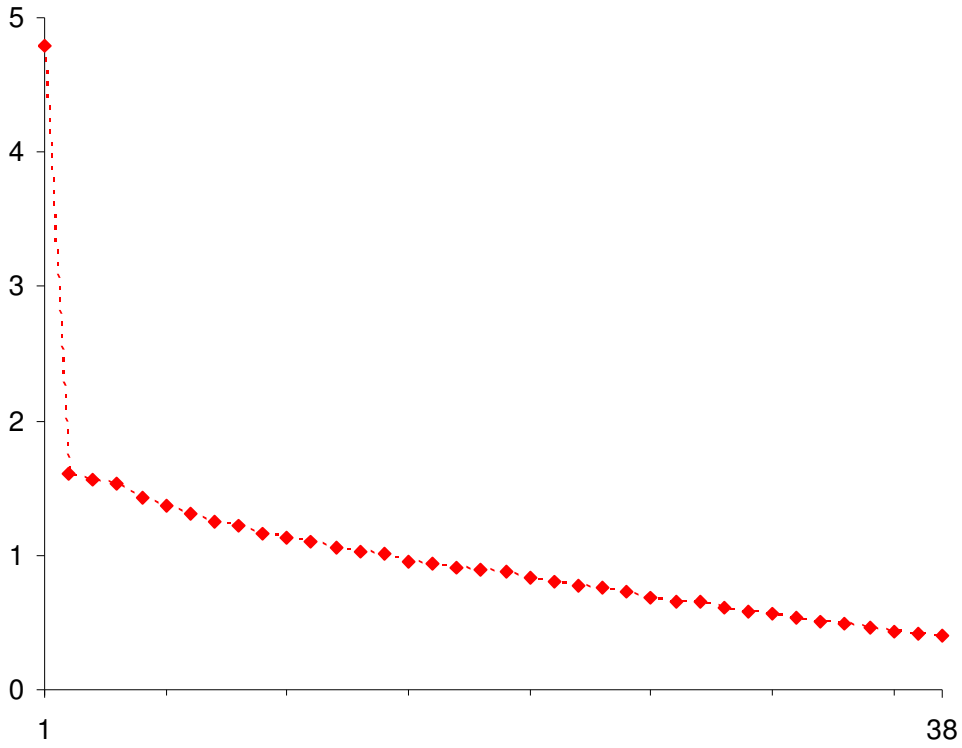


Figure 3: Scree plot, unrotated principal components, Eigenvalue vs. Factor number

To aid in a decision from the scree plot, the above graph is zoomed down to 15 factors and eigenvalues ranging from 1.0 to 1.8. This is depicted in Figure 4. The largest secondary cutoff (after factor 1) is after factor 4. A tertiary cutoff can be taken after factor 9 as well. However, these distinctions are rather fine, and the entire plot could be viewed as scree ("after G, it's all scree").

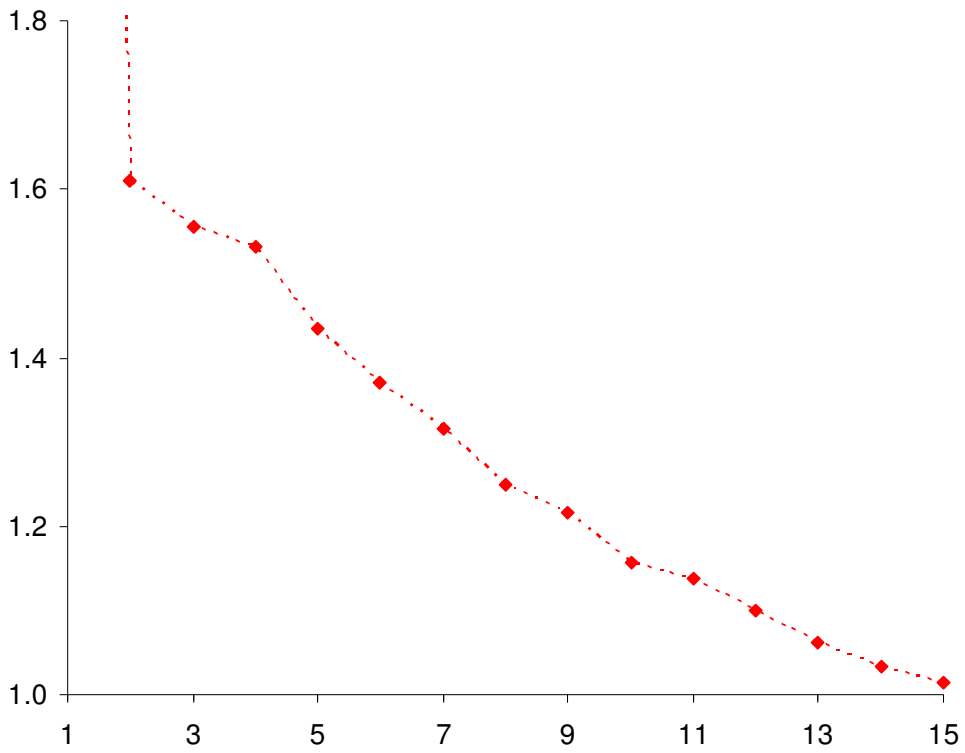


Figure 4: Scree plot, zoomed view of Figure 3

A parallel analysis was conducted by generating random dichotomous data of size 295 × 38 (i.e., subjects who completed the full SCI × the number of items). Four replications were produced using the *rand* function in Excel, with values rounded to whole numbers. These values were then run using an unrotated principal components solution, with the eigenvalues compared to those obtained as in the scree plots.

Table 3 displays the results of the parallel analysis, compared to the SCI solution. The first factor is retained on all four replications. Factors 2 and 3 find no eigenvalues less than the SCI solution, although Factor 4 is mixed with replications two and three saying to retain Factor 4, while replications one and four (and the mean) say to reject Factor 4. No factors above four are retained (showing up to Factor 7), with the difference

between the random solution and the SCI fit generally increasing as the factor number increases. This analysis suggests either a four-dimensional or one-dimensional solution.

Table 3: Comparison between SCI data eigenvalues and random data eigenvalues

Factor	SCI	Random				
		1	2	3	4	Mean
1	4.78	*1.79*	*1.85*	*1.73*	*1.81*	*1.80*
2	1.61	1.64	1.68	1.67	1.68	1.67
3	1.56	1.60	1.62	1.57	1.62	1.60
4	1.53	1.56	*1.50*	*1.51*	1.56	1.54
5	1.43	1.54	1.48	1.46	1.51	1.50
6	1.37	1.44	1.44	1.45	1.42	1.44
7	1.32	1.39	1.42	1.42	1.38	1.40

An interesting approach is to combine the scree plot with the parallel analysis, investigating the correlation between eigenvalue and factor number. Table 4 displays the results. For example, the row "2 to 38" is the correlation between that respective set of eigenvalues and the numerals 2 to 38. For the random data, all 38 factors were used.

Table 4: Correlations between eigenvalues and factor numbers

Factors used	Correlation
Random 1	-0.9871
Random 2	-0.9869
Random 3	-0.9909
Random 4	-0.9855
1 to 38	-0.6980
2 to 38	-0.9880
2 to 15	-0.9874
16 to 38	-0.9985
2 to 4	-0.9751
5 to 9	-0.9952
2 to 5	-0.9692
6 to 9	-0.9922
10 to 15	-0.9942

This analysis confirms what is obvious from the scree plot: all factors after the first are essentially random. The correlation between remaining factors (2 to 38, -0.9880) is essentially equal to those from the random data, while the first factor is not of the same ilk (1 to 38, -0.6980). Any further analysis is not meaningful. However, the fit {2 to 4, 5

to 9} is slightly better than {2 to 5, 6 to 9}. One could therefore say, with very little confidence, that four factors is a more plausible solution.

A meaningfulness decision clearly points to a general "G" factor due to the large first eigenvalue. The previous factor analytic study (Book One) concluded in favor of a five-factor solution, based on a confirmatory model. Thus far, a four-factor solution is preferred, although retention of the fifth factor would not be a stretch and could allow comparison with earlier results.

Assigning items to factors

> *This analysis is based on the five-factor principal components solution with varimax rotation.*

Figure 5 shows the number of factor loadings per item across five values of factor loading threshold. The black portion represents the number of items loading on one factor, while the lighter portions correspond to zero (below) and two-plus (above); for example, using a threshold of 0.1, five items load on only one factor.

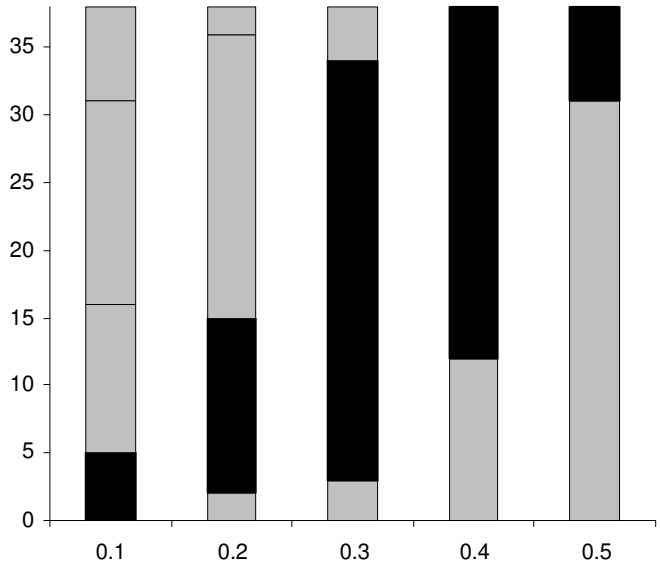


Figure 5: Number of factor loadings per item vs. loading threshold

The optimal value from Figure 5 is 0.3. Finer analysis revealed that values between 0.31 and 0.33 are slightly better but not sufficient to warrant such distinction. The standard error estimate of Harman (1976) is 0.18 (equation 6), which is a poor estimate to use as the threshold.

Rotation

Figures 6a and 6b show the item factor-loading counts for 4-factor and 5-factor solutions, respectively. The unrotated 4-factor and varimax 5-factor each place 31 items on only one factor. As an unrotated solution, the 4-factor places the bulk of items (21) on factor 1, while the varimax 5-factor spreads the items across more evenly across factors (e.g., 11 each on factors 1 and 2).

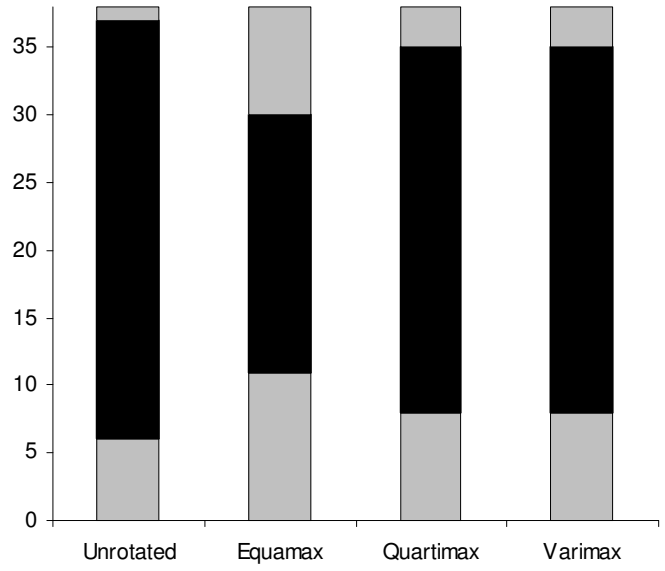


Figure 6a: Number of factor loadings per item vs. rotation method, 4-factor

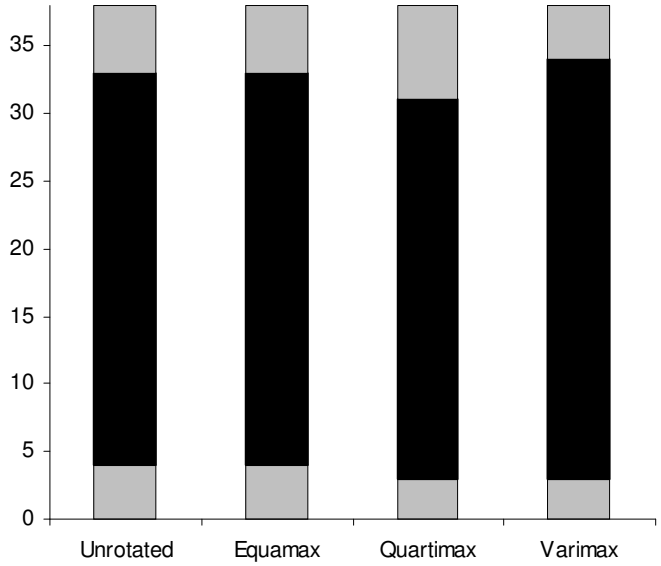


Figure 6b: Number of factor loadings per item vs. rotation method, 5-factor

Figure 7 displays the number of loadings per item vs. the parameter δ for the oblique direct oblimin 5-factor solution. The number of items loading on one factor is relatively constant (range 30 to 32). The average (magnitude) factor correlation varies from 0.098 (δ = 0) to 0.116 (δ = -5).

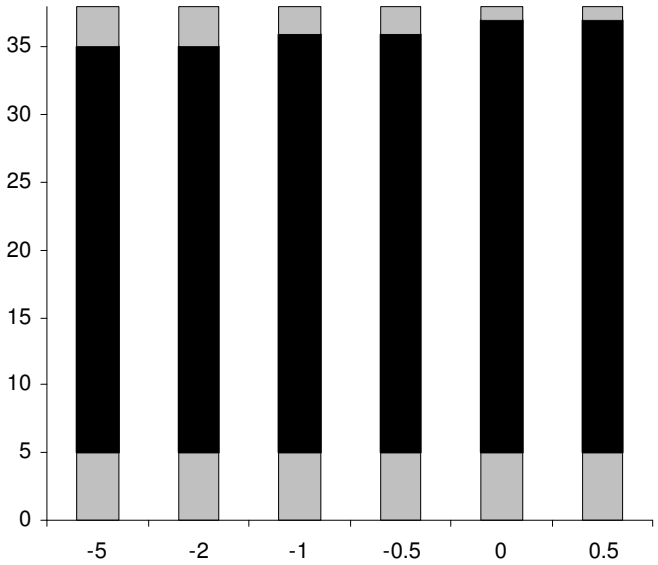


Figure 7: Number of factor loadings per item vs. delta (δ);
5-factor direct oblimin rotation

Analogously, Figure 8 illustrates the promax rotation with the parameter kappa (κ) on the abscissa. The number of items loading on one factor (range 30 to 33) is a tad improved over Figure 8. The average (magnitude) factor correlation shows more variability than the direct oblimin solution, ranging from 0.070 (κ = 2) to 0.206 (κ = 8).

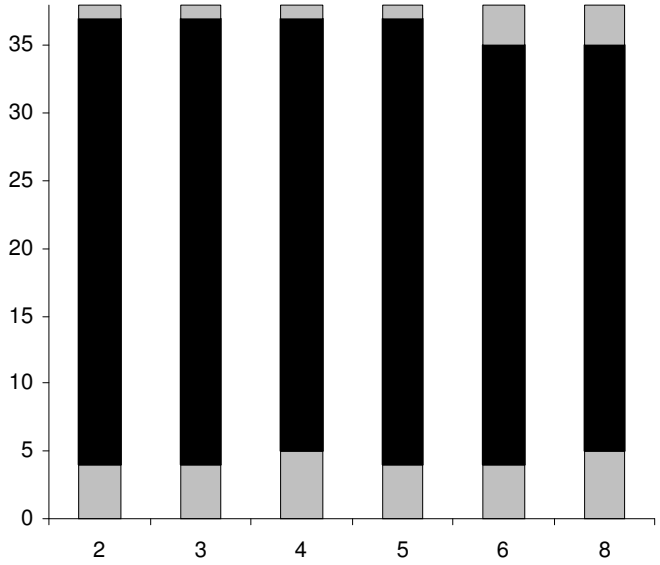


Figure 8: Number of factor loadings per item vs. kappa (κ); 5-factor promax rotation

3.3 Judgment

Closing Arguments

The abundance of solution procedures renders a definitive solution difficult. This exploration best serves as a commencement for the confirmatory analysis to follow. The loading threshold of 0.30 proves to be the best along most combinations of other settings, in many more combinations than were illustrated.

The lingering question is the plausibility of the SCI as a multi-dimensional instrument. Among orthogonal rotations, the preferred four-factor solution is unrotated, which places nearly half the items on factor 1; the optimal five-factor solution is the

varimax rotation, which spaces the items along more factors. The oblique rotations were slight improvements, suggesting that the factors may have slight correlations (at most 0.20), which seems reasonable if a multi-dimensional structure exists. These rotations also placed fewer items (9) on factor 1 in the best fits.

Three solutions are compared as potentially optimal: unrotated 4-factor, varimax (orthogonal) 5-factor, and promax (oblique) 5-factor with $\kappa = 3$. The results can be found in Tables 7 (a, b, c) on the pages following this discussion. All loadings above 0.30 are displayed, along with the question number, proposed topic area, and description. All negative loadings are small-magnitude; the values listed are not absolute values. The horizontal lines group items according to the largest loading of each item, ordered by loading. The final group has no loadings above 0.30. The summary column contains the counts of items on that factor by topic area. For the few items loading on multiple factors, the grouping is listed according to the larger loading, and these items *are* counted twice in the summary column.

The unrotated solution is difficult to interpret because most items load on the first factor, while the remaining factors show no apparent similarities. These results point to the uni-dimensional "G" model for the SCI.

The varimax and promax rotations are nearly identical; only item 22 groups differently (factor 1 in varimax, unclassified in promax). There interpretation can therefore be combined. The gross summary does not point to clustering as hypothesized along the four topic areas. However, items which are highly similar do tend to group along the same factor; these are highlighted in Table 5.

Table 5: Highly similar items grouping in Varimax and Promax solutions

Items	Common topic
20, 27	Height of college students
26, 29	Standard deviation calculation properties
17, 35	Properties of confidence intervals
24, 38	Correlation coefficient (item 37 does not group, however)
2, 36	Choice of test statistic
18, 22	p-value (Varimax only)

The question of a higher-order structure is examined in Table 6, for a 9-factor Varimax rotation. These highly similar items show a lesser degree of clustering compared to the five-factor solution. The simple structure is also less realized, with 28 items loading on one factor, compared with 31 in the five-factor solution.

Table 6: Grouping from Table 5 on a 9-factor Varimax solution

Items	1	2	3	4	5	6	7	8	9
20	0.36							-0.40	
27		0.73							
26	0.47	0.35							
29		0.46							
17	0.54								
35	0.57								
24								0.48	
38	0.45						0.33		
2					0.51				
36					0.63				
18		-0.30		0.38					
22	0.35								

Deciding between varimax and promax is a matter of whether the factors should be allowed to correlate. The average (magnitude) correlation for the promax rotation is 0.120. Factors 1, 2, and 3 correlate most highly ($r_{12} = 0.199$, $r_{13} = 0.311$, $r_{23} = 0.253$), while other interfactor correlations are less than 0.10 in magnitude. A moderate amount of correlation between factors is plausible if a multi-dimensional structure is tenable.

A concern in an exploratory analysis is that items may group along difficulty rather than in a substantively meaningful manner. Figure 9 shows the item difficulties, by

factor, for the promax solution. The first factor contains an abundance of easy items, while the remaining factors ("6" = unclassified items) are generally clustered around 0.50, with no tendency towards a difficulty grouping.

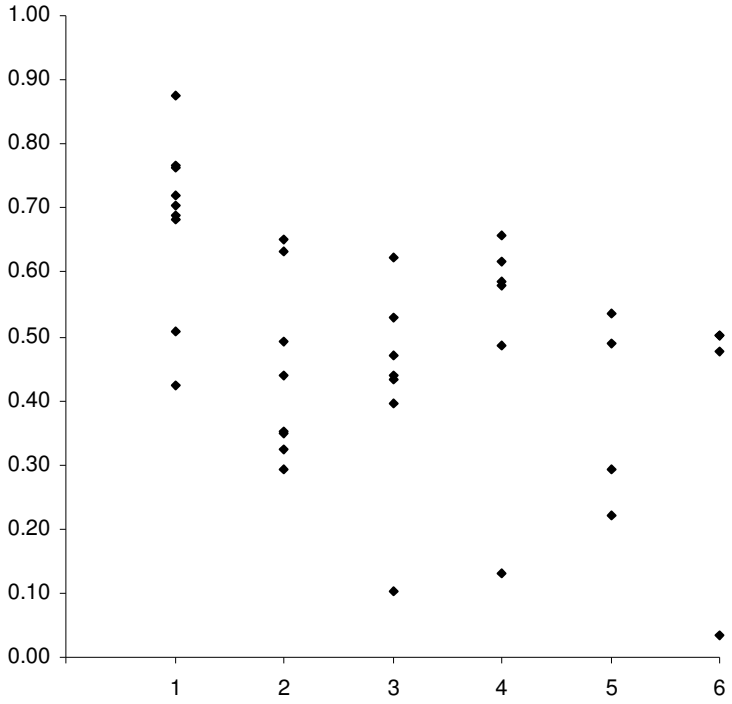


Figure 9: Difficulty vs. factor number, promax rotation

Verdict

Two similar models were found which encouragingly grouped highly similar items along the same factor. However, less than 30% of the overall variance (28.7%) was accounted for, which suggests that SCI items are primarily unique. A definitive conclusion is not warranted at this time. These results can best be used as a guide in designing a confirmatory factor analysis, which is the topic of the next section.

Table 7a: Unrotated, 4-factor solution

Q#	Area	1	2	3	4	Description	Summary
29	D	0.59				Standard deviation equals -2.30	
33	P	0.59				Temperature on October 1	
26	D	0.52				Standard deviation equals zero	
12	D	0.51				Calculating GPA	
22	I	0.49				Effect of sample size on p-value	
8	D	0.49				Percentile	
35	I	0.49				Sample size effect on confidence intervals	
9	D	0.47				Temperatures for a week in August	
31	P	0.45				Error rate in a manufacting process	
4	P	0.43				Babies born in a hospital	P 5
11	D	0.41				Least impacted by outliers	D 9
25	G	0.41	0.40			Parent distribution of a sample	I 6
30	G	0.41				Variability of a histogram	G 3
34	P	0.40				Rolling dice	
23	D	0.40				Which would have a normal distribution?	
21	P	0.40				Chance of rain	
17	I	0.38				Meaning of 95% confidence interval	
20	I	0.38				Height of college students	
38	D	0.36				Correlation engine size	
10	I	0.36				Bottling company	
27	I	0.34				Sampling method for height of college students	
15	D	0.32				Which describes central tendency?	
18	I		0.43			Meaning of p-value = 0.10	P 1
13	P		0.35			Customer service waiting time	D 0 I 1 G 0
14	G			0.44		20 samples of 10 points each	P 0
32	I			0.37		Conclusion of p=0.05	D 1 I 1
6	D			0.31		Olympic track team	G 1
36	I				0.53	Appropriate test for chemical company	
3	D				0.46	Household income	P 0
2	I				0.42	Diet plan	D 1 I 2
37	G				0.40	Effect of outlier on correlation coefficient	G 2
28	G				0.30	Histogram of class grades	
1	P					Testing a disease	
5	P					Coin flipped twelve times	P 2
7	G					Which graph is from a different set of data?	D 0
16	P					Deck of cards	I 0
19	I					Which is true of a t-distribution?	G 1
24	G					Correlation coefficient	

Table 7b: Varimax rotation (orthogonal), 5-factor solution

Q#	Area	1	2	3	4	5	Description	Summary
29	D	0.64	0.30				Standard deviation equals -2.30	
9	D	0.50					Temperatures for a week in August	
34	P	0.50					Rolling dice	
27	I	0.49					Sampling method for height of college students	P 1
20	I	0.48					Height of college students	D 7
12	D	0.46					Calculating GPA	I 4
26	D	0.43					Standard deviation equals zero	G 0
8	D	0.42					Percentile	
23	D	0.33					Which would have a normal distribution?	
22	I	0.31	0.31				Effect of sample size on p-value	
25	G		0.54				Parent distribution of a sample	
33	P		0.47				Temperature on October 1	
18	I		0.46			0.35	Meaning of p-value = 0.10	
7	G		0.42				Which graph is from a different set of data?	P 3
19	I		0.41				Which is true of a t-distribution?	D 1
11	D		0.41				Least impacted by outliers	I 4
10	I		0.40				Bottling company	G 3
30	G		0.35				Variability of a histogram	
21	P	0.30	0.35				Chance of rain	
17	I			0.54			Meaning of 95% confidence interval	
31	P			0.52			Error rate in a manufacting process	
35	I			0.48			Sample size effect on confidence intervals	P 3
38	D			0.46			Correlation engine size	D 1
24	G			0.44			Correlation coefficient	I 2
5	P			0.41			Coin flipped twelve times	G 1
4	P			0.38			Babies born in a hospital	
36	I				0.54		Appropriate test for chemical company	
28	G				0.51		Histogram of class grades	P 0
2	I				0.46		Diet plan	D 1
14	G				0.39		20 samples of 10 points each	I 3
32	I				0.35		Conclusion of p=0.05	G 2
15	D				0.33		Which describes central tendency?	
3	D					0.50	Household income	P 1
37	G					0.44	Effect of outlier on correlation coefficient	D 1
1	P					0.30	Testing a disease	I 0
6	D						Olympic track team	G 1
13	P						Customer service waiting time	P 2
16	P						Deck of cards	D 1
								I 0
								G 0

Table 7c: Promax (κ = 3) rotation (oblique), 5-factor solution

Q#	Area	1	2	3	4	5	Description	Summary
29	D	0.66					Standard deviation equals -2.30	
27	I	0.52					Sampling method for height of college students	
9	D	0.51					Temperatures for a week in August	
34	P	0.49					Rolling dice	P 1
20	I	0.49					Height of college students	D 6
12	D	0.44					Calculating GPA	I 2
26	D	0.39					Standard deviation equals zero	G 0
8	D	0.39					Percentile	
23	D	0.33					Which would have a normal distribution?	
25	G		0.52				Parent distribution of a sample	
18	I		0.50			0.33	Meaning of p-value = 0.10	
19	I		0.44				Which is true of a t-distribution?	
7	G		0.42				Which graph is from a different set of data?	P 2
33	P		0.42				Temperature on October 1	D 1
10	I		0.38				Bottling company	I 4
11	D		0.38				Least impacted by outliers	G 3
21	P		0.32				Chance of rain	
30	G		0.32				Variability of a histogram	
17	I			0.55			Meaning of 95% confidence interval	
31	P			0.53			Error rate in a manufacting process	
35	I			0.46			Sample size effect on confidence intervals	P 3
24	G			0.46			Correlation coefficient	D 1
38	D			0.46			Correlation engine size	I 2
5	P			0.46			Coin flipped twelve times	G 1
4	P			0.36			Babies born in a hospital	
36	I				0.56		Appropriate test for chemical company	
28	G				0.51		Histogram of class grades	P 0
2	I				0.47		Diet plan	D 1
14	G				0.38		20 samples of 10 points each	I 3
32	I				0.34		Conclusion of p=0.05	G 2
15	D				0.31		Which describes central tendency?	
3	D					0.52	Household income	P 1
37	G					0.45	Effect of outlier on correlation coefficient	D 1
1	P					0.30	Testing a disease	I 0
6	D						Olympic track team	G 1
13	P						Customer service waiting time	P 2
16	P						Deck of cards	D 1
22	I						Effect of sample size on p-value	I 1
								G 0

4. Confirmatory Factor Analysis

4.1 Background

The considerations necessary in the exploratory paradigm are superfluous in a confirmatory factor analysis (CFA) because these decisions are made *a priori*. Model comparison lends a certain exploratory flavor to (CFA), although beginning with a theoretical model reduces that feeling of groping about in the dark that EFA instills. With its focus on modeling a proposed structure, CFA is often called structural equation modeling (SEM).

Figure 10 (Loehlin, 2004) displays two similar structural equation models. Part (a) represents is the type commonly seen in a confirmatory analysis, whereas (b) is an exploratory model. Both models attempt to fit three latent variables (circles A-B-C and I-II-III) to six observed variables (squares D to H); error is specified by the lower arrows pointing into the observed variables (e.g., 0.84 for D). The direction of the arrows implies a causal pathway. The exploratory model implies that all three latent variables influence the observed measures. The confirmatory model, in contrast, implies that each latent trait is manifested into three observed variables (e.g., A is observed through D, E, and F). Further, the latent traits B and C are assumed to be correlated (curved arrow), while A is independent. The confirmatory model is thus a more parsimonious representation of the phenomenon: it has nine fewer path coefficients but requires only the one extra correlation to arrive at the same assessment (equal error on the two models).

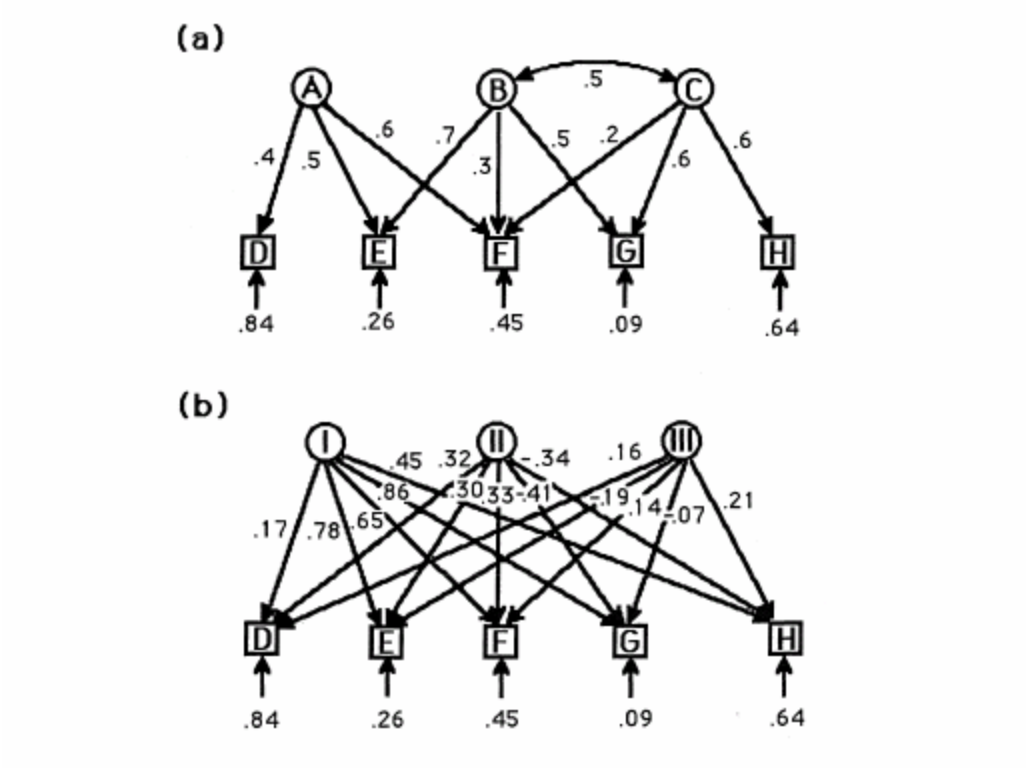


Figure 10: Structural models, (a) confirmatory and (b) exploratory

The numbers along the arrows are *path coefficients*. They are standardized regression weights between the latent and observed variables. The correlations between observed variables are estimated using tracing rules, which essentially amount to summing the multiplicative path coefficients of all possible paths between variables, although the rules can be tricky when more complex models are proposed. For example, the predicted correlation between D and F in (a) is 0.24 (0.4×0.6) through the path D $\rightarrow$ A $\rightarrow$ F. The path E $\rightarrow$ B $\rightarrow$ G is more complex because the correlation r_{BC} allows an extra pathway. The model-implied correlation then is $0.7 \times 0.5 + 0.7 \times 0.5 \times 0.2$ (E $\rightarrow$ B $\rightarrow$ G + E $\rightarrow$ B $\rightarrow$ C $\rightarrow$ G), evaluating to 0.49.

Maximum Likelihood (ML) and Least Squares (LS) can be used to estimate the path coefficients through use of a fit function, defined in equation (7) (Loehlin, 2004).

$$F = (\mathbf{s} - \mathbf{c})' \, \mathbf{W} \, (\mathbf{s} - \mathbf{c}) \quad (7)$$

where: F is the fit function
s is the observed covariance matrix
c is the model-implied covariance matrix
W is a weight function

For multivariate normal data, the above reduces to the following:

$$F = \tfrac{1}{2} \, \mathrm{tr}[(\mathbf{s} - \mathbf{c}) \, \mathbf{V}]^2 \quad (8)$$

where: tr[ ] refers to the trace of a matrix (sum of the diagonals)
V is a weight matrix, where:
$\mathbf{V} = \mathbf{I}$ is ordinary least squares (OLS)
$\mathbf{V} = \mathbf{S}^{-1}$ is generalized least squares (GLS)
$\mathbf{V} = \mathbf{C}^{-1}$ is maximum likelihood (ML)

Model fit is assessed by the following:

$$(N-1)F_{\min} \sim \chi^2 \quad (9)$$

where: N is the sample size
F_{min} is the optimum of the fit function
χ^2 has degrees of freedom $\frac{m(m+1)}{2} - t$, where:
m is the number of observed variables
t is the number of parameters estimated in the model

Being a goodness-of-fit test, the equality in the null hypothesis corresponds to correct estimation (i.e., $\mathbf{s} = \mathbf{c}$). Rejection of the null, therefore, implies poor model fit ($\mathbf{s} \neq \mathbf{c}$). Because the test statistic is a function of sample size, this leads to a contradiction: statistical theory requires large sample size for valid parameter estimation, whereas human nature would prefer a strong fit such that the null is not rejected. In practice, the null is nearly always rejected, in spite of the apparent strong fit of a solution. Therefore, fit indices were developed to provide an alternate means of model assessment.

There appear to be as many (or more) fit indices than there are researchers in the field of structural equation modeling. SAS™, for example, lists around 20. An exhaustive

treatment is beyond the scope of this work. Rather than consider one fit index the quintessence, it is prudent to evaluate the relative values of fit indices from several class for each proposed model (Marsh, *et al.,* 1996). The fit indices chosen for this analysis are discussed below:

- *Goodness of Fit Index (GFI)* – GFI is calculated as the model fit relative to a baseline of no fit (Loehlin, 2004). It is thus quite easy to attain high values, even in excess of 0.99. This index has historical importance but is of little practical value in light of other indices developed since. The Statics Concept Inventory reports a GFI of 0.90 (Steif and Dantzler, 2005).
- *Nonnormed Fit Index (NNFI)* – Along with GFI, NNFI is a member of the class of incremental fit indices because it involves comparison to a baseline model. In a comparison of seven indices, Marsh, *et al.* (1996), preferred NNFI as it "was not systematically related to sample size, appropriately penalized model complexity, appropriately rewarded model parsimony, and systematically reflected differences in model misspecification" (p. 347).
- *Root Mean Square Error of Approximation (RMSEA)* – As a population-based index, RMSEA is "relatively insensitive to sample size" (Loehlin, 2004, p. 68). SAS™ calculates confidence intervals, which is another appealing quality. The Statics Concept Inventory reports a value of 0.067 (Steif and Dantzler, 2005), which is generally acceptable but not outstanding. A value of zero would result from a perfect model-fit.
- *Parsimonious Goodness-of-fit Index (PGFI)* – A parsimonious index adjusts for the degrees of freedom used in a model; a meaningless model could be fit

by assigning one parameter to each estimated value, leaving zero degrees of freedom. PGFI was selected over the adjusted goodness-of-fit index (AGFI) because the latter can take negative values (with low *df*) or can be undefined for a just-identified model (Mulaik, *et al.,* 1989). PGFI is chosen over the parsimonious normed-fit-index (PNFI) along the same lines as NNFI was chosen over NFI.

- *Akaike's Information Criterion (AIC)* – This index also makes a parsimony adjustment, although it is done additively through a penalty term. Various representations exist, and they should be evaluated relative to the model *df*.

4.2 Proposed Models

The significant findings of the exploratory analysis are summarized as follows:

- SCI items are primarily unique; items which *clearly* cover similar topics group along the same factor.
- The hypothesized four topic areas do not appear tenable.
- The uni-dimensional structure cannot be rejected.

A strict one-factor model is shown in Figure 11. The latent construct of Statistics (labeled "G" for general, tip-o'-the-cap to Spearman) is the independent variable influencing item responses. Each item is weighted (w_i) with residual error (e_i).

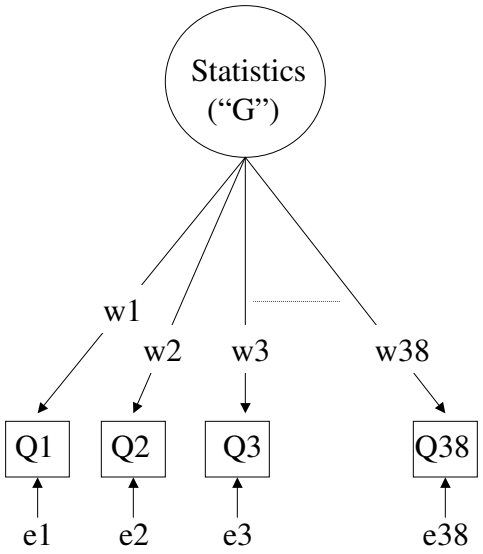


Figure 11: One-factor "G" model for SCI

With the EFA results in mind, a model acknowledging these similarities is desirable. Figure 12 proposes that the errors are correlated for the similar items, denoted by the curved arrow between questions 2 and 36 (choice of appropriate test statistics). Un-grouped items retain the same structure as in Figure 11.

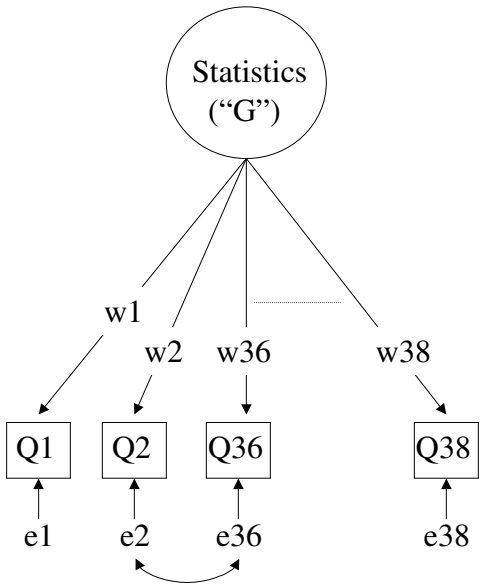


Figure 12: Correlated errors for similar items

Alternately, the similar items can be assessed by an additional construct, as depicted in Figure 13. This is analogous to the four-specific-plus-G model in Book One, although some amount of similarity has been implied by the EFA. Errors have been removed from the diagram but are included in the model.

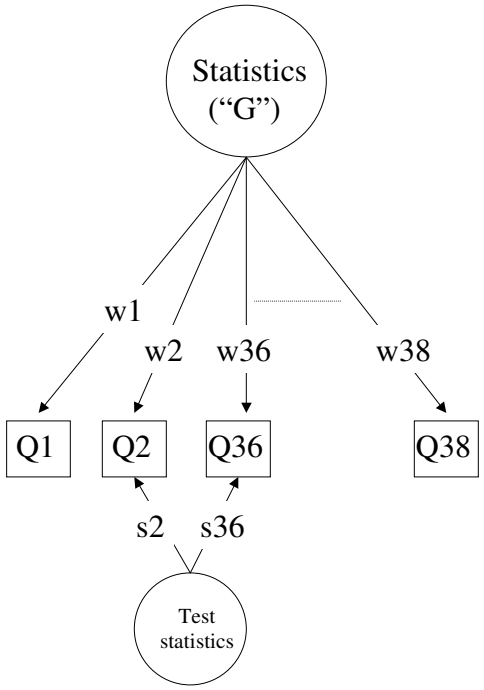


Figure 13: Specific factors modeled as external to "G"

A more complex model, Figure 14, proposes that the similar items form a sub-domain of the parent Statistics construct. This nested model is intuitively appealing because Statistics is acknowledged as the parent construct, again utilizing the guidance of the EFA solution. Each nested construct includes a disturbance term (not pictured). One path from each nested-latent factor to a variable is fixed to one (e.g., *s2* = 1 while *s36* will be estimated) to aid in the scaling of the solution algorithm (Loehlin, 2004).

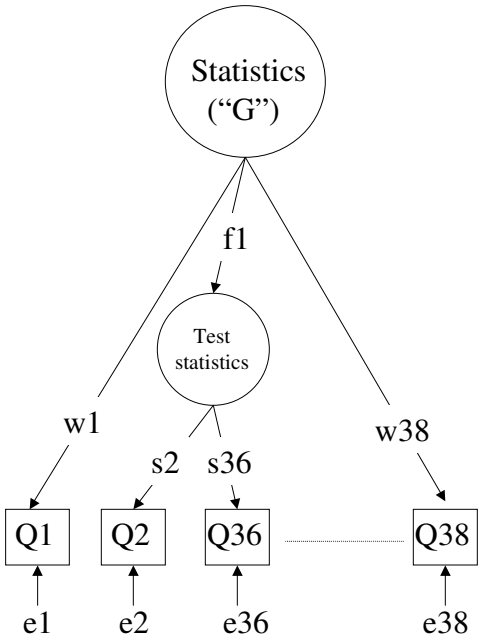


Figure 14: Sub-domain structure, showing relationship between similar items

4.3 Methods

To assess the proposed models, a set of structural equations is written. Each variable with an arrow going in requires an equation. For the one-factor model (Figure 11), these are simply the items. Each item is modeled by the sum of the arrows pointing to it, as below.

$$\begin{aligned} q_1 &= w_1 F + e_1 \\ q_2 &= w_2 F + e_2 \\ &\text{etc.} \end{aligned} \qquad (10)$$

where: w_i are the path coefficients
F is the latent "G" factor
e_i are the residual errors

For the more complex nested model (Figure 14), the latent nested factors require equations as well. The unique items remain in the form depicted above.

$$\begin{aligned} q_1 &= w_1 F + e_1 && \text{(unchanged)} \\ q_2 &= s_2 F2 + e_2 && \text{(nested item)} \\ F2 &= f_2 F + d_2 && \text{(nested factor)} \\ &\text{etc.} \end{aligned} \qquad (11)$$

where: s_i are the path coefficients from items to nested factors
f_i are the path coefficients from "G" to nested factors
F_i are the nested factor
d_i are disturbances for nested factors

The data to be analyzed is the correlation matrix of item scores. This computation was performed in Excel and copied into the data statement of SAS. The SEM analysis was conducted using *proc calis* in SAS.

4.4 Preliminary Results

Table 8 shows the fit function and associated significance tests for six models. Four of these were depicted in section 4.2. Two additional models are included: (4) is the same as (3) except that the specific factors are allowed to correlate with the general factor; (6) is an error-only model, i.e., (1) without the "G" factor. As often occurs in these analyses, all models reject the null hypothesis of strong fit.

Table 8: Fit function and test statistics for six structural models

Model	Fit Function	χ^2	df	$p > \chi^2$
(1) One-factor "G" (Figure 11)	2.6697	784.9	665	0.0009
(2) "G" with correlated errors (Figure 12)	2.6150	768.8	644	0.0005
(3) "G" with specific factors (Figure 13)	2.6177	769.6	652	0.0010
(4) same as (3) with specific correlated to "G" (not pictured)	2.5502	749.7	646	0.0029
(5) "G" with nesting (Figure 14)	2.6209	770.5	659	0.0017
(6) Error only (not pictured)	5.2613	1546.8	703	< 0.0001

Table 9 shows the fit indices for the six models. All models are clearly superior to the error-only model (6), but there is little further distinction.

Table 9: Fit indices for models, keyed to Table 8

Model	GFI	NNFI	PGFI	AIC	RMSEA	*(lower)*	*(upper)*
(1)	0.8805	0.8498	0.8329	-545.1	0.0248	*(0.0166)*	*(0.0315)*
(2)	0.8829	0.8582	0.8251	-545.2	0.0241	*(0.0156)*	*(0.0309)*
(3)	0.8829	0.8497	0.8188	-534.4	0.0248	*(0.0165)*	*(0.0316)*
(4)	0.8849	0.8662	0.8131	-542.3	0.0145	*(0.0145)*	*(0.0304)*
(5)	0.8828	0.8590	0.8275	-547.5	0.0240	*(0.0155)*	*(0.0309)*
(6)	0.6657	0	0.6657	+140.8	0.0639	*(0.0596)*	*(0.0682)*
"Winner"	(4)	(1)	(1)	(5)	(4)	*(4)*	

4.5 Revision

The preliminary results showed little difference between the models. Models (1) and (5) were given further consideration because they represent plausible theoretical models for the instrument and the domain of statistics. Model (4) is rejected because it seems unreasonable that the associated skills should lie *outside* the general domain of statistics, although it is acknowledged as a stronger fit than (5) for this data, based on GFI and RMSEA.

Some inaccuracies were noted in the preliminary solutions. Model (1) can be assessed in an exploratory package. Both SPSS™ and *proc factor* in SAS™ verified the *proc calis* results. Model (5) required slight modification, described below.

- Items 20 and 27 were grouped based on the EFA results. The estimated error variance was greater than one, which outside the bound. The dependency in this case is context (college students' height) rather than concept. This relationship was removed from further consideration, with actually improves the theoretical model.
- The solution was verified by utilizing different optimization techniques in *proc calis*. The weights and communalities were the same to three and most

often four decimals places. The reported results are with the default Dual Quasi-Newton Optimization; the verifying techniques were Double Dogleg, Levenberg-Marquardt, and Trust Region.

The revised fit analysis of (5) is found in Tables 10 and 11, with (1) re-produced for comparison. The fit function and GFI indicate the more complex model is an incremental improvement. However, the complexity comes at the cost of parsimony, which favors the one-factor model (PGFI). In fact, the parsimonious adjustments are positively attenuated by fixing paths in (5), which allows extra degrees of freedom that should be estimated. The results should favor (1) more than is indicated here (e.g., AIC favors (5) but likely erroneously).

Table 10: Fit function and test statistics for preferred structural models

Model	Fit Function	χ^2	df	$p > \chi^2$
(1)	2.6697	784.9	665	0.0009
(5)	2.6386	775.8	660	0.0012

Table 11: Fit indices for preferred structural models

Model	GFI	NNFI	PGFI	AIC	RMSEA	*(lower)*	*(upper)*
(1)	0.8805	0.8498	0.8329	-545.1	0.0248	*(0.0166)*	*(0.0315)*
(5)	0.8826	0.8539	0.8286	-544.2	0.0244	*(0.0161)*	*(0.0312)*

By comparison, Table 12 summarizes the fit of purely exploratory models. The 1-factor solution is model (1). Four factors are required to obtain an un-rejected model by the χ^2. To achieve such a conclusion, the loss of parsimony is great, with the PGFI decreasing by approximately 0.04 for each added factor, while the GFI increases by less than 0.03 across the four factors. The test of no common factors (H_0**:** *no common factors*) is rejected ($\chi^2_{703} = 1476$, $p < 0.0001$).

Table 12: Fit summary for 1- to 4-factor exploratory solutions

Factors	Fit Function	χ^2	*df*	$p > \chi^2$	GFI	PGFI	AIC
1	2.6697	784.9	665	0.0009	0.8805	0.8329	-545.1
2	2.4303	714.5	627	0.0086	0.8906	0.7943	-539.5
3	2.2222	653.3	589	0.0337	0.8995	0.7537	-524.7
4	2.0205	594.0	551	0.0996	0.9084	0.7120	-508.0

This preferred one-factor model is compared to the proposed model of Book One, where each item had a general and specific factor in one of the four content domains. The fit (G + 4, Table 13) is similar to the 3-factor model above, although offering a more parsimonious alternative. By a χ^2-difference test, the G + 4 is a significantly better fit than the one-factor model ($\Delta\chi^2_{48} = 103$, $p < 0.0001$). The one-factor model remains preferred by PGFI. Two expanded G + 4 models are included. These were fit in the vein of a Jöreskog solution, with all items loading on the general factor and other factors defined by the highest-loading item from the Book One analysis. This is similar to a five-factor exploratory solution, with minimal structure pre-imposed. Both uncorrelated (superscript *eu*) and correlated factors (*ec*) yield a better overall fit, but the loss of parsimony is pronounced, more than even the straight four-factor exploratory model.

Table 13: Fit summary comparison for uni-dimensional and G + 4 models

Model	Fit Function	χ^2	*df*	$p > \chi^2$	GFI	PGFI	AIC
(1)	2.6697	784.9	665	0.0009	0.8805	0.8329	-545.1
(G + 4)	2.3194	681.9	617	0.0355	0.8952	0.7857	-552.1
$(G + 4)^{eu}$	1.8325	538.7	525	0.3296	0.9163	0.6843	-511.3
$(G + 4)^{ec}$	1.8268	537.1	515	0.2422	0.9166	0.6715	-492.9

4.6 Conclusions

Table 14 (end of this sub-section) summarizes the communalities and weights (path coefficients) for the preferred models. The columns, as labeled (a) to (g) at the bottom of the chart, are as follows:

(a) Item number. The grouped items are listed first, followed by the unique items. The nested latent factors are listed at the bottom (e.g., N 2-36 is the latent factor for items 2 and 36).

(b) Communalities for model (1). Assessed by the squared multiple correlation, this is the non-unique variance of each item.

(c) Communalities for model (5). The nested latent variables have communalities as well.

(d) Weights for model (1). These are the path coefficients (e.g., *w1* = 0.0931 in Figure 11) in standardized form. These values squared equal the communality (e.g. $0.0931^2 = 0.0087$ for item 1)

(e) Weights for model (5), part 1. These correspond to items which load directly onto the general factor.

(f) Weights for model (5), part 2. These are the loadings from the nested factor to the items (e.g., *s2* = 0.5276 in Figure 14).

(g) Because the nested weights load directly to the general factor, these values squared are the communalities of the nested factor (e.g. 0.4236^2 equals 0.1796 for N 2-36).

(h) Correlation between items and general factor for nested items. By basic tracing rules, these values are the product of the specific weight and the weight from the general factor to the nested factor (e.g., 0.5276 × 0.4238 = 0.2236 for item 2).

The latent nesting factors show high communalities (c) except for N 2 – 36 (items related to test statistics). The latent nesting-to-item weights (f) are higher than the non-

nested-to-item weights (e) (mean 0.39 vs. 0.25). The high path coefficients between the nested and general factors (e) maintain the item-general correlations at nearly the same level as model (1), comparing (d) and (g).

While the one-factor model is a more parsimonious fit for the SCI in its current form, analysis of the nested solution suggests a layered structure exists, although with too few items for a meaningful solution (*cf.* Kim and Mueller, 1978, state that Thurstone recommended at least three items clearly loading on each factor). The structure does not appear as gross as the four-specific-plus-general model proposed in Book One. If one were to cluster all items with highest observed correlations, then presumably a better model could be found; this essentially brings the CFA down to the level of EFA, thus defeating the spirit of defining a plausible structure *a priori*. The possibility of identifying a larger structure is discussed in the forth-coming Extension.

By incorporating both exploratory and confirmatory techniques, the methods of this chapter might be considered a *meeting-in-the-middle* of the two paradigms. A fully exploratory model was utilized, in keeping with the concept inventory literature, as opposed to an exploratory structural model, such as Jöreskog's unrestricted method. The strictly defined model of Book One is not rejected but merely set aside as lacking parsimony for this dataset.

Table 14: Summary of model estimates for models (1) and (5)

Item	Communalities		Weights			
	(1)	(5)	(1)	(5) – G	(5) – N	Corr. G-Item
2	0.0519	0.2784	0.2277	--	0.5276	0.2236
36	0.0200	0.1004	0.1413	--	0.3169	0.1343
17	0.1177	0.1436	0.3431	--	0.3790	0.3386
35	0.1988	0.2442	0.4459	--	0.4941	0.4414
18	0.0069	0.0101	0.0832	--	0.1007	0.0795
22	0.2030	0.3257	0.4506	--	0.5707	0.4503
24	0.0426	0.0877	0.2064	--	0.2962	0.1941
37	0.0071	0.0122	0.0840	--	0.1104	0.0724
38	0.1047	0.2430	0.3236	--	0.4929	0.3230
26	0.2293	0.2319	0.4788	--	0.4816	0.4794
29	0.3106	0.3159	0.5573	--	0.5621	0.5596
1	0.0087	0.0085	0.0931	0.0920		
3	0.0027	0.0025	0.0521	0.0497		
4	0.1567	0.1569	0.3959	0.3961		
5	0.0011	0.0010	0.0326	0.0309		
6	0.0118	0.0123	0.1088	0.1107		
7	0.0624	0.0627	0.2499	0.2504		
8	0.1973	0.1976	0.4442	0.4445		
9	0.1840	0.1856	0.4290	0.4308		
10	0.1016	0.1012	0.3187	0.3181		
11	0.1316	0.1320	0.3627	0.3633		
12	0.2207	0.2213	0.4698	0.4704		
13	0.1008	0.1020	-0.3174	-0.3194		
14	0.0049	0.0049	0.0698	0.0703		
15	0.0800	0.0799	0.2829	0.2826		
16	0.0085	0.0086	0.0922	0.0927		
19	0.0250	0.0251	0.1580	0.1584		
20	0.1190	0.1199	0.3450	0.3463		
21	0.1262	0.1266	0.3553	0.3559		
23	0.1277	0.1285	0.3574	0.3585		
25	0.1367	0.1371	0.3698	0.3703		
27	0.0951	0.0969	0.3084	0.3112		
28	0.0046	0.0044	0.0678	0.0662		
30	0.1360	0.1363	0.3688	0.3692		
31	0.1638	0.1621	0.4047	0.4026		
32	0.0235	0.0229	0.1532	0.1515		
33	0.3008	0.3011	0.5484	0.5487		
34	0.1310	0.1300	0.3620	0.3605		
N 2-36	--	0.1796	--	0.4238		
N 17-35	--	0.7980	--	0.8933		
N 18-22	--	0.6226	--	0.7890		
N 24-37-38	--	0.4295	--	0.6554		
N 26-29	--	0.9910	--	0.9955		
(a)	(b)	(c)	(d)	(e)	(f)	(g)

4.7 Extension

The confirmatory factor analysis can be interpreted in two ways: 1) the Statistics Concept Inventory is best modeled as a uni-dimensional instrument; and/or 2) sub-domains exist within the field of statistics which the SCI partially maps, although there are insufficient topics with multiple items to assess such a broad field as statistics. The first possibility justifies shortening the SCI to a more appealing length, which is discussed in Section 5 (Reliability Revisited), to follow.

The second interpretation is tantalizing as a research proposal, although it is beyond the scope of this work to conduct. Statistics is clearly a broad discipline; approximately 230 universities offer degrees in statistics or related fields (Bureau of Labor Statistics, 2006). The hope to assess the entire domain in one instrument is futile. Figure 15 depicts the substantive relationships identified through the exploratory factor analysis and verified through the subsequent nested structural model. Akin to the original four-factor model presented in Book One, it seems likely that Descriptive (left side) and Inferential (right) are plausible sub-topics, but the dearth of items makes a more-layered structure untenable for the current version of the SCI. This new analysis, which utilized a dataset (Fall 2005) with demographic differences, should be considered a complementary result to Book One, rather than a rejection of the earlier results. These analyses should be conducted on future results, perhaps ultimately concluding in favor of one.

This model can be compared to the field of Physics, which has several related instruments. An analogous model for Physics is shown in Figure 16. The difference between Physics and Statistics is that each sub-domain has an instrument of

approximately 25 items to assess this model, whereas the SCI contains only 2 or 3 items per sub-domain.

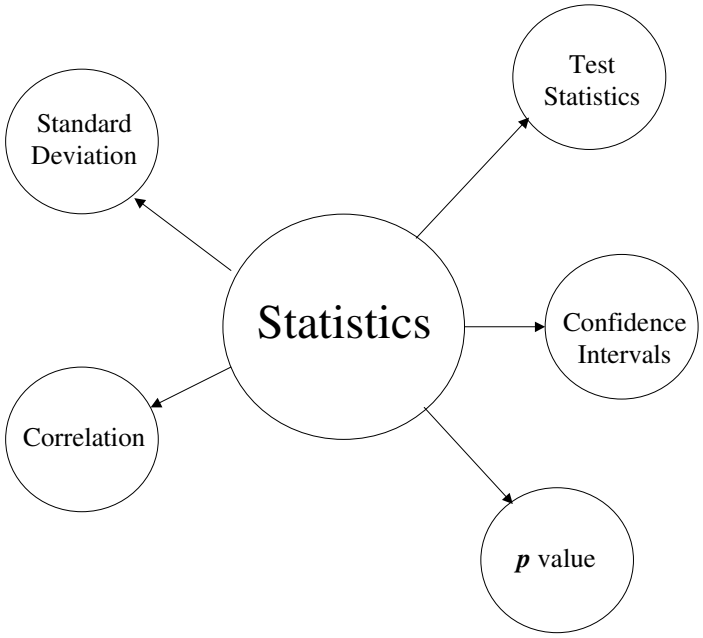


Figure 15: Proposed sub-domains for Statistics discipline

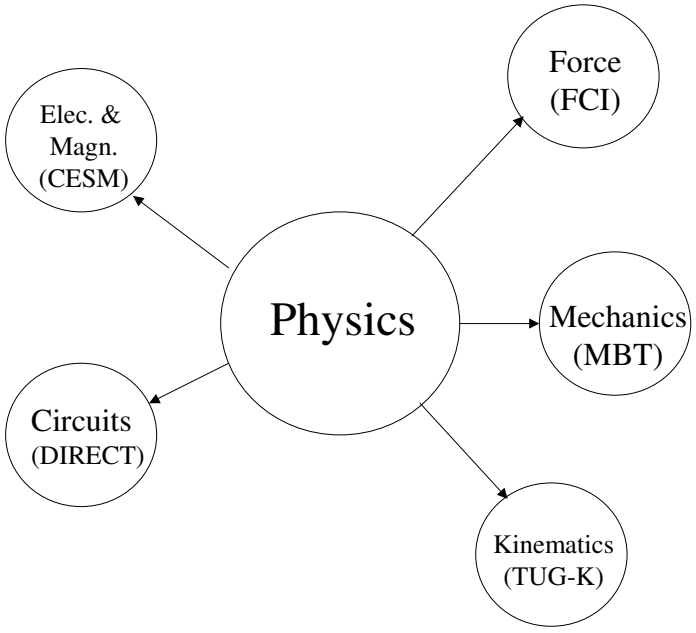


Figure 16: Proposed sub-domains for Physics discipline (with published instruments)

5. Reliability Revisited

If the SCI is assumed to be a uni-dimensional instrument, the question of optimal test length arises. The current 38-item version is longer than other concept inventories, which is a face validity issue: students may not take a low-stakes test seriously if it will require a great deal of effort. To determine an optimal test length, several investigations were conducted.

Using EFA as a guide, six item pairs with highly similar topics were demonstrated to be so analytically. Removing one of each pair reduces the test to 32 items. However, this decreased reliability to 0.7310 from 0.7650 overall. These six pairs (12 items) had generally high discrimination, including two over 0.50 which were dropped. Topically, this strategy appears sound, but the psychometrics of the instrument are hurt.

An alternate strategy focusing directly on reliability was employed next. Using alpha-if-deleted as the criterion, items were removed one-by-one as to provide maximum reliability for any test length. The optimal test length here is 23 items (i.e., 15 deleted), with a reliability of 0.8036. Deleting the 15 lowest-discriminating items was similar (0.8000), with near correspondence between deletion lists. Exclusion of the 15 items with lowest communality from the one-factor structural model is slightly lower (0.7954) but similar. These results are displayed in Figure 17, along with the poor criterion based on the six EFA pairs.

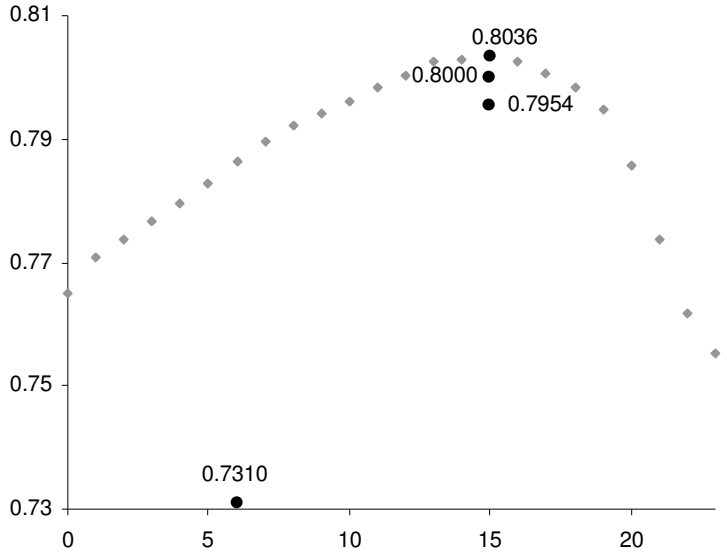


Figure 17: Alpha vs. number of items removed
(highliting maximum alpha, discrimination, communality, and EFA criterion)

5.1 Enquiring Minds Want to Know

A shorter SCI should enhance the face validity of the instrument. Administered in paper form, the packet requires approximately 15 pages, which has elicited groans at times. A nice round number, say 25, would bring the SCI more in line with other concept inventories, whereas the 38 items make it the longest known instrument.

The 15 items deleted to produce Figure 17 were compared across criteria. The lists proved to have 12 items in common. A wise inclusion for the thirteenth item is the waiting time question, which has never had a class score above 25% and often has 0%; this item was assessed a communality slightly below the median but not in the bottom 15, with its loading being the only negative such value. The 13 items chosen for deletion are listed in Table 15, with those remaining in Table 16.

Table 15: Deleted items for 25-item SCI

Topic	Number
Testing a disease	1
Household income	3
Coin flipped twelve times	5
Olympic track team	6
Customer service waiting time	13
20 samples of 10 points each	14
Deck of cards	16
Meaning of p-value = 0.10	18
Which is true of a t-distribution?	19
Histogram of class grades	28
Conclusion of p=0.05	32
Appropriate test for chemical company	36
Effect of outlier on correlation coefficient	37

Table 16: Retained items for 25-item SCI

Topic	Number
Diet plan	2
Babies born in a hospital	4
Which graph is from a different set of data?	7
Percentile	8
Temperatures for a week in August	9
Bottling company	10
Least impacted by outliers	11
Calculating GPA	12
Which describes central tendency?	15
Meaning of 95% confidence interval	17
Height of college students	20
Chance of rain	21
Effect of sample size on p-value	22
Which would have a normal distribution?	23
Correlation coefficient	24
Parent distribution of a sample	25
Standard deviation equals zero	26
Sampling method for height of college students	27
Standard deviation equals -2.30	29
Variability of a histogram	30
Error rate in a manufacting process	31
Temperature on October 1	33
Rolling dice	34
Sample size effect on confidence intervals	35
Correlation engine size	38

The uni-dimensional model was re-analyzed with the retained items. The fit summary is shown in Table 17. The 25-item SCI is a better fit by GFI and PGFI. For publication purposes, this GFI above 0.90 is easier to swallow. The RMSEA, which is a function of *df*, favors the full SCI, but the vastly different *df* make a comparison tenuous.

Table 17: Uni-dimensional model fit summary for 38-original and 25-cut SCI

Model	Fit Function	χ^2	*df*	$p > \chi^2$	GFI	PGFI	RMSEA
(1) – 38	2.6697	784.9	665	0.0009	0.8805	0.8329	0.0248
(1) – 25	1.1529	339.0	275	0.0051	0.9192	0.8426	0.0281

5.2 Cross-validation

A cross-validation was conducted in *MatLab* to assess the accuracy of the alpha-if-deleted criterion. The data divided in half by students ($n \div 2 = 295 \div 2 = 147$ or 148) to serve as training and testing sets. The training set was used to determine the 15 worst items by alpha-if-deleted; the number 15 was chosen to coincide with Figure 17 as to maximize alpha.

The summary statistics for 1000 replications are presented in Table 18. The median alpha (0.7655) is nearly identical that from the 38-item SCI (0.7651). While the estimated 23-item reliability is lower than that from Figure 17 (0.8036), this cross-validation method should yield a less biased estimate of the reliability. It is therefore promising that the reliability maintains its 38-item level.

Table 18: Summary statistics for 1000 replicates of a 15-item-removed SCI

Minimum	*1st Q*	*Median*	*3rd Q*	*Max*	*Mean*
0.6927	0.7517	0.7655	0.7769	0.8133	0.7636

Figure 18 summarizes the item counts for the 1000 replications. The height of each bar is the number of times that each item fell in the bottom 15. The dark bars are those 15 items identified from the full-data exercise. Most importantly, the *exact* list of 15 items was re-produced. The rank-order correlation between alpha-if-deleted (full data)

and bottom 15 count exceeds 0.90. Moreover, the 13 items chosen for deletion are the 13 most-frequent items identified in this cross-validation. The selection of the 13 items for deletion is therefore reinforced.

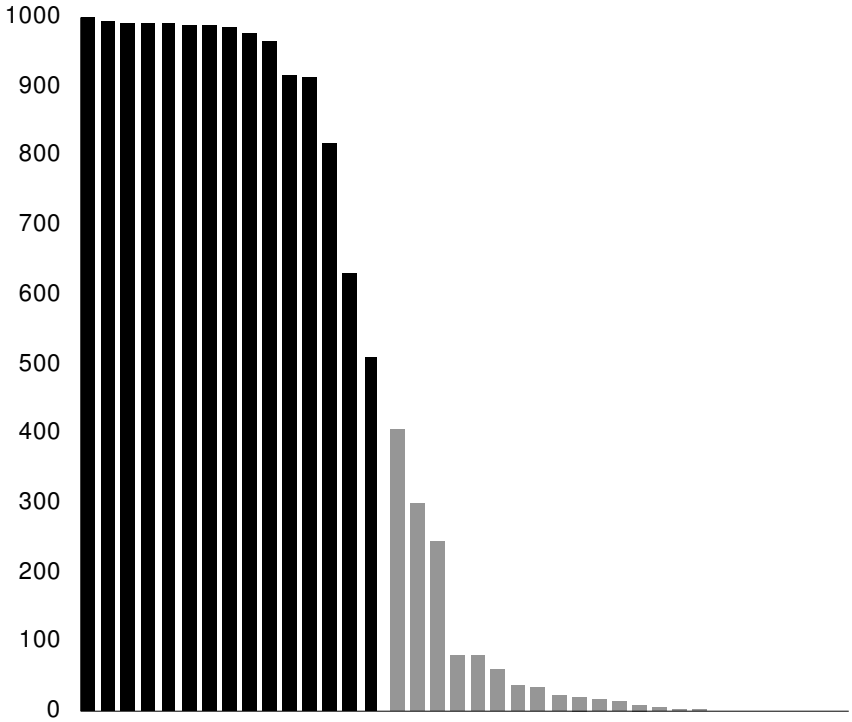


Figure 18: Cross-validation summary, count of items falling in Bottom 15

5.3 Cum Grano Salis

The confirmatory factor analysis used the standard Pearson correlation coefficient. For dichotomous data, these values are attenuated for items differing in difficulty. The tetrachoric correlation is a more appropriate relationship to analyze. Unfortunately, tetrachoric correlation matrices are rarely invertible, which is a requirement for maximum likelihood estimation; the non-invertibility was verified using the tetrachoric correlation matrix from SAS *proc freq* and by a separate program from Enzmann (2001). To correct for this, the tetrachoric matrix was ridged by adding 0.25 to the diagonal elements. The fit statistics and parameter estimates changed, usually by

small amounts. Most importantly, the conclusions from the model were no different: the 13 items chosen for deletion fell in the bottom 15 of communalities. Item 13 (waiting time) seems to be very difficult to assess. When its communality is ignored, the rank-order correlations between communalities (Pearson and tetrachoric data), discrimination index, and alpha-if-deleted all exceed 0.87.

The use of the Pearson correlation coefficient may be considered acceptable if the under-lying correlation between variables is moderate, say, less than 0.70 (Kim and Mueller, 1978; Kim, *et al.,* 1977). This would seem likely for the SCI except for the very few items which are clearly similar. The item clustering will not be affected because the relative magnitude of the correlation coefficients is equivalent across coefficient types.

The sample size of 295 is large by SCI standards but not for this type of analysis. In a meta-analysis, Hoogland and Boomsma (1998) suggest a sample size smaller than five times the model degrees of freedom leads to over-rejection of the model χ^2. A sample of over 3000 would be required to eliminate this erroneous rejection, perhaps implying the overall model fit (i.e. rejection of the null hypothesis) may not be as severe as indicated. Kaplan (2000) comments that non-normality does not affect parameter estimates, which is crucial to the inclusion of communalities as an evaluation criterion.

6. Conclusion

This chapter began by analyzing the reliability of the Statistics Concept Inventory from a multi-dimensional perspective. It was demonstrated that some higher level of structure beyond a uni-dimensional model could account for error variance, thus increasing the reliability estimate. An exploratory factor analysis was conducted to determine a substantively meaningful structure to account for this multi-dimensionality. Six pairs of items with strong similarities were identified. A confirmatory factor analysis sought to verify a theoretical structure. The preferred model for the instrument turned out to be uni-dimensional. Using EFA as a guide, five sets of similar items were verified to have a relatively parsimonious structure, in a nested relationship within the field of statistics.

Promise thus exists for determining a structural model for such a broad field, if more high-quality items could be written; an analogy was drawn to the existing instruments in the field of Physics. The four-factor model of Book One should be re-examined as a plausible alternative when more data is available, although it lacks parsimony for this dataset. Finally, the reliability was re-investigated to arrive at a 25-item SCI as to maximize Cronbach's alpha at a value of around 0.76, at the same level as the full 38-item instrument.

References

Anderson, D.L., K.M. Fisher, and G.J. Norman. 2002. Development and Evaluation of the Conceptual Inventory of Natural Selection. *Journal of Research in Science Teaching*. 39 (10): 952-978.

Armor, D. 1974. "Theta reliability and factor scaling." In Sociological Methodology 1970, H. Costner, Ed., pp.17-50. Jossey-Bass: San Francisco.

Bureau of Labor Statistics. 2006. "Occupational Outlook Handbook, 2006-07 Edition." [http://www.bls.gov/oco/], Accessed March 30, 2006.

Enzmann, D. 2001. "TetCorr 2.1", Accessed March 31, 2006. [http://www2.jura.uni-hamburg.de/instkrim/kriminologie/Mitarbeiter/Enzmann/Software/Enzmann_Software.html]

Gravetter, F.J., and L.B. Wallnau. 1988. Statistics for the Behavioral Sciences. 2nd ed. West Publishing Company: St. Paul.

Green, V., and E. Carmines. 1979. "Assessing the Reliability of Linear Composites." In Sociological Methodology 1980, K.F. Schuessler, Ed., pp.160-75. Jossey-Bass: San Francisco.

Harman, H.H. 1976. Modern Factor Analysis. 3rd ed. The University of Chicago Press: Chicago.

Hoogland, J.J., and A. Boomsma. 1998. "Robustness Studies in Covariance Structural Modeling." *Sociological Methods & Research*. 26 (3 / February): 329-367.

Johnson, R.A. 1994. Miller & Freund's Probability & Statistics For Engineers. 5th ed. Prentice-Hall, Inc.: Englewood Cliffs, NJ.

Johnson, R.A., and D.W. Wichern. 2002. Applied Multivariate Statistical Analysis. 5th ed. Prentice-Hall: Upper Saddle River, NJ.

Kaplan, D. 2000. Structural Equation Modeling. Sage Publications: Thousand Oaks, CA.

Loehlin, J.C. 2004. Latent Variable Models. 4th ed. Lawrence Erlbaum Associates: Mahwah, NJ.

Kim, J.O., N. Nie, and S. Verba. 1977. "A note on factor analyzing dichotomous variables: the case of political participation." *Political Methodology*. 4: 39-62.

Kim, J.O., and C.W. Mueller. 1978. Factor Analysis: statistical methods and practical issues. Sage Publications: Beverly Hills.

Marsh, H.W., J.R. Balla, and K.T. Hau. 1996. "An Evaluation of Incremental Fit Indices: A Clarification of Mathematical and Empirical Properties." In Advanced Structural Equation Modeling, G.A. Marcoulides and R.E. Schumacker, Eds., pp. 315-353.

Mendenhall, W., and T. Sincich. 1995. Statistics for Engineering and the Sciences. 4th ed. Prentice-Hall, Inc.: Englewood Cliffs, NJ.

Mulaik, S.A., L.R. James, J. Van Alstine, N. Bennett, S. Lind, and C.D. Stilwell. 1989. "Evaluation of goodness-of-fit indices for structural equation models." *Psychological Bulletin*. 105: 430-445.

Rummel, R.J. 1970. Applied Factor Analysis. Northwestern University Press: Evanston.

Steif, P.S., and J.A. Dantzler. 2005. A Statics Concept Inventory: Development and Psychometric Analysis. *Journal of Engineering Education*. 33: 363-371.

Zeller, R.A., and E.G. Carmines. 1980. Measurement in the social sciences. Cambridge University Press: London.

CHAPTER X

Content Validity of the Statistics Concept Inventory

1. Introduction

Using objective criteria, Chapter IX proposed a 25-item version of the SCI to optimize the reliability of a shorter instrument. This chapter focuses on the improvement of the retained items, along with suggestions for edits to the 13 deleted items that could allow for their retention in a larger item pool. Incorporating a faculty topics survey, the coverage of the SCI is discussed along with suggestions for enhancing coverage of the aforementioned larger item pool. The potential misconceptions identified in Chapter VIII are analyzed using student interview responses, when possible.

The prior 38-item SCI can be found in Stone (2006), along with item summary statistics up to Summer 2005. The development of these items, again with item statistics, was presented in Chapter IV of this dissertation, up to Spring 2004. The edited items are at the end of this chapter as Appendix 1 (25 retained) and Appendix 2 (13 deleted).

2. Methods

2.1 Student Interviews

Patton (1990) describes three approaches to interviewing. The *informal conversational interview* follows the natural flow of the conversation with the interviewee and may even be conducted in a spontaneous setting. The *general interview guide approach* has a prescribed set of questions which serves as a guideline but need not be followed explicitly, so long as the relevant information is gathered from each subject. The most rigid approach, the *standardized open-ended interview*, is essentially a scripted

interview and varies little from one respondent to another. This approach is most useful when a large number of respondents and interviewers are used, in order to reduce variability in the data gathered.

Regardless of the approach, Patton advises against the use of questions which can elicit a dichotomous response, instead phrasing questions as to be open-ended. An example of poor phrasing is "Did you find that being in the program affected what happened?", while a more effective open-ended structure would be "How do you think your participation in the program affected what happened?" (p. 299). The use of presuppositions also aids in promoting detailed responses. For example, a poorly-worded question in a dichotomous format is "Have you learned anything from this program?", whereas learning can be presupposed by asking "What have you learned from this program?" (p. 304).

Student interviews for the SCI are best conducted in a *general interview guide approach*. For each item, the following will be sought: basic understanding of what the question is asking; the correct answer; consideration given to alternate answers; thought processes. Following Patton's advice, some guidelines for the interview questions are the following:

- "What do you believe this question is asking for?"
- "Which answer to you believe to be correct?" and "Why?"
- "How did you arrive at choice W as opposed to X, Y, and Z?"

Focus groups vs. Individual interviews

Focus groups were originally formulated in the 1950s for use in marketing research. They are appropriate to assess group decision-making. The efficiency is

increased as well by gathering multiple opinions and perspectives simultaneously. The appropriateness of focus groups is less clear in the present setting. The increased efficiency is preferred, but there is no group dynamic to decision-making on the SCI. A balance between the group efficiency and the individual thought processes must be struck. Individual interviews will be the preferred method, but participants can also not simply be turned away.

2.2 Faculty Survey

Results from an earlier faculty survey were consulted as a guideline for topic coverage in constructing the SCI (Chapter IV). This study is essentially a re-validation of that topic list by expanding the audience beyond the College of Engineering at the University of Oklahoma.

The data was gathered using the open-source, freeware PHPSurveyor (Ver. 0.98 stable). [http://www.phpsurveyor.org/]. This package utilizes a series of PHP scripts which interact with a mySQL database to allow creation, management, collection, and analysis of web-based surveys. The same server was used as described in Chapter VII.

The topic list was primarily the same as that used in the earlier survey, but a few additional topics were added based on items the research team felt were not explicit enough. A total of 87 topics were included ranging across 12 broader topic areas; each area contained between 4 and 13 topics. The list can be found in Tables 4 which includes the results of the survey.

To elicit participation, an email was sent to an Industrial Engineering *listserv* and parties who had expressed interest in the SCI. The message contained a link to the survey and simple instructions. Upon visiting the link, subjects first responded to an informed

consent form (yes/no). Those who agreed to participate were first directed to a short demographics survey which requested current position, type of institution, statistics teaching experience, experience with statistics as a student, highest degree, and field of highest degree

Topics were listed in groups by area one screen at a time (e.g., Figure 1). Respondents were asked to rate the topic for its importance to their curricular needs in a 4-point scale: “Not at all important”; “Somewhat important”; “Important”; “Very important”; “N/A” was the default choice and offered in the even that topics were unfamiliar. Radio buttons were used to make selection, as these are a commonly encountered format in web surveys and should be familiar to participants. A screen shot of one area is shown below.

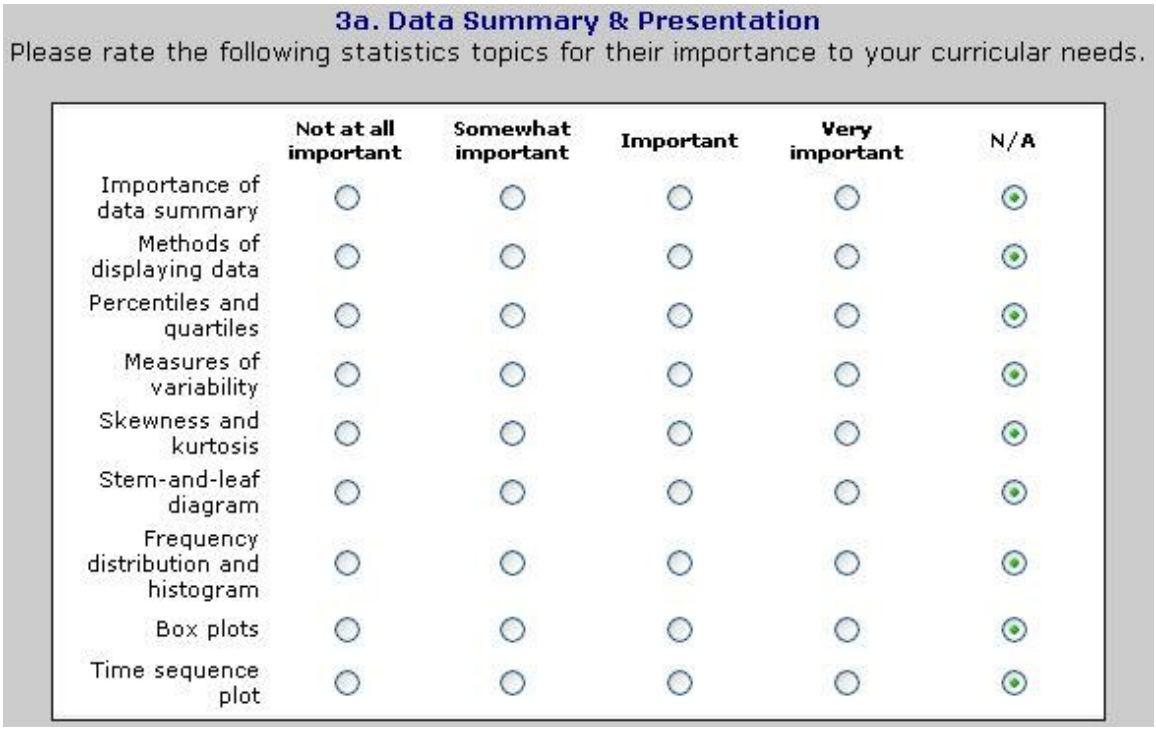


Figure 1: Sample of online survey format

3. Results

Table 1 (next page) summarizes the objective metrics used to arrive at a 25-item SCI. Sections 3.1 and 3.2 contain comments on the retained and cut items, respectively. Interviews were conducted with four Industrial Engineering graduate students who have taken multiple statistics courses. Section 3.3 assesses the topic coverage of the SCI with the faculty survey results.

Table 1: Item Summary Statistics for full 38-item SCI

		Corr.	Values			Ranks		
Old #	New #	(%)	Comm.	Disc.	Alpha	C.	D.	A.
1	cut	49%	0.0087	0.17	-0.0024	31	34	32
2	1	52%	0.0519	0.38	+0.0047	25	21	22
3	cut	22%	0.0027	0.14	-0.0024	37	35	32
4	2	40%	0.1567	0.41	+0.0086	10	19	11
5	cut	10%	0.0011	0.07	-0.0012	38	37	31
6	cut	50%	0.0118	0.21	-0.0027	30	30	36
7	3	29%	0.0624	0.36	+0.0040	24	24	23
8	4	76%	0.1973	0.47	+0.0102	7	13	7
9	5	77%	0.1840	0.46	+0.0090	8	15	10
10	6	44%	0.1016	0.38	+0.0075	20	22	17
11	7	65%	0.1316	0.53	+0.0093	13	9	9
12	8	72%	0.2207	0.54	+0.0115	4	8	4
13	cut	3%	0.1008	-0.13	-0.0057	21	38	38
14	cut	13%	0.0049	0.08	-0.0008	35	36	29
15	9	66%	0.0800	0.37	+0.0052	23	23	21
16	cut	48%	0.0085	0.18	-0.0027	32	32	36
17	10	44%	0.1177	0.46	+0.0071	18	14	18
18	cut	29%	0.0069	0.21	-0.0011	34	31	30
19	cut	33%	0.0250	0.23	+0.0000	27	28	28
20	11	42%	0.1190	0.51	+0.0068	17	11	19
21	12	63%	0.1262	0.57	+0.0086	16	5	11
22	13	50%	0.2030	0.55	+0.0111	5	7	6
23	14	51%	0.1277	0.53	+0.0078	15	10	15
24	15	53%	0.0426	0.39	+0.0025	26	20	25
25	16	49%	0.1367	0.48	+0.0086	11	12	11
26	17	69%	0.2293	0.63	+0.0119	3	2	3
27	18	87%	0.0951	0.23	+0.0036	22	29	24
28	cut	62%	0.0046	0.26	-0.0026	36	27	35
29	19	71%	0.3106	0.63	+0.0136	1	3	2
30	20	35%	0.1360	0.45	+0.0086	12	16	11
31	21	47%	0.1638	0.58	+0.0094	9	4	8
32	cut	58%	0.0235	0.29	+0.0001	28	26	27
33	22	35%	0.3008	0.68	+0.0161	2	1	1
34	23	68%	0.1310	0.44	+0.0078	14	18	15
35	24	43%	0.1988	0.57	+0.0115	6	6	4
36	cut	48%	0.0200	0.32	+0.0009	29	25	26
37	cut	54%	0.0071	0.17	-0.0025	33	33	34
38	25	62%	0.1047	0.44	+0.0068	19	17	19

Key:

Old #	keyed to full SCI
New #	keyed to retained items
Corr.	percent correct
Comm. / C.	communality from uni-dimensional structural model
Disc. / D.	discriminatory index
Alpha / A.	<overall alpha> minus <alpha-if-deleted> (+ is good item)
Values	Numerical values of these columns
Ranks	Rank order of these columns (high number = bad item)

3.1 Retained Items

These items were retained on the basis of strong discrimination, a positive influence on reliability, and high communalities. Therefore, any major changes are discouraged, with the editing focused on enhancing clarity and other minor clean-ups. The items marked "(Misconception)" are those falling in the over-confidence region from Chapter VIII (Table 6).

New #1 (Old #2)

Students who are familiar with dependence (e.g., "before and after") reported no difficulty (e.g., it was apparent "right away"). One subject distinguished between *t* (small sample) and *Z* (large sample), but unfamiliarity with dependence led to guessing between B and C. These comments indicate a well-written item.

New #2 (Old #4) (Misconception)

Interviewees focus on the location without considering sample size. One subject felt he would interpret the question differently if numbers were given (e.g., 10 total babies in rural, 100 in urban). Such a change, however, inhibits assessment of transference. Another subject correctly keyed on size rather than location.

New #3 (Old #7)

This item requires a great deal of thought. The optimal strategy seems to be noticing the empty "12" stem and the corresponding flat 12-to-13 portion of the cumulative frequency; these two can therefore be considered equivalent. The histogram and stem-and-leaf seem to provide the easiest comparison (e.g. histogram "4" has 10 whereas stem "4" has only one). Thus, the histogram is different from the other three.

To aid students in arriving at this conclusion, the boxplot option is deleted. This option was least-chosen for Fall 2004, Spring 2004, and Summer 2005 (Stone, 2006). One student felt the scale of the cumulative frequency was too fine, suggesting that labeling the bars would help. This is a reasonable suggestion, but it was not carried out for this document.

New #4 (Old #8)

This item is a basic definition recall. One incorrect student commented that he did not remember the definition, was in turn tricked by answer B, thus arriving at D. A non-definitional percentile question is advisable for an expanded item pool. Prior discussions amongst the research group proposed interpreting percentile differences in the tails versus the center of a normal distribution.

New #5 (Old #9)

This is among the easier items; no useful comments were offered during interviews.

New #6 (Old #10) (Misconception)

Students feel comfortable with the textual hypothesis definition as opposed to the traditional symbolic format. One student erroneously chose B as a more conservative statistical test than the correct D, which is sound statistical reasoning for someone grounded in conservativism; this option could be deleted if other high-level students possess similar reasoning. Another student did not view the scenario statistically but rather "what is good for the company" (i.e., under-filling the bottles saves money).

New #7 (Old #11)

This is among the easier items, and no useful comments were offered during interviews.

New #8 (Old #12)

Again, this item is relatively simple for most students. One subject initially considered B because the numerator is correct, although he settled correctly on C. This led to a lower confidence rating of 2.

New #9 (Old #15)

Subjects recognize that the mean is inaccurate because the distribution is "too spread out," with 36 and 83 identified as potential "outliers." It was observed that the median and mode have the same numeric value (10), which could cause confusion. Therefore, one additional "5" was added to the list in place of a "10."

New #10 (Old #17) (Misconception)

In general, subjects are evenly split along options B and C with minimal consideration of A and D, either of which could be deleted if desired. One correct respondent recalled the graphical depiction of confidence intervals typical of introductory statistics textbooks. Another re-considered his incorrect B during discussion upon recalling that the population mean "does not move around."

New #11 (Old #20) (Misconception)

Students tend to reason correctly on this item ("think about groups first"; "use standard deviation"). For example, one student evaluated each option relative to its distance from one-sigma. However, he ignored the sample size information, providing evidence of this item's diagnosticity for failure to consider sample size. Wholly correct reasoning will recognize options A and B as individual "outliers" compared to such a large sample as 100.

New #12 (Old #21)

This item effectively evaluates conceptual understanding of probability. One incorrect student took a deterministic view, considering 40% to be a prediction of no rain ("less than 50"), thus choosing B. Another student correctly held that decisions cannot be made individually, with the probabilities acting "like a coin."

New #13 (Old #22)

Interview subjects failed to see a connection between sample size and *p*-value, despite recalling the Central Limit Theorem or commenting "sample size more: test is better." One student focused on *p*-value as a decision criterion relative to alpha, not viewing *p* as an indicator of the strength of the evidence against a null hypothesis. This item merits more discussions and is a good candidate for a complementary item. The strong objective metrics trump potential edits.

New #14 (Old #23)

One student commented that a normal distribution should extend from $-\infty$ to $+\infty$. While technically correct, the tails become so small at even $Z = \pm 3$ that this thinking might be considered a *practical misconception*.

New #15 (Old #24)

Interview comments may warrant an additional item along similar lines, although the present version is unchanged due to the thoughtful consideration of all possibilities. For example, correlation is not distinguished from slope ("gradient"); a re-scaling of one axis could help evaluate this mistake by providing a wider range of values (e.g., 16 and 9 rather than 1.6 and 0.9). A slightly stronger correlation is also recommended, as the

addition of even one point near the lower right could yield a zero correlation. One student did not recall correlation, suggesting that it should be covered in the curriculum.

New #16 (Old #25)

The general strategy is to work by process of elimination, until only the correct B (uniform) remains. One student ascertained the correct answer immediately, however, because the distribution has "no tail."

New #17 (Old #26)

All interviewees felt this item was easy and answered correctly. One commented that the set of numbers had "no spread."

New #18 (Old #27)

This is the easiest item on the SCI and barely made the cut from Table 1. Interview subjects commented that B is correct as the "more [*sic*: most] random sample," while A has "bias" and C is "not representative." One student over-looked the *little word* "not," a problem cited in Force Concept Inventory research and earlier SCI interviews.

New #19 (Old #29)

The summary statistics rate this as one of the top items. Interviews yielded no useful comments.

New #20 (Old #30) (Misconception)

This item is a strong indicator of students' understanding of histograms. One student with correct reasoning focused on A and D as potential responses, favoring A due to its larger range. Incorrect students either favor C for its normality ("most comfortable with normal") or B because the bin counts show the most variability (i.e., failing to interpret a histogram as a summary of a set of numbers).

New #21 (Old #31)

A conceptual understanding of independence seems to be accurately assessed. Incorrect students chose A either on a gambler's fallacy or sampling without replacement, while a coin flip was a justification for C. A correct respondent keyed on the terms "groups" and "not independent" as guides.

New #22 (Old #33) (Misconception)

This item is perplexing. Objectively, it is the best item on the SCI. It also rates as the third most-difficult item among those retained, suggesting its importance at reaching high-level students. Subjectively, however, this item is poor in that it relies more on definitional recall than statistical reasoning. This is apparent by inspection and verified in interviews. An expanded item pool should contain a complementary item that ideally measures this concept less recollectively.

New #23 (Old #34)

Editing is called for here. Interviewees misinterpreted the spirit of the "average of 1.5" statement. One subject considered sample space (e.g. {1, 1.5, 2, …, 5.5, 6} for 2; {1, 1.1, 1.2, …, 5.9, 6} for 10). The intent of the item is actually "less than or equal to 1.5," but such complex possibilities made this respondent "not want to think about it." Another subject considered the sum rather than the mean, which is not possible for any rolls.

The new version is more general, referring to "which scenario" rather than "which roll." The direction was changed to help with wording, to "at least 4.5," while the term "mean" was substituted for "average" for consistency throughout the stem.

New #24 (Old #35)

Interviews bespeak of this item's ability to probe sample-size reasoning skills. Two subjects did not see the relation between sample size and confidence intervals, choosing C due to "same standard deviation."

New #25 (Old #38)

Interviewees possess correct conceptualization. One utilized a visual heuristic ("graph in my head"), while others related negative to opposite or "reverse."

3.2 Deleted Items

Compared to the previous section, it is less important for these items to maintain a similar form. Some items clearly do not work (e.g., #13) and need to be deleted. Some items are wise to retain for a future expanded item pool, especially those previously labeled as misconceptions (e.g., #14).

Old #1

Interview subjects had difficulty reconciling the text with the chart. A suggested edit is to devise two items of differing context: one textual, one chart-based. The item may prove worth retaining in one form but not the other.

Old #3 (Misconception)

Interviews revealed carelessness as a cause of high confidence for the incorrect B. One subject missed the word "random" in C, while another admittedly was "not careful" to distinguish between B and C. In the Fall 2005 data, the correct answer had the lowest confidence, and the item was quite difficult (23%). These results suggest unfamiliarity with stratified sampling; the item could be retained as part of an advanced item pool. One correct student commented that the stratified sample will yield "more variability."

Old #5 (Misconception)

The notion that a coin is a 50-50 proposition seems ingrained in students. Interview subjects used the terms "always" and "fixated" to describe their views of coins. One nearly arrived at the correct answer by considering the question of control ("General [*sic*: Western] Electric rules") but then erred in the vain of a gambler's fallacy (reversal of prior happenings) rather than expecting a continued out-of-control process. Another felt the "probability has to be different" because 12 is "very big" but still settled on D. A different context is recommended for this item.

Old #6

From interviews, students appear to respond thoughtfully to this item, perhaps too much so. The intent of this item is for students to consider the numerical values of the running times listed (e.g. A could be 25.2, 25.7, 25.9, …; C could be 11.3, 26.0, 53.2 for one set of the events). A further distinction lies in assuming that an individual's times will show less variability than those for the team as a whole. Students commented that sampling is a consideration in the sense that the smallest variability will occur in the largest sample (*cf.* central limit theorem). The times for multiple events (C and D) were interpreted as sums rather than individuals measurements of varying range.

Old #13 (Misconception)

Despite universally low scores, some students reason statistically, by considering what type of distribution is relevant. However, the focus tends to settle on normality, with expectation of variance and perhaps sample size given. The absence of these values leads to answer A as the simplest consideration. One student identified the "exponential"

distribution but did not arrive at the correct answer. The memoryless property is at times recalled when queried, but it is not recognized as relevant to the scenario.

Old #14 (Misconception)

Students have difficulty reconciling the parent χ^2 distribution with the sample histograms. Even students who possess an understanding of the central limit theorem can be diverted by misestimating the population mean, leading to "Graph 2" as the most normal. One student correctly identified the population mean when queried but did not appear to make this consideration during the testing. The population mean has been added parenthetically to the stem as an aide.

Old #16 (Misconception)

One interviewee initially erred to A as the most "different" but reconsidered in favor of D during discussion. The "least likely" phrasing did not pose problems for the subjects. It is unclear why this item fell into the over-confident region in the large dataset; more research may help.

Old #18

There was an oversight in conversion to the online test: option A originally contained an alpha level in symbolic form but was omitted from the online test. This has been corrected. A possibly related source of confusion lies in that the stem states the null was rejected, implying an alpha of at least 0.10. Given that the hypothesis is rejected, the re-evaluation of the decision in option A may be an unfamiliar thought process. If the edited item is retained, reasoning about option A should be evaluated.

Old #19

Confusion arises in interpreting the "less area" portion of option C. One student inferred that for a given area, the *t* statistic will fall further from the mean than the corresponding *Z*. This option has been edited in an incorrect form to respect this reasoning which is more in line with how *t* and *Z* are used in statistical tests.

Old #28

This item is perhaps too tricky, requiring a keen eye for the scale rather than correct statistical reasoning. One interviewee noticed the scale differences but "got lazy" to find the correct response, while another considered the distributions equal because the average bin count is equal due to equal sample sizes.

Old #32

Interviews did not yield useful comments.

Old #36

This item evoked thoughtful responses. One subject felt that the use of the same reactor introduced dependency but realized during discussion that this is not the same level of dependency inherent in a paired comparison. Correct reasoning was found in one subject who ruled out B as "not before and after" in favor of C as "two different means."

Old #37

The first sentence of the stem could imply that correlation coefficients are to be listed with the graphs. To avoid this potential confusion, estimated values are listed in an edited version of the item.

3.3 Topic Coverage

Topic coverage is assessed via the faculty survey. Twenty-four subjects participated (n = 24). The demographics are summarized in Table 2. The "typical" participant is a professor at a doctoral university who has teaching experience with introductory statistics and probably at least one related course. Coursework as a student was similar while also employing statistics in research, as he pursued a Ph.D. in Industrial Engineering.

Table 2: Demographics summary, with $n = 24$ (multiple responses allowed)

Categories	*Count*
Current Position	
Professor	21
Adjunct faculty	3
Graduate student	1
Work in industry	0
High school teacher	0
Other	1
Type of Institution	
Doctoral/research university	18
Master's college/university	5
Baccalaureate college	2
Two-year/Associate's college	1
High school	0
Other	0
Statistics teaching experience	
Introductory statistics	20
Advanced statistics	14
Course with Stats as a pre-req	14
Course which is a pre-req for Stats	2
Other course that uses statistics	7
Other	3
Experience with Statistics as a student	
Introductory statistics course	17
At least one advanced statistics course	15
Used statistics in research (e.g. thesis/dissertation)	17
Other	2
What is your highest degree?	
Ph.D. or other doctorate	21
Master's	2
Bachelor's	0
Other	0
What is the field of your highest degree?	
Engineering	16
Mathematics	1
Social Sciences	1
Health/Medicine	1
Education	2
Physical Sciences	2
Other	1
Degree specifics	
Industrial Engineering / Operations Research	12
Other Engineering / Mathematics	4
Psychology / Education / Health	3
Statistics / Applied Statistics	2

The results of the new survey ("New") are compared to the prior study ("Old") conducted at OU, which was discussed in Book One. Table 3 offers the summary statistics for the two studies. New has higher values at every point, suggesting it is unwise to strictly combine the results. The surveys will be considered in tandem, using rank-orders to avoid re-scaling.

Table 3: Summary statistics for New and Old surveys

	Minimum	*1st Q*	*Median*	*3rd Q*	*Max*	*Mean*	*St. Dev.*
New	1.95	2.59	2.95	3.33	3.86	2.94	0.51
Old	1.88	2.28	2.61	2.86	3.75	2.63	0.42

Figure 2 displays the relationship between New and Old mean topic rankings. The dashed line represents equal ranks, as a visual guide. The correlation is moderately strong ($r = 0.69$ for numbers; $r = 0.67$ for ranks). Over half of the topics (48 of 87) fall within 10 of equality.

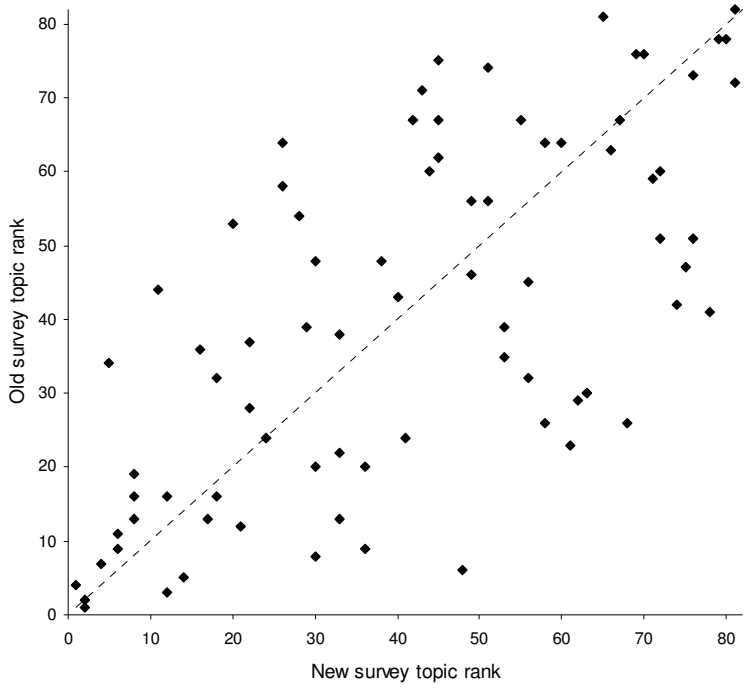


Figure 2: Comparison of New and Old topic rankings

Table 4, spread across the three subsequent pages, summarizes the results of the topics surveys cross-checked with the SCI. The table's components are the following, from left to right:

- *Topics*. The general areas are in larger font and italicized. The specific topics are listed underneath, up to the next general area.
- *Rank New*. For areas, this is the ranking of the mean for all topics in that area (12 total). For topics, this is the ranking of that topic's mean out of the whole 87.
- *Rank Old*. Equivalent to the Rank New for the earlier study.
- *Retained*. For areas, this is the count (in parenthesis) of unique items in that area. For topics, this is the item number based on the 38-item SCI (to avoid confusion of numbers being listed in the Cut column as well). Many items are listed in multiple topics.
- *Cut*. These are the 13 items removed from the SCI, again with numbering maintained to the original 38 for clarity.

For example, the topic "Importance of data summary" was ranked 2nd highest on both New and Old. There are four retained items and zero cut in this topic. The parent area "Data Summary and Presentation" was ranked 3rd highest out of 12 areas on both New and Old, with eleven items retained and four cut.

The most important topic in most areas has at least one item, typically multiple items. One exception is *Linear Regression*, which may be encountered in other courses but not covered in introductory statistics. The two Design of Experiments and *Time Series* are the only areas with zero topics covered on the SCI, but these are rated as

unimportant, perhaps because they are specific to upper-division Industrial Engineering rather than a broad engineering base.

The coverage of the twelve broad topic areas is described as follows:

- *Data Summary and Presentation*. There is strong coverage within this area, which contains most of the basic concepts of descriptive statistics and graphical displays.
- *Probability*. The Retained and Cut portions contain an equal number of items, suggesting this area is difficult to assess. A more liberal classification could place more items in "Interpretation of probability" (rank 4 new / rank 7 old) and "Independence" (17 / 13), which are the two most important topics.
- *Random Variables*. The most important topic, "Expected values" (21 / 12), is covered, while other topics are relatively unimportant, especially on the new survey.
- *Discrete Probability Distributions*. The "Poisson distribution" has proven very difficult to assess, with the item focusing on the memory-less property. A different approach is recommended for this topic.
- *Continuous Random Variables and Probability Distributions*. The "Normal distribution" is adequately covered and justifiably so. A notable gap is "Standardized normal," which is suggested as a complementary item to broaden coverage of this topic area. "Continuous uniform distribution" could be added more explicitly, based on Old (rank 6), although it could be implied by one item (new #16 / old #25)
- *Joint Probability Distributions*. Again, the most important topic is covered, "Covariance and correlation," with two items. However, both items are

correlation, suggesting an item to distinguish covariance and correlation to be sensible.

- *Parameter Estimation*. The "Central Limit Theorem" has ample coverage with five items. Sampling is assessed in only one retained item, while it is relatively important ("Random sampling" and "Sampling distributions"). Another item to assess these aspects is reasonable.
- *Linear Regression*. This area is clearly important (rank 2 New and Old). Less apparent is whether this topic is appropriate for a test assessing introductory statistics topics. The topic of correlation is often encountered in other coursework, such as fitting lines in Excel for Chemistry lab data. A faculty survey of topics taught could help determine the coverage rate for statistics courses.
- *Time Series*. The lack of items is appropriate for this unimportant area.
- *Confidence Intervals and Hypothesis Testing*. This area was ranked surprisingly low on the older survey (9^{th}), but it is the most important on the new survey. Topic coverage could be better spread across topics aside from "Inference on the mean of a population" to areas such as "Inference on a population proportion" and "Type I (alpha) error." With four items cut, this area proved difficult to assess.
- *Single Factor Experiments* and *Multi-factor Designs*. These areas contain no topics of high importance, although the former rates relatively high on average at 5^{th}. These topics are typically taught in a second statistics course, such as Design of Experiments in an Industrial Engineering Department.

Table 4a: Topics surveys results (part 1)

	Rank New	Rank Old	Retained	Cut
Data Summary and Presentation	*3*	*3*	*(11)*	*(4)*
Importance of data summary	2	2	9, 11, 12, 15	--
Methods of displaying data	14	5	7, 25, 30	37
Percentiles and quartiles	40	43	8, 11	--
Measures of variability	2	1	26, 29	6
Skewness and kurtosis	70	76	25	--
Stem-and-leaf diagram	83	80	7	--
Frequency distribution and histogram	6	11	7, 25, 30	14, 28
Box plots	66	63	7	--
Time sequence plots	33	13	--	--
Probability	*4*	*1*	*(3)*	*(3)*
Sample space and events	18	16	21, 34	--
Axiomatic rules	58	26	--	--
Interpretation of probability	4	7	21	1
Addition rules	53	35	--	--
Conditional probability	33	22	31	--
Multiplication and total probability rules	24	24	21	16
Independence	17	13	31	5
Counting concepts	62	29	--	16
Bayes' Theorem	53	39	--	1
Random Variables	*10*	*4*	*(1)*	*(0)*
Expected values	21	12	34	--
Moment generating functions	85	55	--	--
Functions of random variables	63	30	--	--
Linear combinations	68	26	--	--
Discrete Probability Distributions	*8*	*8*	*(3)*	*(2)*
Discrete uniform distribution	63	30	25, 30	--
Binomial distribution	30	20	4	5
Geometric and negative binomial distribution	60	64	--	--
Hypogeometric distribution	71	59	--	--
Poisson distribution	30	8	--	13
Continuous Random Variables and Probability Distributions	*6*	*7*	*(3)*	*(0)*
Continuous uniform distribution	48	6	--	--
Normal distribution	1	4	23, 30, 33	--
Standardized normal	8	19	--	--
Normal approximations	28	54	--	--
Exponential distribution	30	48	--	--
Weibull distribution	79	78	--	--
Beta distribution	76	73	--	--
Lognormal distribution	78	41	--	--

Table 4b: Topics surveys results (part 2)

	Rank New	Rank Old	Retained	Cut
Joint Probability Distributions	*9*	*6*	*(0)*	*(0)*
Two discrete random variables	56	32	--	--
Multiple discrete random variables	75	47	--	--
Covariance and correlation	6	9	--	--
Bivariate normal distribution	75	51	--	--
Parameter Estimation	*7*	*5*	*(6)*	*(3)*
Random sampling	8	16	27	3
Properties of estimators	61	23	--	--
Sampling distributions	18	32	--	19
Central Limit Theorem	8	13	4, 20, 22, 34, 35	14
Estimators and their properties	56	45	--	--
Chebyshev's Inequality	81	72	--	--
Maximum Likelihood estimation	74	42	--	--
Linear Regression	*2*	*2*	*(2)*	*(1)*
Simple linear regression	12	3	--	--
Properties of the least squares	36	9	--	--
Confidence intervals for regression coefficients	41	24	--	--
Hypothesis tests in regression	33	38	--	--
F test of the regression model	44	60	--	--
Assessing the adequacy of regression	16	36	--	--
Use of regression for prediction	36	20	--	--
Correlation	12	16	24, 38	37
Time Series	*12*	*10*	*(0)*	*(0)*
Trend analysis	51	56	--	--
Seasonal and cyclic behavior	49	56	--	--
Ratio-to-moving-average method	83	48	--	--
Exponential smoothing methods	72	51	--	--
Confidence Intervals and Hypothesis Testing	*1*	*9*	*(6)*	*(4)*
Inference on the mean of a population	5	34	10, 17, 20, 22, 35	18
Inference on the variance of a population	29	39	--	--
Inference on a population proportion	11	44	--	5
Testing for goodness of fit	22	28	--	--
Contingency table tests	45	75	--	--
Inference on the means of two normal populations	20	53	20	36
Paired comparisons	26	58	2	--
Inference on the variance of two normal populations	55	67	--	--
Inference on two population proportions	26	64	--	--
Sample size determination	22	37	--	--
Type I (alpha) error	15	--	--	18, 32
Type II (beta) error	25	--	--	--
Power	39	--	--	18

Table 4c: Topics surveys results (part 3)

	Rank New	Rank Old	Retained	Cut
Single Factor Experiments	*5*	*11*	*(0)*	*(0)*
Analysis of fixed effects	45	67	--	--
Estimation of model parameters	49	46	--	--
Model adequacy check	45	62	--	--
Comparison of treatment means	43	71	--	--
Sample size	38	48	--	--
Non-parametric ANOVA	67	67	--	--
ANCOVA	69	76	--	--
Multi-factor Designs	*11*	*12*	*(0)*	*(0)*
Randomized complete block design	58	64	--	--
Latin square design	80	78	--	--
Graeco-Latin square design	81	82	--	--
Two-factor factorial design	42	67	--	--
General factorial design	51	74	--	--
Factorial design with random factors	65	81	--	--
Expected mean squares	72	60	--	--

Table 5 (next page) summarizes the coverage of the SCI relative to the 25 most important topics; the number 25 is arbitrary. The first section of Table 5 contains topics rated in the top 25 on both New and Old surveys, sorted by New. The coverage is very strong, with 14 of 16 topics covered, most by multiple items. The second section of the table contains items rated in the top 25 on New but not Old, while the third section is analogous for Old but not New. These lower sections suggest areas which may merit additional coverage. Cut items, such as Poisson distribution and Type I error, were poor items and re-specification is advised.

Table 5: Coverage of Top 25 Important Topics, for 25-item SCI

Topic		New	Old
Top 25 New and Old			
Normal dist.	√	1	4
Measure of variability	√	2	1
Importance of data summary	√	2	2
Interpretation of prob.	√	4	7
Frequency dist and histograms	√	6	11
Covariance and correlation	√	6	9
The central limit theorem	√	8	13
Standardized normal		8	19
Random sampling	√	8	16
Simple linear regression		12	3
Correlation	√	12	16
Methods of displaying data	√	14	5
Independence	√	17	13
Sample space and events	√	18	16
Expected values	√	21	12
Multiplication and total prob rules	√	24	24
Summary	*14 of 16*		
Top 25 New only			
Inference on the mean of a pop.	√	5	34
Inference on a pop prop.		11	44
Type I (alpha) error		15	--
Assessing the adequacy of reg.		16	36
Sampling dist.		18	32
Inference on means of 2 norm pop.	√	20	53
Testing for a goodness of fit		22	28
Sample size determination		22	37
Type II (beta) error		25	--
Summary	*2 of 9*		
Top 25 Old only			
Continuous uniform dist.		48	6
Poisson dist.		30	8
Properties of the least squares		36	9
Time sequence plot		33	13
Use of the reg. for prediction		36	20
Binomial dist.	√	30	20
Conditional prob.	√	33	22
Properties of estimators		61	23
Confidence intervals for the reg.		41	24
Summary	*2 of 9*		

4. Conclusions and Proposals

This chapter assessed the content validity of the Statistics Concept Inventory. The prior chapter proposed a shortened version of the instrument with 25 items rather than 38, intended to maximize the psychometric properties. The results of this chapter suggest the editing maintained an acceptable level of content validity, covering 14 of the 16 most important topics from a synthesis of two faculty surveys.

4.1 Proposals

Further assessments of the content validity of the items are important as well. The author independently conducted the classification presented herein. Hambleton (1980) describes three approaches to assess item validity by gauging expert opinion. The first of these methods asks raters to assign one of three values to each item: -1 (the item does not measure the objective), 0 (neutral), and +1 (the item definitely measures the objective). Each item is rated along all possible objectives and ideally attains +1 ratings on the hypothesized objective and -1 on all other objectives. Unfortunately, this task can be very time consuming (e.g., 500 ratings for a 50-item test with 10 objectives). The second method is a simpler rating scale, such as 1-to-5, asking if the item measures the intended objective. Low-level analysis, such as median or mean, is often sufficient to assess the item validity with this method. A third method is a matching task in which raters are presented with a list of items and a list of objectives, with instructions to match the items to their objectives. A contingency table analysis (raters × items) can be used to assess both item validity and inter-rater agreement.

A preliminary study of the SCI item effectiveness was conducted by publishing a link at the end of the topics survey. Respondents were asked “How appropriate is this

item to the given topic?" with responses on scale of 1 ("Not appropriate at all") to 4 ("Highly appropriate"). The "given topic" refers to the four sub-scales (descriptive, probability, graphical, inferential), which is listed with each item and its multiple choice options. Due to lack of time and publicity, the survey garnered only four participants.

To allow comparison with the faculty survey, items could be classified into topics-survey areas. It may prove tedious to category with 87 categories. A two-stage method could be employed, whereby subjects place an item first into one of the twelve areas and then choose a specific area therein. Alternatively, subjects could be offered multiple choice of proposed topics for each item and choose the "best" classification. An example is the following:

Which topic does this item assess?
(show the item)
a) Importance of data summary
b) Percentiles and quartiles
c) Measures of variability
d) Box plots

Although only four students were interviewed, useful comments were offered on nearly every item. It is clear that an appropriate subject pool will give thoughtful responses to the instrument: all subjects took at least 40 minutes to complete the online test and discussed with the author for approximately the same amount of time. Individual interviews were successful without the logistics of planning focus groups and offering incentive such as free pizza.

Additional interviews should especially focus on those items previously labeled misconceptions. Five of the ten were cut due to poor item statistics, which perhaps meshes with prior studies on reliability (Chapter VI) that "trick" items lower reliability and have poor discrimination.

References

Hambleton, R.K. 1980. Test Score Validity and Standard-Setting Methods. Chapter 4 (pp. 80-123) in Criterion-Referenced Measurement. ed. R.A. Berk. The Johns Hopkins University Press: Baltimore.

Patton, M.Q. 1990. Qualitative Evaluation and Research Methods. 2nd Ed. Sage Publications: Newbury Park, CA.

Stone, A. 2006. A Psychometric Analysis of the Statistics Concept Inventory. Dissertation, University of Oklahoma.

Appendix 1 25 retained items (slightly edited)

New #1 (Old #2)

A certain diet plan claims that subjects lose an average of 20 pounds in 6 months on their plan. A dietitian wishes to test this claim and recruits 15 people to participate in an experiment. Their weight is measured before and after the 6-month period. Which is the appropriate test statistic to test the diet company's claim?

a) two-sample Z test
b) paired comparison t test
c) two-sample t test

Correct Answer: B
Topic Area: Inferential

New #2 (Old #4)

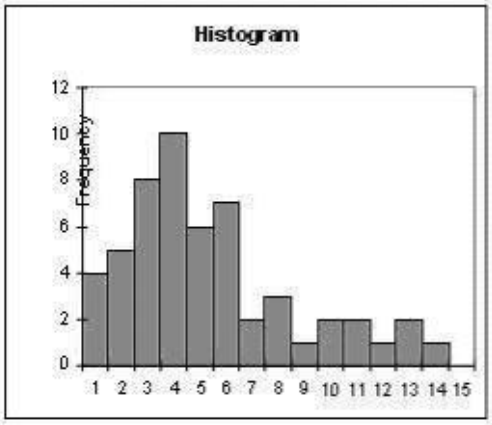
Which would be more likely to have 70% boys born on a given day: A small rural hospital or a large urban hospital?

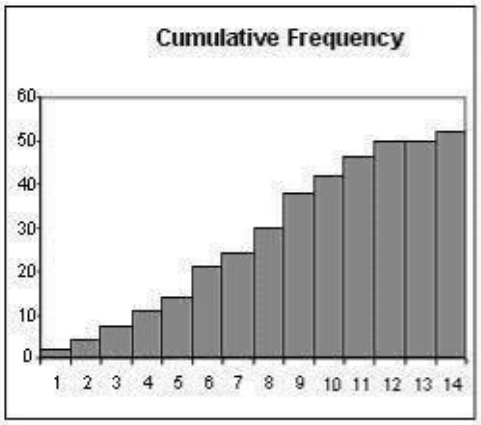
a) Rural
b) Urban
c) Equally likely
d) Both are extremely unlikely

Correct Answer: A
Topic Area: Probability

New #3 (Old #7)

Two of the following are graphical presentations of the same set of data. Which graph is of a different data set?





Stem	Leaf
0	55
1	79
2	124
3	1355
4	6
5	00235679
6	179
7	033348
8	0136799
9	00358
10	2679
11	1455
12	
13	22

a) Histogram
b) Cumulative Frequency
c) Stem and Leaf

Correct Answer: A
Topic Area: Graphical

New #4 (Old #8)

A student scored in the 90th percentile in his Chemistry class. Which is always true?

a) His grade will be an A
b) He earned at least 90% of the total possible points
c) His grade is at least as high as 90% of his classmates
d) None of these are always true

Correct Answer: C
Topic Area: Descriptive

New #5 (Old #9)

The following are temperatures for a week in August: 94, 93, 98, 101, 98, 96, and 93. By how much could the highest temperature increase without changing the median?

a) Increase by 8°
b) Increase by 2°
c) It can increase by any amount
d) It cannot increase without changing the median.

Correct Answer: C
Topic Area: Descriptive

New #6 (Old #10)

A bottling company believes a machine is under-filling 20-ounce bottles. What will be the alternate hypothesis to test this belief?

a) On average, the bottles are being filled to 20 ounces.
b) On average, the bottles are not being filled to 20 ounces.
c) On average, the bottles are being filled with more than 20 ounces.
d) On average, the bottles are being filled with less than 20 ounces.

Correct Answer: D
Topic Area: Inferential

New #7 (Old #11)

Which of the following statistics is least impacted by extreme outliers?

a) Range
b) 3rd quartile
c) Mean
d) Variance

Correct Answer: B
Topic Area: Descriptive

New #8 (Old #12)

A student attended college A for two semesters and earned a 3.24 GPA (grade point average). The same student then attended college B for four semesters and earned a 3.80 GPA for his work there. How would you calculate the student's GPA for all of his college work? Assume that the student took the same number of hours each semester.

a) $\frac{3.24 + 3.80}{2}$

b) $\frac{3.24(2) + 3.80(4)}{2}$

c) $\frac{3.24(2) + 3.80(4)}{6}$

d) It is not possible to calculate the students overall GPA without knowing his GPA for each individual semester.

Correct Answer: C
Topic Area: Descriptive

New #9 (Old #15)

For the following set of numbers, which measure will most accurately describe the central tendency? 3, 4, 5, 5, 6, 8, 10, 12, 19, 36, 83

a) Mean
b) Median
c) Mode
d) Standard deviation

Correct Answer: B
Topic Area: Descriptive

New #10 (Old #17)

A researcher conducts an experiment and reports a 95% confidence interval for the mean. Which of the following must be true?

a) 95% of the measurements can be considered valid
b) 95% of the measurements will be between the upper and lower limits of the confidence interval
c) 95% of the time, the experiment will produce an interval that contains the population mean
d) 5% of the measurements should be considered outliers

Correct Answer: C
Topic Area: Inferential

New #11 (Old #20)

The mean height of American college men is 70 inches, with standard deviation 3 inches. The mean height of American college women is 65 inches, with standard deviation 4 inches. You conduct an experiment at your university measuring the height of 100 American men and 100 American women. Which result would most surprise you?

a) One man with height 79 inches
b) One woman with height 74 inches
c) The average height of women at your university is 68 inches
d) The average height of men at your university is 73 inches

Correct Answer: D
Topic Area: Inferential

New #12 (Old #21)

A meteorologist predicts a 40% chance of rain in London and a 70% chance in Chicago. What is the most likely outcome?

a) It rains only in London
b) It rains only in Chicago
c) It rains in London and Chicago
d) It rains in London or Chicago

Correct Answer: D
Topic Area: Probability

New #13 (Old #22)

You perform the same two significance tests on large samples from the same population. The two samples have the same mean and the same standard deviation. The first test results in a p-value of 0.01; the second, a p-value of 0.02. The sample mean is equal for the 2 tests. Which test has a larger sample size?

a) First test
b) Second test
c) Sample sizes equal
d) Sample sizes are not equal but there is not enough information to determine which sample is larger

Correct Answer: A
Topic Area: Inferential

New #14 (Old #23)

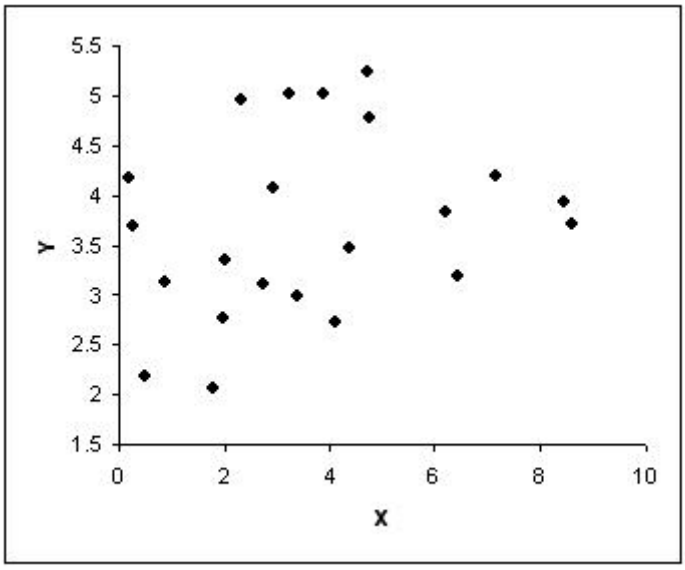
Which statistic would you expect to have a normal distribution? I) Height of women II) Shoe size of men III) Age in years of college freshmen

a) I & II
b) II & III
c) I & III
d) All 3

Correct Answer: A
Topic Area: Descriptive

New #15 (Old #24)

Estimate the correlation coefficient for the two variables X and Y from the scatter plot below.

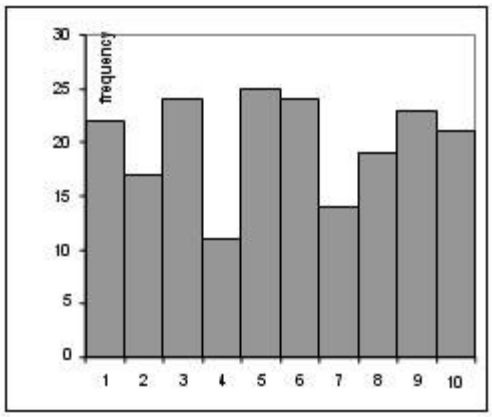


a) -0.3
b) 0
c) 0.3
d) 0.9
e) 1.6

Correct Answer: C
Topic Area: Graphical

New #16 (Old #25)

Consider the sample distribution below. This sample was most likely taken from what kind of population distribution?



a) Normal
b) Uniform
c) Skewed
d) Bimodal

Correct Answer: B
Topic Area: Graphical

New #17 (Old #26)

You have a set of 30 numbers. The standard deviation from these numbers is reported as zero. You can be certain that:

a) Half of the numbers are above the mean
b) All of the numbers in the set are zero
c) All of the numbers in the set are equal
d) The numbers are evenly spaced on both sides of the mean

Correct Answer: C
Topic Area: Descriptive

New #18 (Old #27)

In order to determine the mean height of American college students, which sampling method would not introduce bias?

a) You randomly select from the university basketball team
b) You use a random number table to select students based on their student ID
c) You roll a pair of dice to select from among your friends
d) None of the methods will have bias

Correct Answer: B
Topic Area: Inferential

New #19 (Old #29)

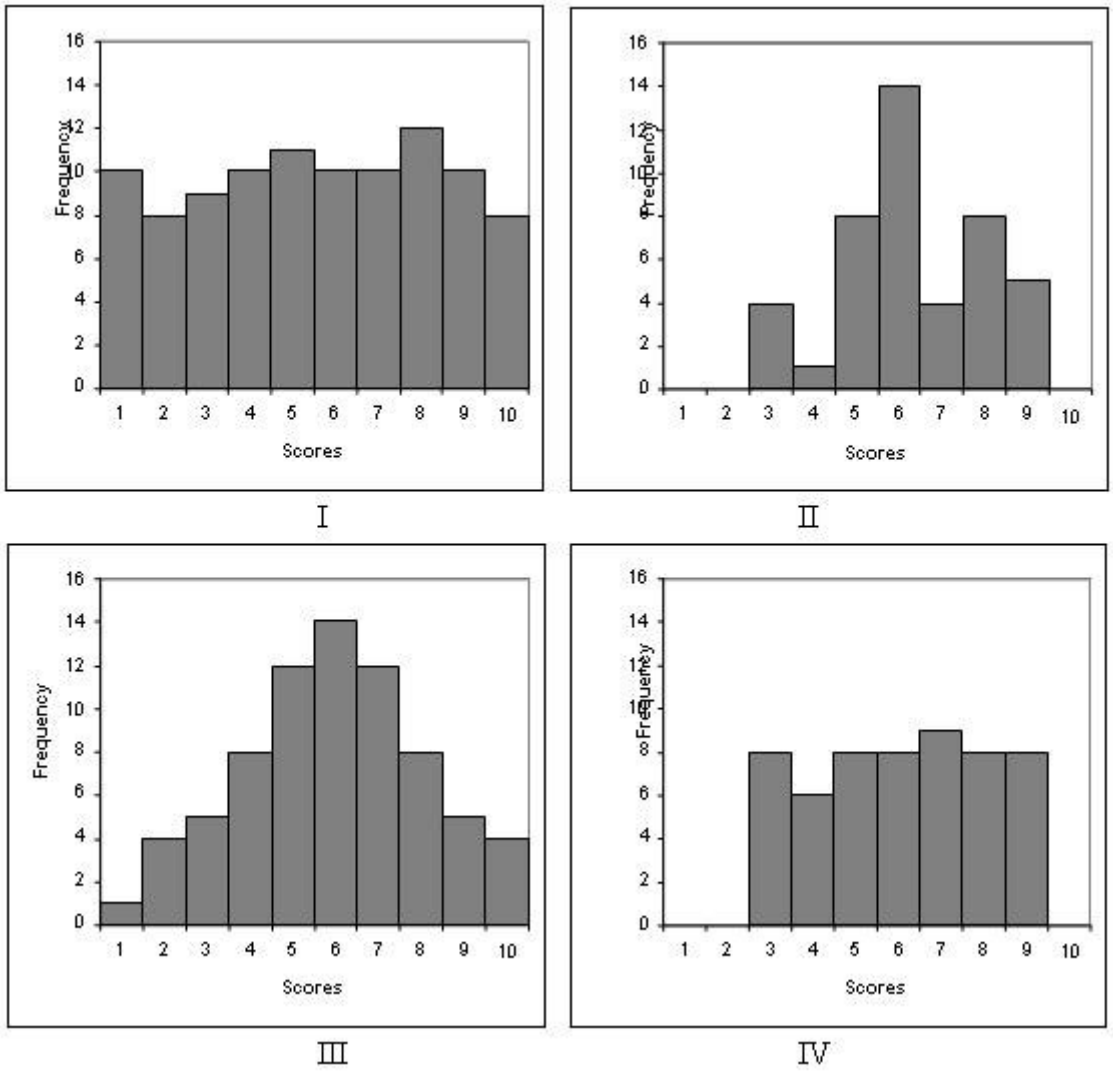
A scientist takes a set of 50 measurements. The standard deviation is reported as -2.30. Which of the following must be true?

a) Most of the measurements were negative
b) All of the measurements are less than the mean
c) All of the measurements were negative
d) The standard deviation was calculated incorrectly

Correct Answer: D
Topic Area: Descriptive

New #20 (Old #30)

The following are histograms of quiz scores for four different classes. Which distribution shows the most variability?



a) I
b) II
c) III
d) IV

Correct Answer: A
Topic Area: Graphical

New #21 (Old #31)

In a manufacturing process, the error rate is 1 in 1000. However, errors often occur in groups, that is, they are not independent. Given that the previous output contained an error, what is the probability that the next unit will also contain an error?

a) Less than 1 in 1000
b) Greater than 1 in 1000
c) Equal to 1 in 1000
d) Insufficient information

Correct Answer: B
Topic Area: Probability

New #22 (Old #33)

For the past 100 years, the average high temperature on October 1 is 78° with a standard deviation of 5°. What is the probability that the high temperature on October 1 of next year will be between 73° and 83°?

a) 0.68
b) 0.95
c) 0.997
d) 1

Correct Answer: A
Topic Area: Probability

New #23 (Old #34)

You are rolling dice. You roll 2 dice and compute the mean of the numbers rolled, then 6 dice and compute the mean, then 10 dice and compute the mean. Under which scenario would you be most surprised to find a mean of at least 4.5?

a) Rolling 2 dice
b) Rolling 6 dice
c) Rolling 10 dice
d) There is no way this can happen

Correct Answer: C
Topic Area: Probability

New #24 (Old #35)

Two confidence intervals are calculated for two samples from a given population. Assume the two samples have the same standard deviation and that the confidence level is fixed. Compared to the smaller sample, the confidence interval for the larger sample will be:

a) Narrower
b) Wider
c) The same width
d) It depends on the confidence level

Correct Answer: A
Topic Area: Inferential

New #25 (Old #38)

Information about different car models is routinely printed in public sources such as Consumer Reports and new car buying guides. Data was obtained from these sources on 1993 models of cars. For each car, engine size in liters was compared to the number of engine revolutions per mile. The correlation between the two was found to be -0.824. Which of the following statements would you most agree with?

a) A car with a large engine size would be predicted to have a high number of engine revolutions per mile.
b) A car with a large engine size would be predicted to have a low number of engine revolutions per mile.
c) Engine size is a poor predictor of engine revolutions per mile.
d) Engine size is independent of revolutions per mile.

Correct Answer: B
Topic Area: Descriptive

Appendix 2 13 deleted items (slightly edited)

Old #1

You are a doctor testing a blood-born disease. You know that in the overall population, 2 out of 100 people have the disease. All positives are accurately detected. You also know that the test returns a positive for 5 out of 100 people tested who do not have the disease. Portions of the related contingency table are given below. What is the probability that a patient will test positive?

	Has the disease (+)	Does not have the disease (-)
Tests positive (+)		
Tests negative (-)		0.95*0.98
	0.02	

a) 0.02
b) 0.05*0.98
c) 0.02 + 0.05*0.98
d) 0.95*0.98
e) 0.02+0.05

Correct Answer: C
Topic Area: Probability

Old #3

In practice, which data collection strategy would be the best way to estimate the mean household income in the United States?

a) every household within the United States
b) 1500 randomly selected households in the United States
c) 10 random households within each of 150 random US counties
d) 1500 is not a large enough sample

Correct Answer: C
Topic Area: Descriptive

Old #5

A coin of unknown origin is flipped twelve times in a row, each time landing with heads up. What is the most likely outcome if the coin is flipped a thirteenth time?

a) Tails, because even though for each flip heads and tails are equally likely, since there have been twelve heads, tails is slightly more likely
b) Heads, because this coin has a pattern of landing heads up
c) Tails, because in any sequence of tosses, there should be about the same number of heads and tails
d) Heads and tails are equally likely

Correct Answer: B
Topic Area: Probability

Old #6

An Olympic track team consists of 6 sprinters (2 compete in the 100 meter event, 2 compete in the 200 meter event, and the remaining 2 compete in the 400 meter event). For which of the following samples would you expect to calculate the largest variance?

a) A randomly selected sprinter's running times for 15 trials of the 200 meter event
b) The track team's (all six members) running times for the 200 meter event
c) A randomly selected sprinter's running times for 5 trials each of the 100 meter, 200 meter and 400 meter events
d) The track team's running times for the 100 meter, 200 meter, and 400 meter events, each person running all three events

Correct Answer: D
Topic Area: Descriptive

Old #13

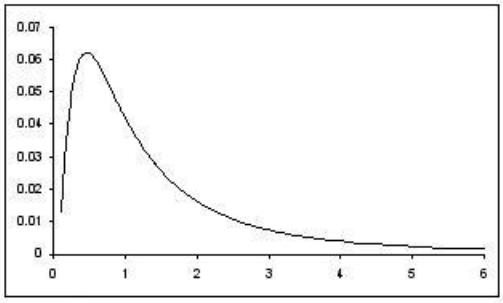
You have called your cell phone provider to discuss a discrepancy on your billing statement. Your call was received and placed on hold to 'await the next available service representative.' You are told that the average waiting time is 6 minutes. You have been on hold for 4 minutes. How many more minutes do you anticipate you will have to wait before speaking to a service representative?

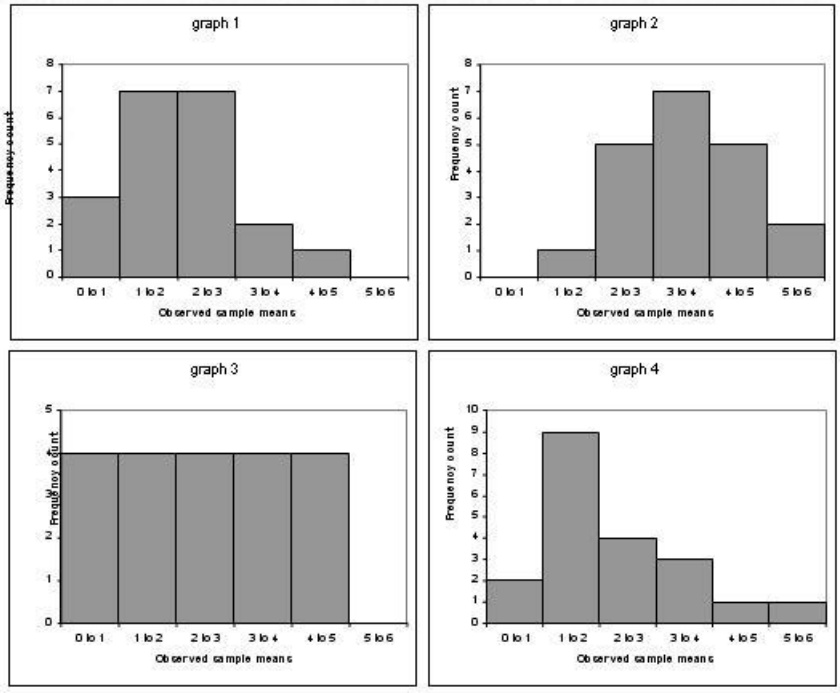
a) 2
b) 4
c) 6
d) there is no way to estimate

Correct Answer: C
Topic Area: Probability

Old #14

From the probability density function shown below ($\mu = 1.97$), 10 random data points are drawn and the mean is computed. This is repeated 20 times. The observed means were placed into six bins to construct a histogram. Which of the following histograms is most likely to be from these 20 sample means?





a) Graph 1
b) Graph 2
c) Graph 3
d) Graph 4

Correct Answer: A
Topic Area: Graphical

Old #16

A standard deck of 52 cards consists of 13 cards in each of 4 suits: hearts (H), diamonds (D), clubs (C), and spades (S). Five separate, standard decks of cards are shuffled and the top card is drawn from each deck. Which of the following sequences is least likely

a) HHHHH
b) CDHSC
c) SHSHS
d) All three are equally likely.

Correct Answer: D
Topic Area: Probability

Old #18

A researcher performs a t-test to test the following hypotheses: He rejects the null hypothesis and reports a p-value of 0.10. Which of the following must be correct?

$$H_0 : \mu \leq \mu_0$$
$$H_1 : \mu > \mu_0$$

a) The test statistic fell within the rejection region at the alpha = 0.05 significance level
b) The power of the test statistic used was 90%
c) Assuming the null is true, there is a 10% possibility that the observed value is due to chance
d) The probability that the null hypothesis is not true is 0.10

Correct Answer: C
Topic Area: Inferential

Old #19

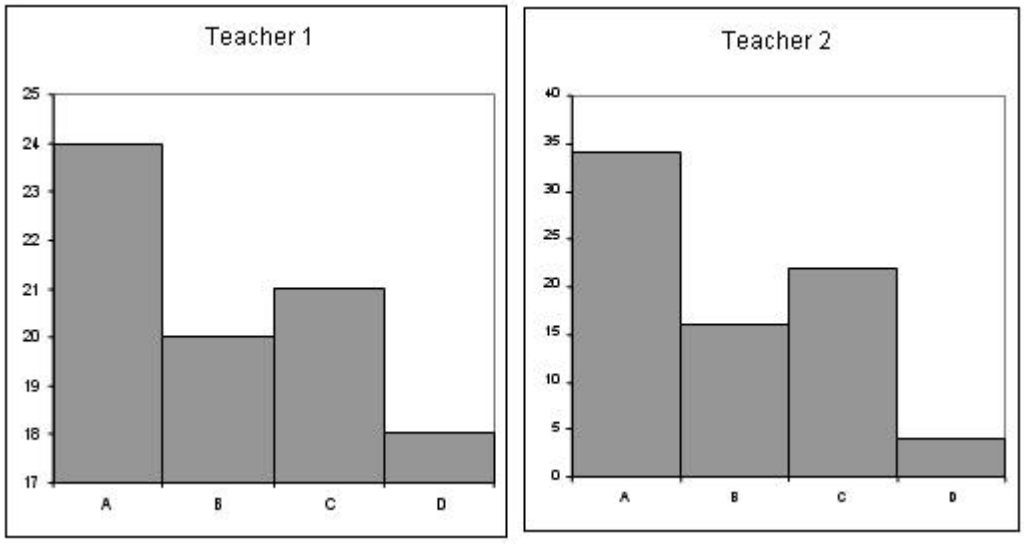
Which is true of a t-distribution?

a) It is used for small samples
b) It is used when the population standard deviation is not known
c) It has less extreme critical values than a Z-distribution for a given significance level
d) a & b are both true
e) a, b & c are all true

Correct Answer: D
Topic Area: Inferential

Old #28

The following histograms show the number of students receiving each letter grade for two separate physics classes. Which conclusion about the grades is valid?



a) Teacher 1 gave more B's and C's but approximately the same number of A's and D's as Teacher 2
b) Teacher 2 gave more A's and fewer D's than Teacher 1
c) Teacher 2 gave more B's and C's than Teacher 1
d) The overall grade distribution for the two Teachers is approximately equal

Correct Answer: B
Topic Area: Graphical

Old #32

An engineer performs a hypothesis test and reports a p-value of 0.03. Based on a significance level of 0.05, what is the correct conclusion?

a) The null hypothesis is true.
b) The alternate hypothesis is true.
c) Do not reject the null hypothesis.
d) Reject the null hypothesis

Correct Answer: D
Topic Area: Inferential

Old #36

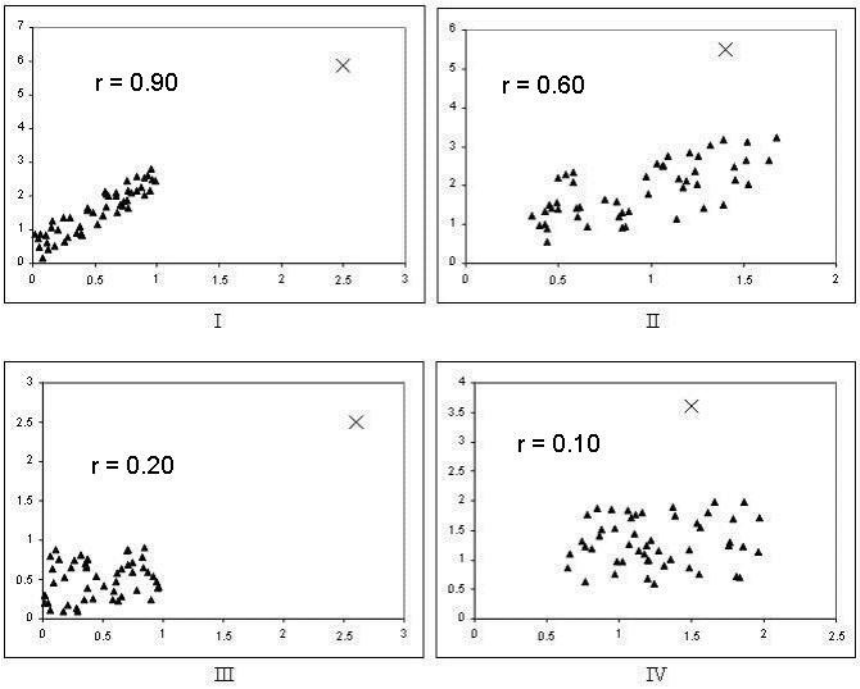
A chemical company has decided to begin producing a new product. They want to use existing equipment. An engineer is assigned to determine which of two reactor settings will yield the most pure product. He performs ten runs at each of the settings and measures the purity. Which test is most appropriate for this analysis?

a) Two-sample Z test
b) Paired comparison t test
c) Two-sample t test
d) One-sample t test

Correct Answer: C
Topic Area: Inferential

Old #37

Consider the correlation coefficients of the scatter plots below. If the data point that is marked by an X is removed, which of the following statements would be true?



a) correlation of (I) decreases, correlation of (II) stays the same
b) correlation of (III) increases, correlation of (IV) increases
c) correlation of (I) stays the same, correlation of (III) decreases
d) correlation of (II) increases, correlation of (III) increases

Correct Answer: C
Topic Area: Graphical

CHAPTER XI

The Concept Inventory Cookbook: A Comparative Study of Methods

con·cept
(kŏn′sĕpt′) *n.*

A general idea derived or inferred from specific instances or occurrences.

in·ven·to·ry
(ĭn′vən-tôr′ē) *n.*

A detailed, itemized list, report, or record of things in one's possession, especially a periodic survey of all goods and materials in stock.

Abstract

Ingredients: 1 part topics survey
2 parts focus groups
Psychometrics, to taste
A dash of personal opinion

Instructions: Combine ingredients as seen fit.
Repeat, if desired.

What is a concept inventory?
Who writes them? Who uses them?
How are they analyzed?
How are they used?

1. Introduction

Concept inventories began with the highly successful Force Concept Inventory in the early 1980's. The engineering education community followed suit starting around the late 1990's, with 15 instruments known to be in various stages of development. Other science fields have caught on as well, although not to the extent of engineering.

This chapter serves as a comparative review of the state of the genre – a sort of *greatest hits* if you will. Exhaustive documentation of the instruments was presented as Chapter 3, to introduce the reader. Taking a more critically evaluative stance, the following questions are asked:

- What is a concept inventory?
- Who uses concept inventories?
- How are concept inventories analyzed?
- How else could concept inventories be used?

With possible answers to these questions, the Statistics Concept Inventory can be held up to these standards in the grand finale.

2. Force Concept Inventory

There is little doubt that the Force Concept Inventory has a lasting and continuing influence not only in physics education research but expanding into other fields as well. A cited reference search (January 1, 2006) yielded 127 hits for the original article of what was at that time an unnamed diagnostic instrument (Halloun and Hestenes, 1985) and 109 hits for what was formally named the FCI (Hestenes, *et al.,* 1992).

Despite its popularity and success, it is surprising that little mention is made of what might be considered psychometrics: the 1985 article contained correlation analysis akin to concurrent and predictive validity, and Cronbach's alpha was calculated presumably across the combined dataset but even this is unclear. Although only "about half" of the items on the 1992 FCI were the same, the most formal assessment is comparison between scores on the 1985 and 1992 instruments, and student interviews which yielded compatible results between their respective test answers as well the earlier

findings. In fact, “formal procedures” are eschewed because “the test designs are so similar and such diverse data are presented here” (Hestenes, *et al.,* 1992, p. 151), referring to the interviews and scores. Each article contains nine references, all physics textbooks or other research about student (pre-,mis-)conceptions; there are no references related to test theory or analysis.

This should not be viewed as a potshot, however. The content validity is meticulously documented through use of previous research, student interviews, and item taxonomy. The fact that it continues to be used to assess learning and pedagogy is what makes it a milestone, over 20 years since its first publication.

3. Other Concept Inventories

The success of the FCI led researchers to develop concept inventories in a number of other fields, primarily other Physics and Engineering disciplines. Table 1 (next page) lists the concept inventories reviewed in Chapter III.

Table 1: List of Concept Inventories (CI) and similar instruments

Instrument	Abbreviation	Authors (year)
p h y s i c s		
Physics diagnostic instrument	none	Halloun and Hestenes (1985)
Force CI	FCI	Hestenes, *et al.* (1992)
Mechanics Baseline Test	MBT	Hestenes and Wells (1992)
Test of Understanding Graphs in Kinematics	TUG-K	Beichner (1994)
Force and Motion Conceptual Evaluation	FMCE	Thornton and Sokoloff (1998)
Conceptual Survey of Electricity and Magnetism	CESM	Maloney, *et al.* (2001)
Determining and Interpreting Resitive Electric Circuit Concepts Test	DIRECT	Engelhardt and Beichner (2004)
e n g i n e e r i n g		
Materials CI	MCI	Krause, *et al.* (2003 and 2004b)
Heat Transfer CI	HTCI	Jacobi, *et al.* (2003)
Fluid Mechanics CI	FMCI	Martin, *et al.* (2003 and 2004)
Statics CI	none	Steif (2003 and 2004) Steif and Dantzler (2005) Steif and Hansen (2006)
Thermal and Transport Science CI	TTSCI	Olds, *et al.* (2004)
Dynamics CI	DCI	Gary, *et al.* (2003 and 2005)
Wave CI	WCI	Roedel, *et al.* (1998) Rhoads and Roedel (1999)
Circuits CI	CCI	Helgeland and Rancour
Computer Engineering CI	CPECI	Michel, *et al*
Electromagnetics CI	EMCI	Notaros
Electronics CI	ECI	Simoni, *et al.* (2004)
Signals and Systems CI	SSCI	Wage, *et al.* (2002 and 2005)
Strength of Materials CI	SOMCI	Richardson, *et al.* (2003); Morgan and Richardson
Thermodynamics CI	none	Midkiff, *et al.* (2001)
Chemistry CI	CCI	Pavelich, *et al.* (2004) Krause, *et al.* (2004a)
o t h e r s		
CI of Natural Selection	CINS	Anderson, *et al.* (2002)
Chemical equilibrium (Test to Identify Students' Conceptualizations)	TISC	Voska and Heikkinen (2000)
Geoscience Concept Inventory	GCI	Libarkin and Anderson (2005)

4. Best Practices

4.1 General Considerations

Beichner (1994), developer of the TUG-K, offers a general blueprint for concept inventory construction. His flowchart is re-produced in Figure 1. This should not be taken as gospel. For example, a validity check of the objectives can be conducted before writing items, depending on the method in which the objectives were formulated.

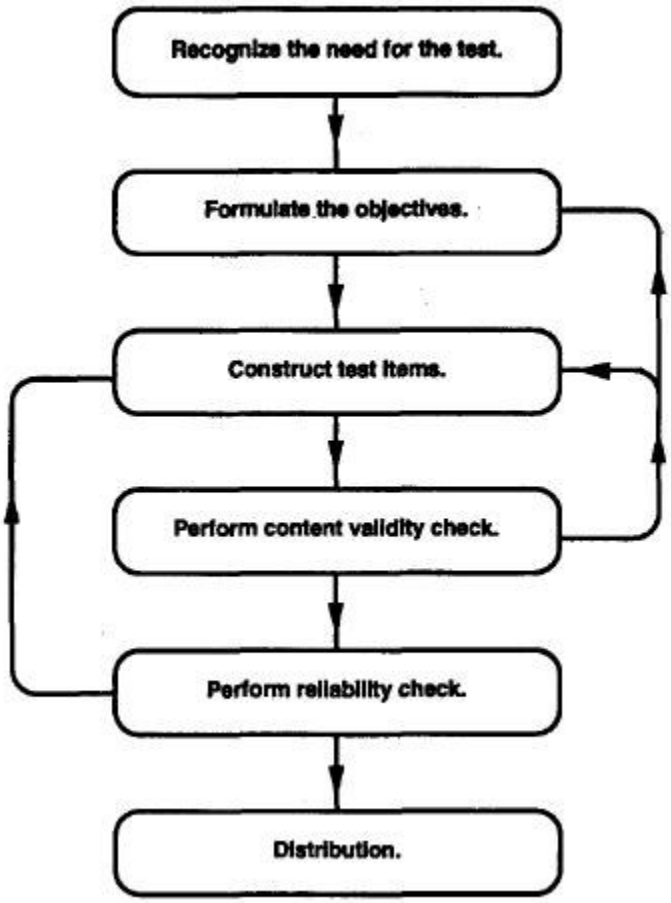


Figure 1: General model for concept inventory development (Beichner, 1994)

Students' scores are certainly an important consideration for making further judgments about a concept inventory. Table 2 shows the most recent pre- and post-test scores reported in the papers reviewed.

Table 2: Concept Inventory scores

Instrument	Pre-Test	Post-Test	Comment
FCI	low 50%	mid 60%	Traditional, calculus-based
MBT	--	around 30% 40% to 60%	High schools Universities
MCI	--	gain 15% to 20% gain 38%	Traditional Active
CINS	--	46%	--
TISC	--	64% 49%	Answers only Reasons
Statics CI	--	39%	--
TTSCI	--	52%	--
DCI	31% to 35%	56% to 63%	--
WCI	10.4 11.8	11.9 15.2	Traditional Integrated
SSCI	--	+20% gain +37% gain	Traditional Interactive
Chem. CI	27% 36%	53% 55%	Chem I Chem II
CESM	25% 31% 41%	44% 47% 69%	Algebra-based Calculus-based Honors Calc-based
TUG-K	--	40%	--
FMCE	--	+20% gain > +50% gain	Traditional Interactive & Lab
DIRECT	--	41%	--

The Geoscience Concept Inventory (GCI) is the only instrument known to utilize item response theory (IRT) and test equivalency (Libarkin and Anderson, 2005). The novel test construction consisted of 11 core items and two sets of 9 items, yielding two versions with 20 items total. An IRT Rasch model was fit to items and used to compute an ability estimate for each examinee. More complex models, which can account for item discrimination (slope of the logistic curve) and guessing (lower asymptote > 0), were not discussed. A test-equating step was performed to convert the raw GCI scores onto a 0 to 100 scale (it is stated that a method paper is in preparation). The test scores, as often

encountered, were not encouraging (pre-test 43, post-test 47, based on 930 matched students). Publication of IRT is an important step forward for concept inventories.

4.2 Teaching Implications

As noted in Table 2, concept inventory scores are sometimes used to make inferences about the effectiveness of teaching styles. Hake (1998) demonstrated, using the FCI, that students have little conceptual gain in traditional lecture-based physics courses. However, those students whose instructors engage in active learning practices gain much more conceptual knowledge, attaining average normalized gains twice as high in his study. Similar results have been found using the MCI, WCI, SSCI, and FMCE.

4.3 Similar Concept Inventories

Some concept inventories overlap in coverage or have similar areas of focus, while other concept inventories could be used in a sequence of related courses. In Physics, the content of the FCI, MBT, and TUG-K is quite similar. The TUG-K is the most specific, as it focuses only on kinematics graphs (i.e., position / velocity / acceleration vs. time). The FCI also includes graphs and items about velocity, etc., but the overall focus is on force. The MBT contains some items very similar to the TUG-K. The MBT was also based on the FCI, but it requires calculations on many problems. These three instruments could be used in introductory physics to assess student understanding of different types of problems and representations.

Many of the engineering concept inventories could be used in a sequenced curriculum and perhaps incorporate concept inventories from other fields. For example, the DIRECT seems most appropriate for use in a second-semester of introductory physics. The Electronics Concept Inventory or Circuits Concept Inventory would provide

good follow-up in a first electrical engineering course, although both instruments lack sufficient documentation at this point to be useful. The Electromagnetics, Systems and Signals, and Waves concept inventories could be used in the next level of coursework. Similar sequencing may be possible with Chemistry → Thermodynamics → Heat Transfer / Fluid Mechanics / Thermal & Transport Science; and FCI / MBT → Statics → Dynamics / Strength of Materials.

4.4 Content Validity

Content validity refers to the extent to which items are (1) representative of the knowledge base being tested and (2) constructed in a "sensible" manner (Nunnally, 1978). Achieving content validity requires finding an adequate sampling of possible topics and incorporating value judgments as to which topics to include in the instrument.

The Dynamics Concept Inventory (DCI) (Gary, *et al.,* 2003) provides a good blueprint for conducting a faculty survey to identify content for an exam. Twenty-five faculty members ranging from two-year colleges to research universities were asked to identify topics which they felt students had conceptual difficulty learning, as opposed to difficulty with problem-solving. This initial survey yielded 24 topics which were passed on to a second step in which the participants were asked to give the percentage of students they believed adequately learn the topic (re-scaled on 0 to 10) and the importance of the topic (also 0 to 10). The list was pared to 13 topics by eliminating those with average importance scores below 8. Several topics which have similarities were combined and an extra topic was included due to the authors' lists of the important topics, to yield a final list of 11 topics.

Focus groups are a crucial element for validating content from a student perspective. The Materials Concept Inventory (MCI) authors documented the focus group efforts used to revise the instrument (Krause, *et al.,* 2004b). Focus groups held for their initial study were not found to be as informative as desired. For the new focus groups, students were given only selected MCI questions, which helped guide the discussions. Group size was increased from two or three to six to ten, which made students more comfortable and willing to speak. The students were first given their 10 to 12 selected questions to answer individually. They then met in their group to discuss why they had chosen certain answers but were not told by the moderator which answer was correct until discussions had concluded. This format proved informative to the authors for developing and validating distracters and it also helped students gain a deeper understanding of the material.

The DIRECT authors (Engelhardt and Beichner, 2004) divided student interviews into three sections: 1) identification of symbols in the test; 2) definitions of terms in the test; and 3) answering the items, providing reasons and stating their confidence. The interviewer had access to each student's original answer, asking the student to recall his original reasoning if his answer changed from the original.

4.5 Construct Validity

A test is "constructed" to measure some latent ability of its subjects (Thorndike, 1982). This is often applied to personality scales where the desired measure may be a quality not directly observable, such as aggression or courage. This concept is extended to achievement tests to define sub-tests within a larger instrument. The hope is to find specific abilities within a larger domain of knowledge.

The Concept Inventory of Natural Selection (CINS) (Anderson, *et al.,* 2002) provides an example of factor analysis used to establish construct validity. A principal component analysis was conducted on the matrix of item phi correlation coefficients. A varimax rotation was used, and solutions with two to eight components were examined. The solution with seven components was found to be optimal because it had (a) a large proportion of the variance explained (53%); (b) all items loaded at least 0.40 on at least one component; (c) only one item loaded at least 0.40 on multiple components; and (d) nine of the ten evolution concepts grouped along the same respective component.

A confirmatory factor analysis was conducted for the Statics Concept Inventory by attempting to fit each item to one of eight hypothesized constructs (Steif and Dantzler, 2005). The overall model fit was not significantly different from the observed data (χ^2 = 0.22, d.f. 296, p = 0.22). The Goodness of Fit Index (GFI = 0.90) and Comparative Fit Index (CFI = 0.91) are at the bottom of the acceptable range, while the root mean square approximation (RMSEA = 0.067) is acceptable. The fit indices suggest that some item revision could improve the Inventory's structure, but these results are considered "acceptable" to the authors.

The Statics model can be viewed with some skepticism. By analyzing the degrees of freedom, the model apparently had eight latent factors with each item loading on only one of these, thus implying around 3 items per factor. The eight factors are *a priori* content domains, and no alternative models are presented for comparison.

4.6 Predictive and Concurrent Validity

Predictive validity refers to a test's ability to accurately predict future performance. This is often discussed in the context of training, such as whether a training

program can be considered valid at increasing job performance (Thorndike, 1982). This requires a decision to be made as to what constitutes success in a future endeavor. In an academic setting, future performance may be the grade earned for a course or graduation.

Concurrent validity is "assessed by correlating the test with other tests" (Klein, 1986). This requires a decision as to what constitutes the "other test." Of course, if another test already exists, it raises the question of whether the test being constructed is even necessary. Therefore, the term "other test" should be loosely interpreted. On concept inventories, a logical selection is the course grade or final exam score, with the caveat that a concept inventory does not focus on computational aspects of a field.

To assess validity of the Systems and Signals CI (Wage, *et al.,* 2005), a number of correlations between SSCI scores (pre-test, post-test, gain) and course grades (e.g., CT systems & signals, DT systems & signals, calculus, overall GPA) are computed. To highlight several of these, both CT-SSCI and DT-SSCI post-test and gain have significant positive correlations with their respective course letter grades (4.0 scale), a measure of concurrent validity. Pre-requisite course grades are used as a measure of predictive validity of the SSCI. This is backwards of how predictive validity may normally be assessed, but it is interesting nonetheless. For example, the CT course grade correlates significantly with the DT-SSCI pre-test and post-test and the Circuits course grade correlates positively with the CT-SSCI pre-test. GPA correlates positively with both SSCI (CT and DT) pre-test scores. The authors feel this type of correlation analysis provides valuable insight into the role of course sequencing and can evaluate whether signals and systems courses are conceptual in nature.

4.7 Reliability

A reliable instrument is one in which measurement error is small, which can also be stated as the extent that the instrument is repeatable (Nunnally, 1978). There are several types of reliability: test-retest measures answer stability on repeated administrations; alternative forms requires subjects to take two similar tests on the same subject; and internal consistency is based on inter-item correlations and describes the extent to which the test measures a single attribute.

Internal consistency is the most common measure because it requires only one test administration. This reduces administration costs and eliminates the issue of students gaining knowledge between test administrations. Internal consistency is measured using Cronbach's alpha (1951), which is a generalized form of Kuder-Richardson Formula 20 (1937). Typically, a test is considered to be reliable if alpha is above 0.80 (Nunnally, 1978). Other sources consider a value of 0.60 to 0.80 to be acceptable for classroom tests (Oosterhof, 1996). Table 3 shows the reported alpha values for concept inventories.

Table 3: Cronbach's alpha for concept inventories

Instrument	**Pre-Test**	**Post-Test**
FCI	0.86	0.89
CINS	--	0.58, 0.64
TISC	--	0.79
Statics CI	0.72	--
Chem. CI	--	Chem I: 0.7135
	--	Chem II: 0.4188
CESM	--	"around 0.75"

4.8 Discrimination

Discrimination refers to a test's ability to produce a wide range of scores. It is a desirable property because tests are designed to look for differences between subjects. The discriminatory power depends on the shape of the score distribution. For example, if

scores are normally distributed, it is easiest to differentiate between scores at the tails because there are few extreme scores; the middle scores are hard to differentiate because they are clustered. (Ausubel, 1968)

Discriminatory power of a full instrument can be measured by Ferguson's delta, which ranges from 0 (all scores the same) to 1 (each person has a unique score). A test is considered discriminating if delta is above 0.90 (Kline, 1986). Few concept inventories report Ferguson's delta, which is not a major oversight since it appears to be very easy to attain values over 0.90. The TUG-K (Beichner, 1994) reports a value of 0.98.

It is most important for each item to discriminate. Item discrimination is measured by the discriminatory index, which compares the top-scoring students to the low-scoring students. For example, if 75% of the top students and 30% of the bottom students get a question correct, the item has a discriminatory index of 0.45. To determine the top and bottom students, the optimal split is considered to be 27% at each end (Kelley, 1939). In practice, the first and third quartiles are used as the partition, due to ease of calculation.

The Strength of Materials CI (Richardson, *et al.,* 2003) offers a good presentation of discriminatory indices for all items. The graphic, reproduced below (Figure 2), allows a simple visual interpretation of which items are poor. However, it might be more suitable ordered by discriminatory index rather than by item number.

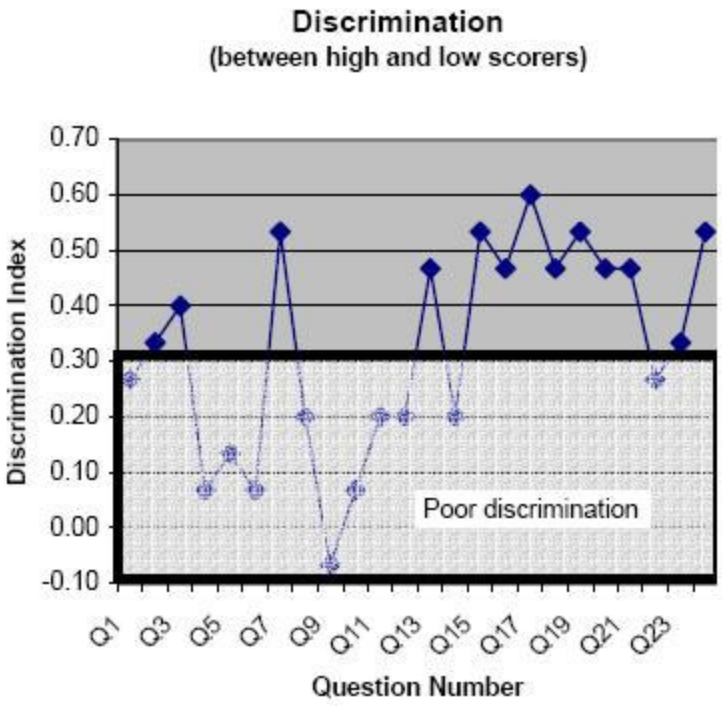


Figure 2: Discriminatory Indices from the Strength of Materials CI (Richardson, *et al.,* 2003)

5. Dissemination

The face validity of a concept inventory project is often gauged by the number of students who take the test and the length of time that the project exists. For engineering concept inventories, those with multiple publications were reviewed to evaluate these criteria. In most cases, exact numbers are provided or easily calculated from the publications. The first version of this graphic shows the extreme success of the Statics Concept Inventory relative to the others. The second version has this instrument removed to allow clearer comparison among the others. These data are based almost exclusively on what has been published, with the exception of the Year 4 datapoint for Statics, which was found on his website but is only for one semester (Fall 2005). It is therefore possible that more students have been surveyed for other concept inventories, but these are

unavailable due to lag in publication time; it may also be the case that the projects lacked momentum to move beyond the second year.

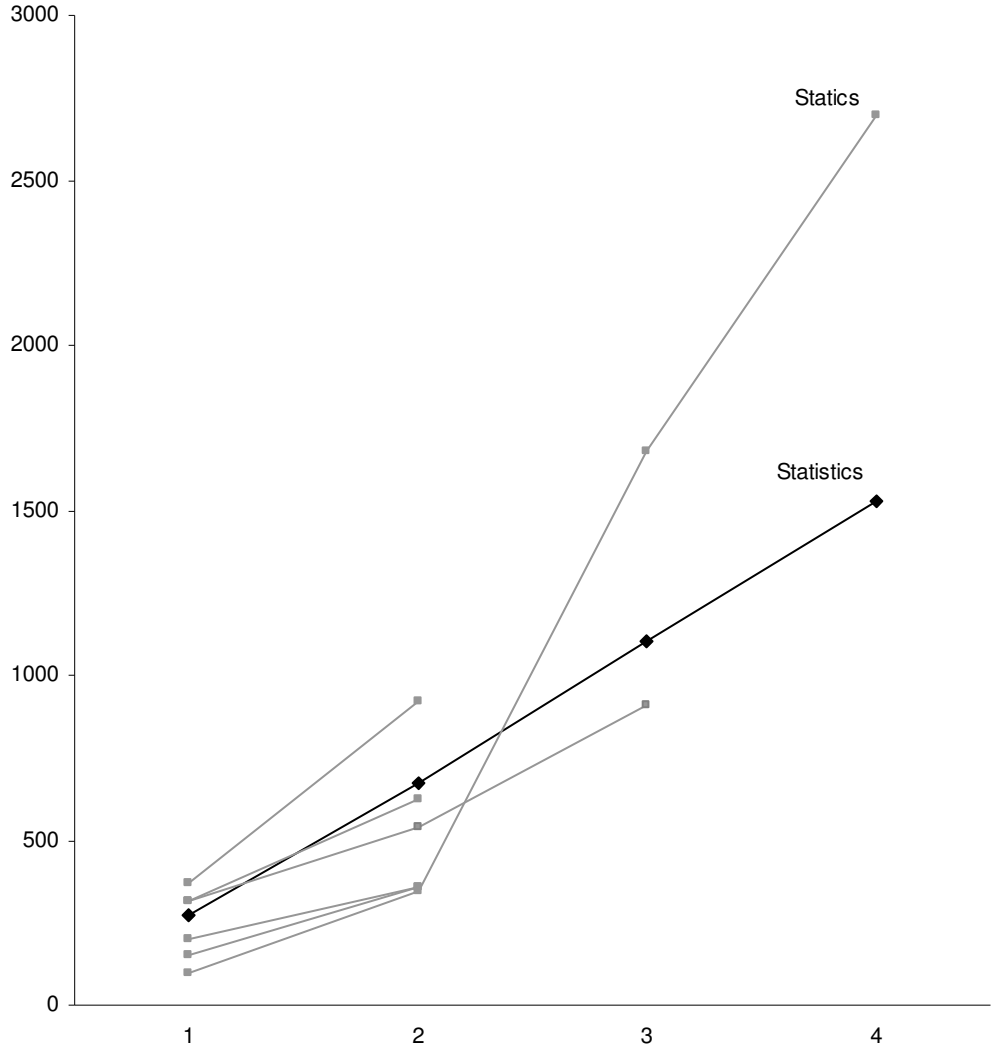


Figure 3a: Cumulative examinees across project years for engineering concept inventories (Statics dominates: see Figure 3b for other labels)

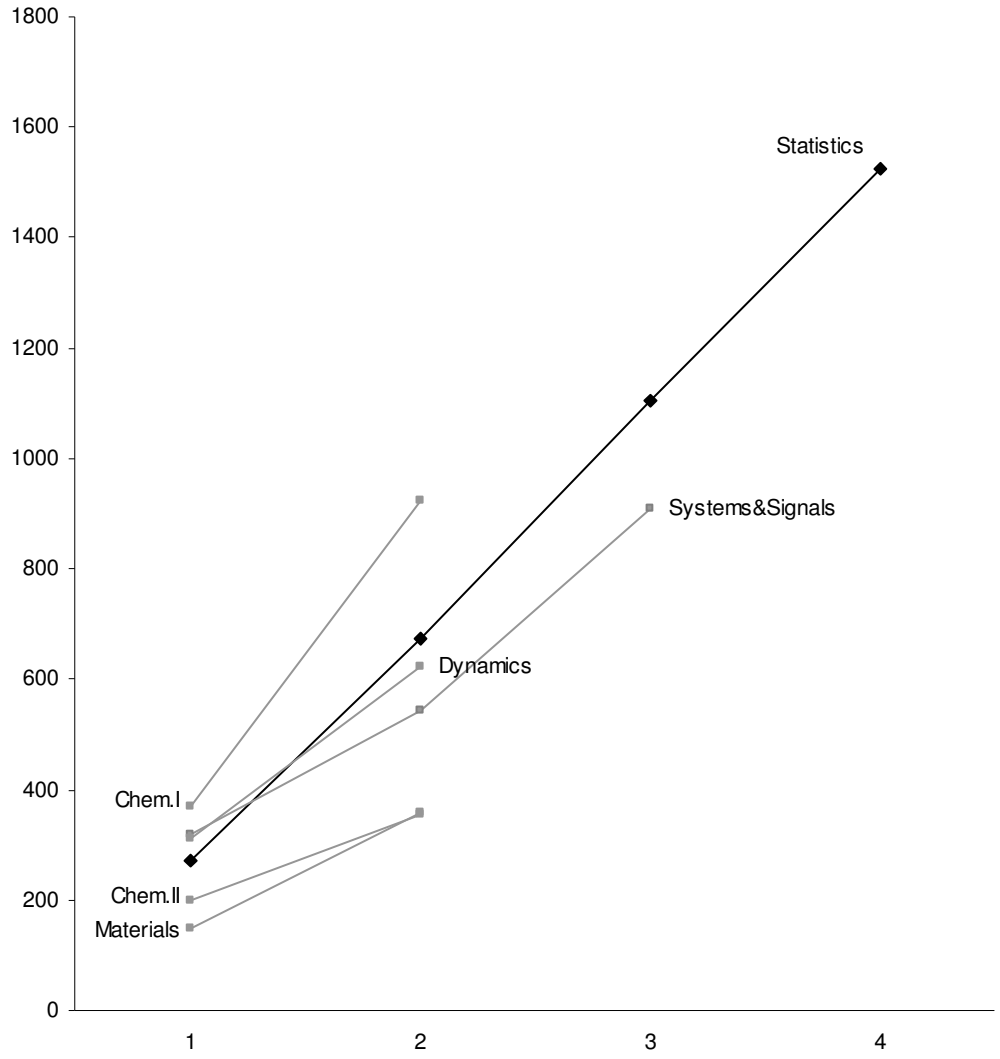


Figure 3b: Cumulative examinees across project years for engineering concept inventories (Statics removed)

6. Conclusions

What is a concept inventory?

A concept inventory is a multiple choice cognitive assessment instrument (read: "test") designed to assess students' conceptual understanding of a specific field or topic of science. The authors are educational researchers within their respective fields who are very likely writing their first test for an audience broader than their own classroom.

Who uses concept inventories?

This question is difficult to answer precisely. It is generally assumed that the authors of the inventories use them to assess their own courses. The reported sample sizes make it apparent that others use them as well:

How many of these are colleagues at the same university?

Or professional contacts through previous research?

Did the authors initiate contact? ("Hey! Take my test.")

Or did the participating instructors? ("That looks interesting; I'll give it a shot.")

These unanswered questions and often-reported low inventory scores raise the question of student motivation for what is typically a low-(or no-) stakes assessment.

How are concept inventories analyzed?

Content validity, through focus groups or faculty topics surveys, is mentioned for nearly every instrument. The discrimination index is often reported as an objective item metric. Advanced psychometric techniques of factor analysis (TISC), structural equation modeling (Statics), and IRT with test equating (GCI) are not wide-spread.

Correlational studies are a favorite method to assess the predictive and concurrent validity of the instruments, either by comparison to course grades or exam scores. With courses often focusing on computational ability, these analyses show that inventory

scores serve a complementary role to traditional assessment and are not intended as a proxy.

Scores are often used as a pre-test vs. post-test measure to assess students' conceptual knowledge gain. Teaching implications point to interactive engagement being crucial to increasing concept inventory scores.

How else could concept inventories be used?

With the proliferation of engineering instruments, wide-spread use within an institution could allow comparison longitudinally through a curriculum (e.g. FCI / MBT → Statics → Dynamics / Strength of Materials). If instruments can be validated extensively to ensure the teaching implications, a single institution who buys into the inventories could utilize them across departments to assess teaching methods and philosophies.

And finally…

How does the Statistics Concept Inventory hold up to these standards?

To be continued…

References

A full listing of concept inventory references can be found in the Book One references. This listing refers to those citations found in the text of this chapter.

Beichner, R.J. 1994. Testing student interpretation of kinematics graphs. *American Journal of Physics*. 62 (8): 750-755.

Engelhardt, P.V., and R.J. Beichner. 2004. Students' understanding of direct current resitive electrical circuits. *American Journal of Physics*. 72 (1): 98-115.

Gary, G.L., D. Evans, P. Cornwell, F. Costanzo, and B. Self. 2003. Toward a Nationwide Dynamics Concept Inventory Assessment Test. *Proceedings of the 2003 American Society for Engineering Education Annual Conference & Exposition*. Session 1168.

Hake, R. 1998. Interactive-engagement vs. traditional methods: A six-thousand student survey of mechanics test data for introductory physics courses. *American Journal of Physics*. 6 (1): 64-75.

Halloun, I. and D. Hestenes. 1985. The initial knowledge state of college physics students. *American Journal of Physics*. 53 (11): 1043-1055.

Hestenes, D., M. Wells, and G. Swackhamer. 1992. Force Concept Inventory. *The Physics Teacher*. 30 (March): 141-158.

Kelley, T. 1939. The Selection of Upper and Lower Groups for the Validation of Test Items. *Journal of Educational Psychology*. 30: 17-24.

Libarkin, J.C., and S.W. Anderson. 2005. Assessment of Learning in Entry-Level Geoscience Courses: Results from the Geoscience Concept Inventory. *Journal of Geoscience Education*. 53 (4, September): 394-401.

Nunnally, J. 1978. Psychometric Theory. McGraw-Hill: New York.

Richardson, J., P. Steif, J. Morgan, and J. Dantzler. 2003. Development Of A Concept Inventory For Strength Of Materials. *Proceedings of 33rd ASEE/IEEE Frontiers in Education Conference*. Session T3D-29.

Wage, K.E., J.R. Buck, C.H.G. Wright, and T.B. Welch. 2005. The Signals and Systems Concept Inventory. *IEEE Transactions on Education*. 48 (3): 448-461.

Book Four

Table of Contents

List of Tables 454
List of Figures 454

XII The Statistics Concept Inventory:
A Tool for Measuring Cognitive Achievement
in Introductory Statistics 455

1. Introduction 455
    1.1 Participation 456
    1.2 Scores and Reliability 458
2. Results 459
    2.1 Early Results 459
    2.2 Transitional Results 459
    2.3 Final Results 464
    2.4 Other Results 469
3. Conclusions and Recommendations 470
    3.1 Directions for Future Research 471
4. Process Model Re-Visited 473
5. Final Word 474

References 475

Full References 476
Statistics Concept Inventory 476
Concept Inventories 477
Statistics and Probability Reasoning and Assessment 481
Test Theory and Practice 483
Factor Analysis 486
Statistics Textbooks 487
Others 488

List of Tables

Chapter XII

Tables 1: Classes by Semester 457
Table 2: Comparison of Confirmatory Factory Analysis for Statistics and Statics 464
Table 3: Coverage of Top 25 Important Topics, for 25-item SCI 468

List of Figures

Chapter XII

Figure 1: Development of Statistics Concept Inventory, after Spring 2004 455
Figure 2: Sample size comparison between SCI and FCI 456
Figure 3: Post-test mean (left axis) and reliability (α, right axis) across semesters 458
Figure 4: Item discrimination index (bars) and difficulty (stars) for Fall 2005 Post-test 462
Figure 5: Confidence vs. fraction correct, rank orders 463
Figure 6: Proposed sub-domains for Statistics discipline 465
Figure 7: Item discrimination index (bars) and difficulty (stars) of 25 retained items 467
Figures 8: Multiple-response IRT curves 470
Figure 9: Future directions for the SCI 471
Figure 10: Test creation process 474

CHAPTER XII

The Statistics Concept Inventory: A Tool for Measuring Cognitive Achievement in Introductory Statistics

1. Introduction

Development of the Statistics Concept Inventory began in Fall 2002. A summary of the first two years' work was presented as Book One (Chapter V, Figure 1). Figure 1, below, takes the Spring 2004 version as the starting point for expansion. The first result was a determination of the factors influencing test reliability (Alpha node). Development of the Online Test was the next endeavor and led to the Confidence study. Factor Analysis was the next project, with the dashed arrow acknowledging that most of the data was gathered using the online test. The conclusions of the factor analysis led to a re-assessment of the content validity (Interviews, Faculty Survey). The interviews also delve into the reasons for confidence differences across items.

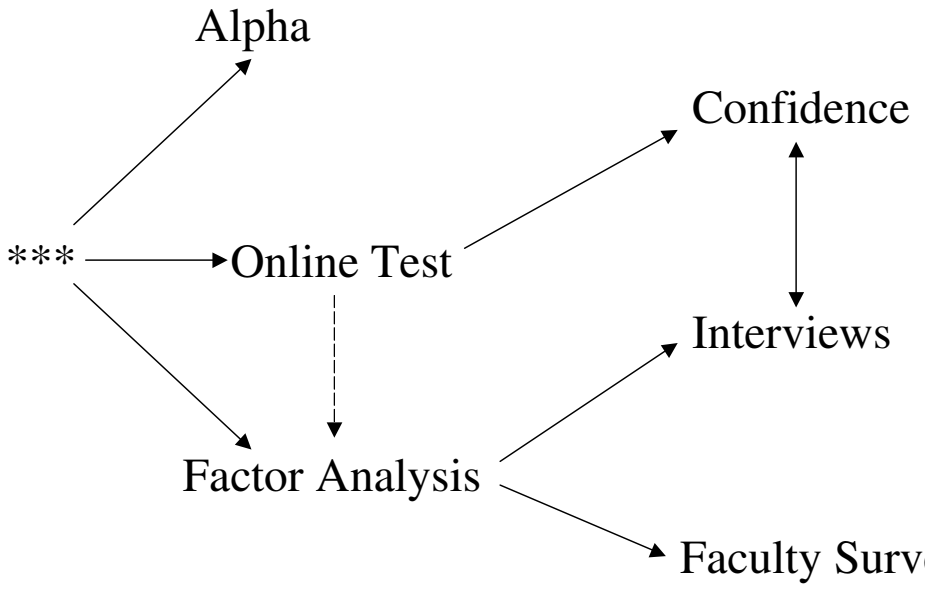


Figure 1: Development of Statistics Concept Inventory, after Spring 2004

1.1 Participation

In four years, the Statistics Concept Inventory has been administered to over 1500 students. Figure 2 offers a comparison of this progress to the Force Concept Inventory. At this stage, the SCI holds its own with the FCI. The SCI participation rate is extrapolated with the dashed line, for comparison with later FCI publications.

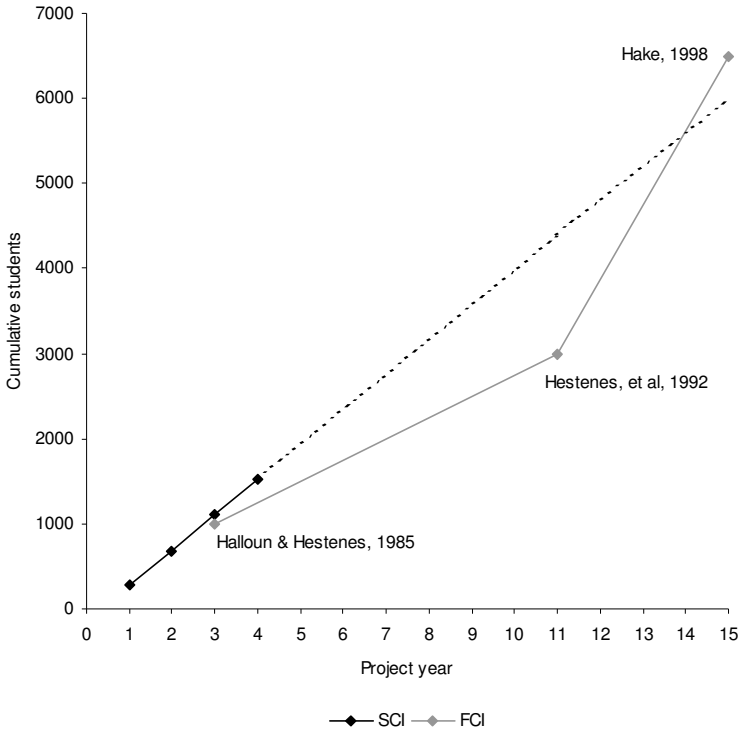


Figure 2: Sample size comparison between SCI and FCI

The length of time represented in this graphic is substantial: the first FCI publication was the result of three years' development, with seven and then six years between major publications. The development presented in Book One, for the first two years of the SCI, is comparable to the documented development of the FCI. The progress of the SCI relative to other engineering concept inventories (Chapter XI, Figures 3a and 3b) shows the SCI to be among the top instruments in this genre.

Sample size speaks of the acceptance of the instrument and the generalizability of the results. By these standards, the work of this dissertation certainly ranks the Statistics Concept Inventory among the most advanced engineering concept inventories.

Statistics is a broader field than is assessed by other concept inventories. Courses are typically taught in multiple departments at multiple levels. Tables 1a and 1b summarize the courses assessed for this dissertation. Seven outside universities (External #x) have participated along with ten courses in five departments and one summer research program at the University of Oklahoma.

Table 1a: Classes by Semester (Book One)

Course	**Level**	**Fa 02**	**Su 03**	**Fa 03**	**Sp 04**
Communications	Intro	√			
Engr	Intro	√	√	√	√
Math #1	Intro	√ (2)	√ (2)	√ (2)	√ (2)
DOE	Advanced	√		√	
REU	Varies		√		
External #1	Intro		√	√	
External #2	Intro			√ (2)	
External #3	Intro, 2-yr			√	

Table 1b: Classes by Semester (Later)

Course	**Level**	**Su 04**	**Fa 04**	**Sp 05**	**Su 05**	**Fa 05**	**Sp 06**
REU	Varies	√					
Engr	Intro		√	√	√	√	
Math #1	Intro		√	√ (2)	√ (2)	√ (2)	√ (2)
Math #2	Advanced		√			√	
External #1	Intro		√				
External #4	Unsure		√ (2)				
Quality	Advanced			√			
Psych #1	Intro			√	√	√	
Psych #2	Junior					√	
Psych #3	Jr / Sr					√	
Meteorology	Intro					√	
External #5	Intro					√ (3)	
External #6	Unsure					√	
DOE	Advanced					√	
External #7	Freshman					√	

1.2 Scores and Reliability

Figure 3 shows the aggregate post-test scores across semester (black: left axis). The results are consistently close to 50%. Lack of variability has been demonstrated between courses as well (Stone, 2006). Not all courses participate as a pre- and post-test. For those who do, the gains are small, typically less than 10%. Post-test data are more reliable (gray: right axis) than pre-tests, suggesting a lesser degree of guessing. Wider dissemination is needed to assess teaching styles able to produce the highest gains.

From Chapter XI, these results are in line with the limited findings presented from other concept inventories: scores around 50% (Chapter XI, Table 2) and reliability around 0.70 (Chapter XI, Table 3).

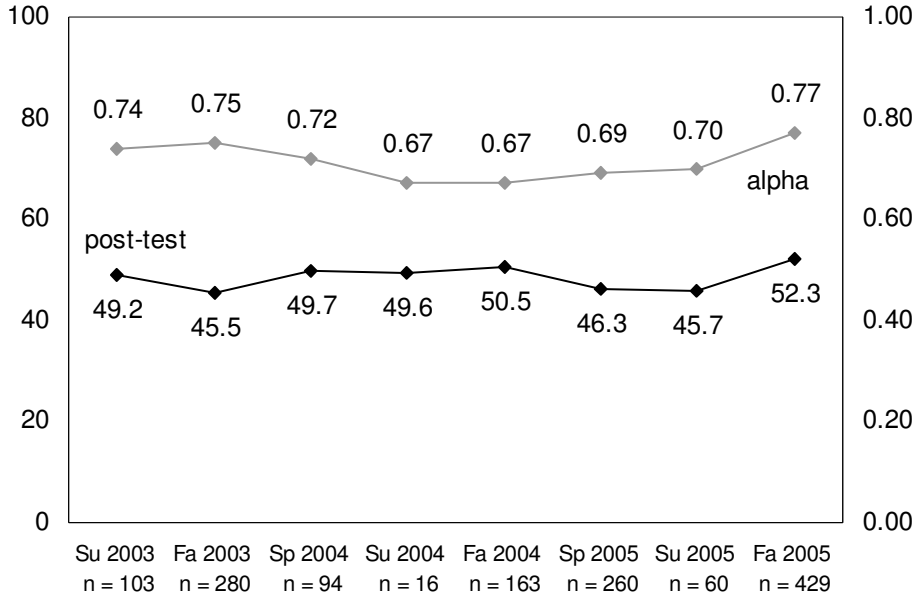


Figure 3: Post-test mean (left axis) and reliability (α, right axis) across semesters

2. Results

In Chapter XI (Figure 1), a flowchart of concept inventory development was proposed. The general process is approximately the same as that of the SCI. One important difference is that steps were often conducted simultaneously. For example, both content validity (e.g., interviews) and reliability (with other objective techniques) were consulted as a means to re-formulate objectives and construct new items. Further, distribution has been perpetual, as a means to gather data to be fed-back to the prior steps. The approximate phases of the SCI project are described in the following sections.

2.1 Early Results

Book One documented the development of the SCI in the first two project years. The focal point was item analysis, primarily the discriminatory index and alpha-if-deleted. Student focus group comments were incorporated to identify and improve items that caused confusion.

As related to later work, an important step was the factor analysis. These results established four sub-tests on the *a priori* domains of Probability, Descriptive Statistics, Inferential Statistics, and Graphical Interpretation. Due to the breadth of Statistics, adequate topic coverage mandated few redundant items, with the resulting model yielding a relatively low 30% explained variance. These four sub-test labels have been maintained throughout the dissertation.

2.2 Transitional Results

The remainder of the dissertation primarily utilized data from the Fall 2005 post-test, which was the largest one-semester administration to date (Figure 3). These data

were partially gathered using an online version of the SCI. The system was designed by the author using PHP interfacing with a mySQL database.

Two studies were conducted to assess the equivalency of the online SCI with the paper version. On the Fall 2005 post-test, with 429 participants (308 online, 121 paper). To allow a measure of control, non-engineering majors were removed from the equivalency analysis, leaving 194 and 116. Analysis of responses revealed nine items with significant differences between online and paper versions of the test, with five of these occurrences in the Inferential sub-test; the paper group had higher percent correct in every instance. Significant differences ($p < 0.05$) were found on the overall mean, as well as three sub-tests (excepting Probability). These differences were not large, however, with only Inferential having a difference greater than 10% (online 43%, paper 57%).

Nearly all of the paper students were at one university, in three sections of one course, including the section with the highest-ever SCI mean. A more controlled study was possible on the Spring 2006 pre-test: two sections of the same course, taught by the same professor, on the same date. To achieve such a level of control, the sample size was necessarily much smaller (14 online, 16 paper). Four items were found to have significant differences in response patterns, with another two manifesting significant differences in percent correct. Unlike the larger study, there was no tendency for one group to out-perform the other on these items. In terms of overall scores, the Inferential sub-test once again significantly favored paper version (+11%).

In a summative comparison, only the aforementioned Inferential sub-test and two items (one Descriptive, one Inferential) had significant differences across version for both studies. Given the large number of comparison points, some correspondence can be

expected by random error. At best, the online version cannot be said to be *vastly* different from the paper version. Given other research showing it is possible to construct an equivalent online version of a paper test (e.g., Cole, *et al.,* 2001 for the FCI), there is sufficient implication to conclude that the online SCI is a satisfactory method of gathering data, when one further takes into account the increased sample sizes which can be obtained.

Figure 4 shows the discrimination index of the items on the SCI (bars), along with the item difficulty (stars); the items have been ordered by discrimination index because question order is arbitrary. The shading refers to low (< 0.20: 5 items), medium (0.20 to 0.40: 13 items), and high (> 0.40: 20 items) discrimination; one item is not plotted because it attained negative discrimination. There are 24 items with discrimination above 0.30. Proportionally, these results are similar to the shorter Statics Concept Inventory (4; 10; 13) (Steif and Dantzler, 2005). However, Steif (2006, with Hansen) obtained much better results (only one item < 0.30) in a more recent publication.

These results are based on all students who took the SCI in either paper or online format. For the online format, students who did not participate in the Inferential sub-test are included; this makes little difference in the results, as the correlations between discrimination indices (r = 0.95) and difficulty (r = 0.99) are very high between datasets with and without these students.

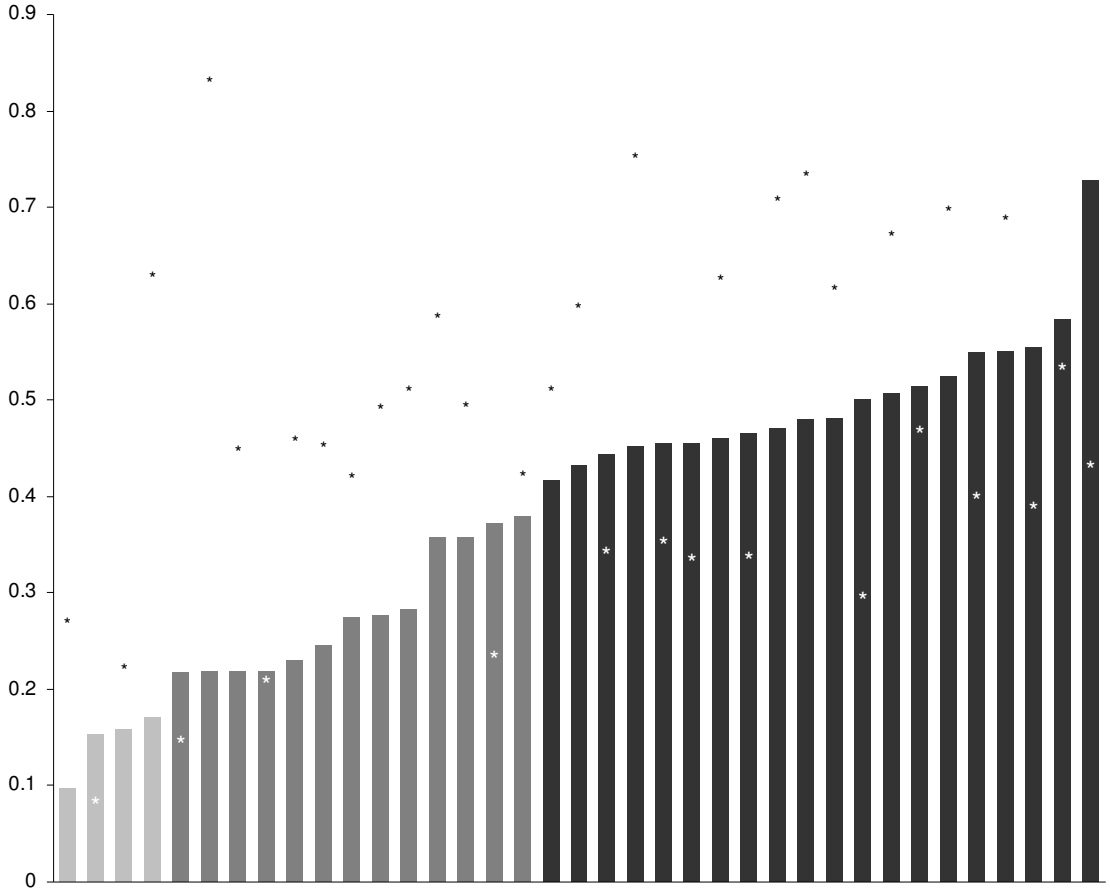


Figure 4: Item discrimination index (bars) and difficulty (stars) for Fall 2005 Post-test

An inherent assumption of concept inventories is that students hold some conceptual framework for the content. Interviews are the standard model for verifying these thought processes. Anyone who has conducted educational research appreciates the logistical difficulties of recruiting subjects for interviews. As an alternative, the Fall 2005 online SCI contained an additional wrinkle for each item: subjects were asked to rate the confidence in their answers on a simple 1 (low) to 4 (high) scale.

The results were fascinating in many ways. Figure 5 depicts the answer confidence vs. percent correct (ranks orders) for the 38 items. The "+" region represents items where confidence and correct ranks differ by at least 10 (over-confidence), while

the "-" region is analogous for under-confidence. The Probability sub-test yielded a disproportionate number of over-confident items, which is plausible given previous research showing that probabilistic reasoning skills develop as early as childhood, thus allowing informal heuristics to come into play. The under-confident region was more specifically dominated by items assessing correlation, which is often encountered in prior coursework but often not reached during an introductory statistics course. This may cause a vague understanding of the concept, although lack of formal instruction prevents solidification.

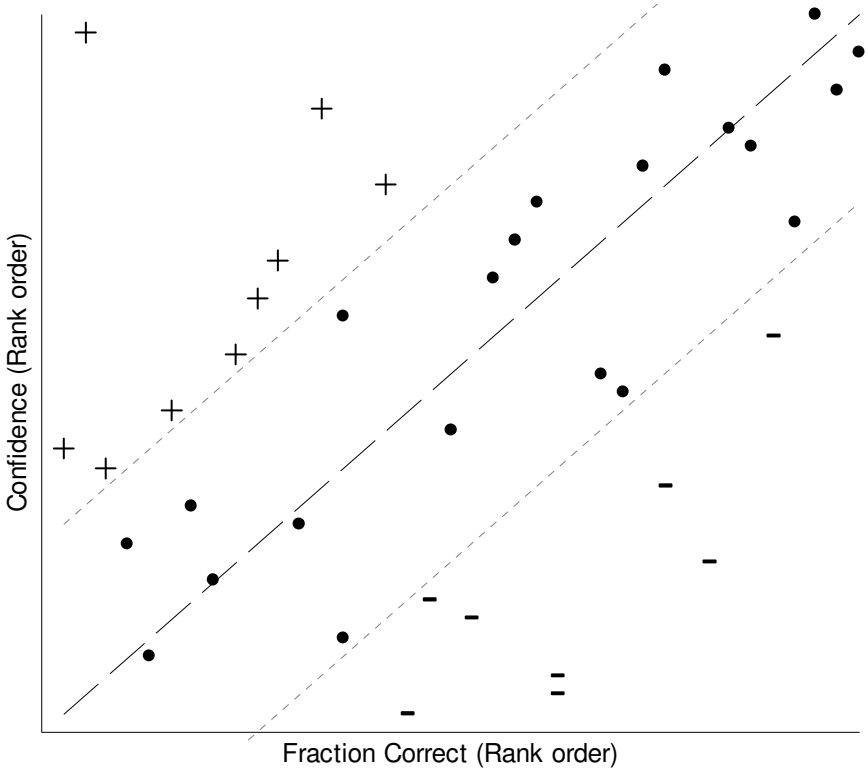


Figure 5: Confidence vs. fraction correct, rank orders

2.3 Final Results

As the development moved into its final stages, it became imperative to package the instrument in a publishable form. Specifically, there was concern that the length would be a detriment to widespread use of the SCI relative to other concept inventories, which tend to have 20 to 30 items.

To determine an optimal length, the dimensionality of the 38-item SCI was assessed in a factor analytic study. An exploratory analysis suggested a gross uni-dimensional structure, with some small clusters of similar items. Using these results as a guide, several confirmatory models were assessed. The uni-dimensional structure was retained as the most parsimonious among several alternatives. The model-fit was in the low end of what could be considered acceptable, although the results are quite similar to those on the Statics Concept Inventory (Steif and Dantzler, 2005). The results from the two instruments are compared in Table 2, reporting the same metrics for comparison.

Table 2: Comparison of Confirmatory Factory Analysis for Statistics and Statics

	Model Significance Test			Fit Assessment		
	χ^2	d.f.	$p > \chi^2$	GFI	CFI	RMSEA
Statistics	785	665	< 0.01	0.88	0.86	0.025
Statics	315	296	0.22	0.90	0.91	0.067

key: χ^2 = test statistic for overall model fit
d.f. = degrees of freedom for χ^2 $p > \chi^2$ = p-value for χ^2 statistic
GFI = Goodness-of-fit Index CFI = Comparative Fit Index
RMSEA = Root-mean-squared Error

The first group of metrics assesses the model's ability to predict inter-item correlations relative to those observed (H_0: model fits). By failing to reject the null hypothesis (p = 0.22), the Statics instrument appears to be a much-better-fitting model. Being a function of sample size, the significance test can be over-powered. Fit indices are reported as an alternate measure. By these standards, the SCI is only slightly below on GFI and CFI and actually better on RMSEA.

A nested model with similar items below the parent construct of Statistics was proposed as an intuitively appealing structure of the field as a whole (Figure 6). This intriguing possibility requires an expansion of the item pool and subsequent validation, which is beyond the scope of the current research (i.e., it would likely require several years and another NSF grant). The nesting was at a finer level than the earlier proposal of four sub-tests (Probability, Descriptive, Inferential, Graphical). Such a model may re-appear with additional items, but it currently serves best as a labeling scheme rather than a formal assessment tool; the four-specific-plus-general model is not wholly rejected.

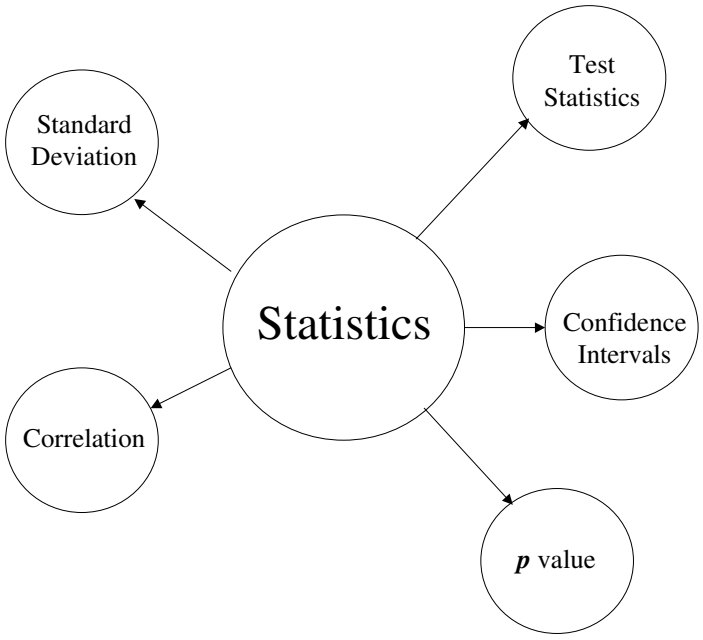


Figure 6: Proposed sub-domains for Statistics discipline

With a uni-dimensional model in hand, the task turned to shortening the SCI. Focusing first on reliability, the optimal test length was found to be 23 items, removing the 15 worst by alpha-if-deleted. This list of items was compared to the discrimination indices, and it was found that 13 of the 15 lowest-discriminating items were common with the 15-lowest by alpha-if-deleted. The communality estimates from the structural

model were compared as well, with 12 of 15 coinciding on all three lists. These 12 were deleted; a 13th item was included which missed the communality list but was the only item with a negative factor loading and also fell on the reliability and discrimination lists. Therefore, a final 25-item cut of the SCI was produced. From Figure 7, the retained items demonstrate strong discrimination (bars) and fall predominantly in a moderate range of difficulty (stars: between 0.3 and 0.8, save two items). When the discrimination indices are re-calculated based on subject scores to the 25 retained items, there is little change in either discrimination (median change –0.02, $r = 0.91$ between original and re-specified) or especially difficulty (< 0.01, $r = 0.98$).

A cross-validation was conducted on the Fall 2005 data, using 50% of the students as training and the remainder as testing. Using 1000 random replications, it was determined that removing 15 items could maintain the shortened-test reliability at nearly exactly the same level as the full-length SCI (median simulation $\alpha = 0.7655$, full test $\alpha = 0.7651$). Moreover, the 13 items chosen for deletion proved to be those most often falling in the worst 15 of the simulations.

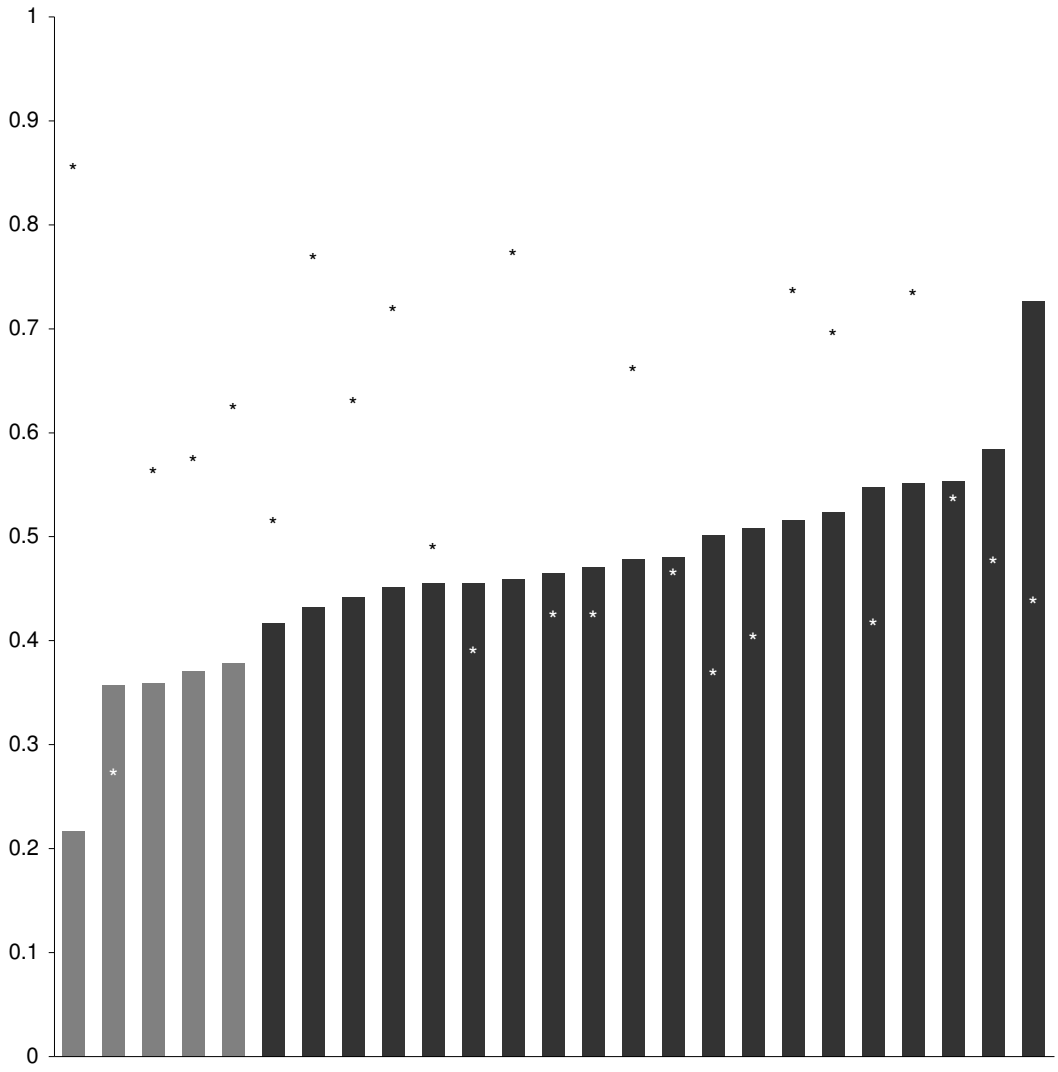


Figure 7: Item discrimination index (bars) and difficulty (stars) of 25 retained items

Finally, the content validity of the leaner 25-item SCI was assessed by comparing the topic coverage to two faculty topics surveys. The first was conducted in 2001 at the University of Oklahoma; the latter was conducted in 2006 by soliciting Industrial Engineering departments and other parties who had expressed interest in the SCI.

Table 3 (next page) compares the SCI item coverage to the 25 most-important items on the two faculty surveys. Coverage is very strong among items which rated

highly on both surveys. Some improvement could be made in the areas which fell in the top 25 on only one of the surveys.

Table 3: Coverage of Top 25 Important Topics, for 25-item SCI

Topic		New	Old
Top 25 New and Old			
Normal dist.	√	1	4
Measure of variability	√	2	1
Importance of data summary	√	2	2
Interpretation of prob.	√	4	7
Frequency dist and histograms	√	6	11
Covariance and correlation	√	6	9
The central limit theorem	√	8	13
Standardized normal		8	19
Random sampling	√	8	16
Simple linear regression		12	3
Correlation	√	12	16
Methods of displaying data	√	14	5
Independence	√	17	13
Sample space and events	√	18	16
Expected values	√	21	12
Multiplication and total prob rules	√	24	24
Summary	*14 of 16*		
Top 25 New only			
Inference on the mean of a pop.	√	5	34
Inference on a pop prop.		11	44
Type I (alpha) error		15	--
Assessing the adequacy of reg.		16	36
Sampling dist.		18	32
Inference on means of 2 norm pop.	√	20	53
Testing for a goodness of fit		22	28
Sample size determination		22	37
Type II (beta) error		25	--
Summary	*2 of 9*		
Top 25 Old only			
Continuous uniform dist.		48	6
Poisson dist.		30	8
Properties of the least squares		36	9
Time sequence plot		33	13
Use of the reg. for prediction		36	20
Binomial dist.	√	30	20
Conditional prob.	√	33	22
Properties of estimators		61	23
Confidence intervals for the reg.		41	24
Summary	*2 of 9*		

Student interviews were conducted to identify improvements and assess reasoning of the full SCI. Unfortunately, due to lack of time, only four students volunteered. However, their responses were thoughtful both in completing the test (online: minimum 40 minutes) and in discussions (similar times). Only minor changes were made to the items, and some suggestions for complementary items were presented (Chapter X).

2.4 Other Results

In a separate dissertation, Stone (2006) conducted item-response-theory (IRT) analyses of the SCI. Pedagogical implications were identified by fitting multiple-response item curves. Two examples are given as Figures 8a and 8b. The *Theta* scale is an estimate of subject ability, scaled similar to Z-scores where negative implies low ability with zero as an average subject. The *P(X=response)* is the probability of choosing a given multiple-choice option at a certain theta. The first figure indicates that low-ability subjects are most likely to choose option A, with D and B to a lesser extent. The values decrease as ability increases, while the correct response C increases to nearly 100% for high-ability students, with the incorrect responses correspondingly approaching zero.

The second figure reveals a possible misconception, as high-ability students indicate a preference for the incorrect B in favor of the correct C. Both figures have pedagogical implications, suggesting areas of focus for reaching low-ability students (Figure 8a: highlight why A is incorrect) and high-ability students (Figure 8b: distinguish between B and C). Only the Geosciences Concept Inventory has published IRT analyses (Libarkin and Anderson, 2005), and it was not at such a fine level.

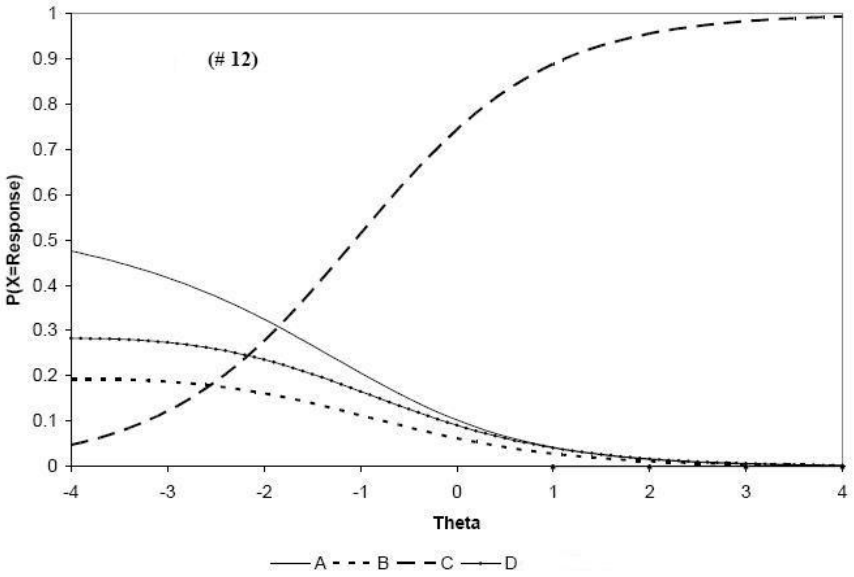


Figure 8a: Multiple-response IRT curves (question #12 on 38-item SCI)

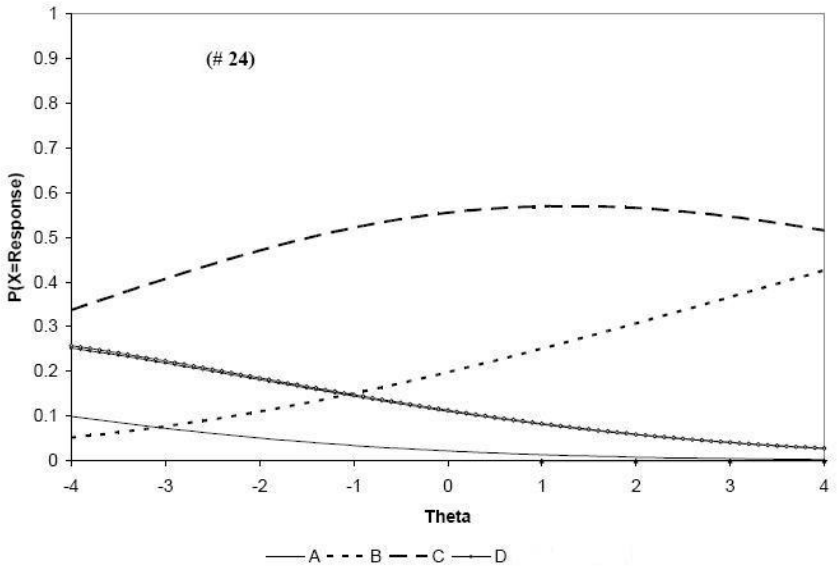


Figure 8b: Multiple-response IRT curves (question #24 on 38-item SCI)

3. Conclusions and Recommendations

The analysis presented in this dissertation, both in breadth and depth, is comparable and in most cases exceeds that presented for other concept inventories (Chapter XI). In tandem with the Stone dissertation, the SCI is likely the most-analyzed concept inventory. One uncertainty in this claim is the unpublished work in the 20+ years

of the FCI and pending analyses from well-developed engineering concept inventories such as Statics and Systems & Signals. To aid in establishing the SCI as a leader in this field, the work after Book One was presented in a form to foster publications without the usual steps of pulling together material from multiple chapters.

One unanswerable question, which affects the validity of the findings, is subject motivation. As the instrument gains wider acceptance, with interest more often initiated by the instructor rather than the research team, the validity should improve. Publication is a key to achieving this goal.

3.1 Directions for Future Research

Figure 9 depicts the recommendations for further work, taking end-point nodes of Figure 1 as expansion points. An holistic integration of all future steps can provide valuable feedback for improving statistics education.

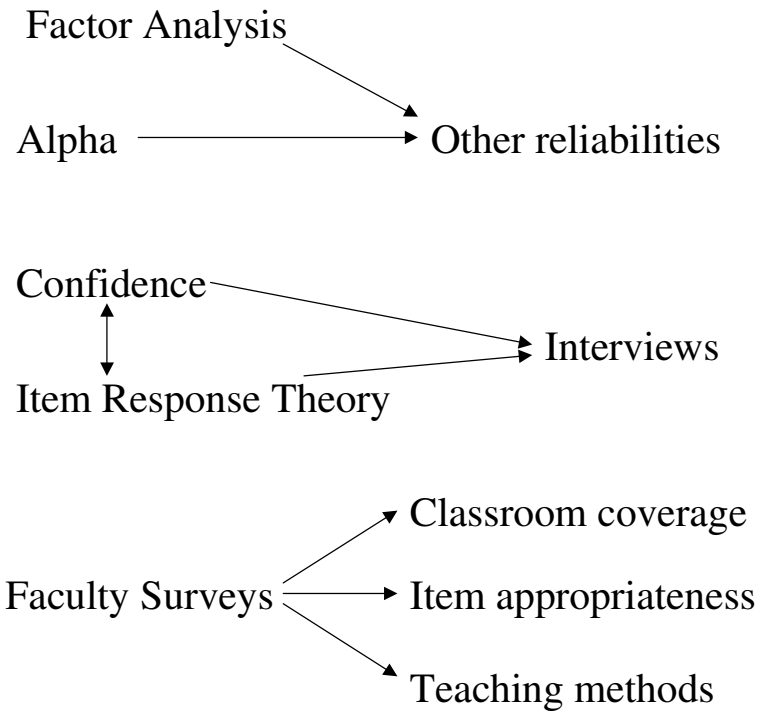


Figure 9: Future directions for the SCI

Reliability assessment can be enhanced using results of a factor analysis as related to dimensionality of the SCI; some work has been presented which makes these connections. An exposition similar to Chapter VI could introduce other concept inventory authors to these methods.

Content validity should continue to be assessed, through student interviews and faculty input. The effort exerted by the few subjects in Chapter X suggests that an appropriate audience of experienced statistics students can provide insightful feedback even in a shortened timeframe. Future pursuits should be perpetual and not handcuffed by the pre-test / post-test timeframe. These interviews should also take advantage of the confidence rating scale and associated literature review of Chapter VIII. The Item Response Theory of Stone (2006) can be integrated with the confidence and interviews; connections can be made between IRT and confidence before interviews are conducted.

Additional faculty input should integrate the topics survey with the instrument. In Chapter X, it was proposed to have faculty classify items nominally in terms of the topics survey. A second stage could include ordinal ratings of the appropriateness of the items to the given topic(s). A survey of classroom topic coverage could enhance the content validity of the instrument. Further faculty involvement of any form is likely to increase the "buying-in" factor necessary for dissemination. Inclusion of non-engineering faculty in these processes could foster inter-departmental communication, which is crucial given the scattering of statistics courses across departments.

The edited 25-item SCI requires dissemination to verify its psychometric properties. The contacts established throughout the prior four years should be contacted as a starting point for dissemination, to avoid missing semesters with the publication lag.

To assess the construct model (Figure 6), more items are needed in multiple topics. The findings indicate the four-area model of Book One could be appropriate, such as Descriptive (Figure 6, left side) and Inferential (right). With sufficient effort, this leads to multiple instruments to assess the whole of the Statistics field, although requiring many more years of development.

Finally, little has been presented about students' gain scores on the SCI. The gains are generally quite small, less than 10%, which is not atypical of concept inventories in their early stages. A combination of Stone's IRT with the confidence assessment from this work could have profound pedagogical implications.

4. Process Model Re-Visited

In Chapter I (Figure 1), this author's general model for test creation was presented; this is re-produced as Figure 10. In contrast to Beichner (1994) (Chapter XI, Figure 1), this model considers administration a first-step following item development. This is a more data-driven approach (e.g., item-level discrimination, test-level reliability), whereas other concept inventories generally focus on qualitative analyses (e.g., content validity: item-level, focus groups; test-level, faculty surveys) prior to broad student testing. The validation points in the figure are similarly weighted by quantitative methods, such as the factor analysis of Chapter IX. To avoid becoming a highly reliable but invalid measure, the next iteration of development should examine the shortened SCI (Chapter X) in comparison with an expanded item pool, whether through sub-tests or alternate forms. Dissemination will occur simultaneously.

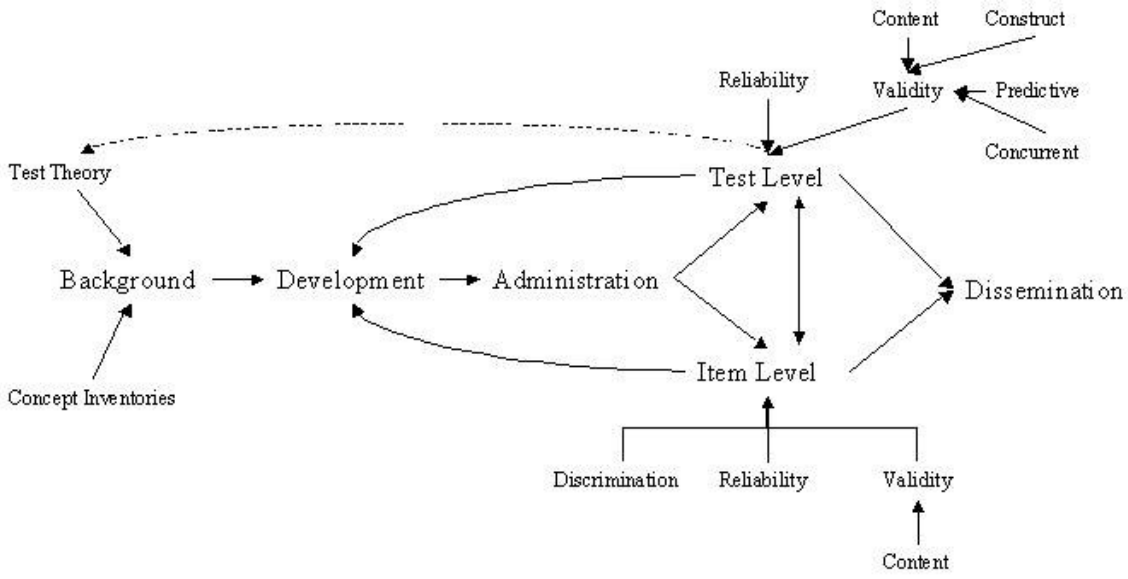


Figure 10: Test creation process

5. Final Word

As with the FCI and other educational research, the ultimate goal is to increase understanding and awareness of what, why, and how students learn. A re-centering along with qualitative-quantitative spectrum will help this happen. The work on the Statistics Concept Inventory thus far has perhaps raised more questions than it has answered. Most importantly, however, the foundation is solid for continuing to search for answers.

References

Cole, R.P., MacIsaac, D., and Cole, D.M. 2001. A Comparison of Paper-Based and Web-Based Testing. *Annual Meeting of the American Educational Research Association*. ERIC Document ED453224.

Hake, R. 1998. Interactive-engagement vs. traditional methods: A six-thousand student survey of mechanics test data for introductory physics courses. *American Journal of Physics*. 6 (1): 64-75.

Halloun, I. and D. Hestenes. 1985. The initial knowledge state of college physics students. *American Journal of Physics*. 53 (11): 1043-1055.

Hestenes, D., M. Wells, and G. Swackhamer. 1992. Force Concept Inventory. *The Physics Teacher*. 30 (March): 141-158.

Libarkin, J.C., and Anderson, S.W. 2005. Assessment of Learning in Entry-Level Geoscience Courses: Results from the Geoscience Concept Inventory. *Journal of Geoscience Education*. 53 (4, September): 394-401.

Steif, P.S., and Dantzler, J.A. 2005. A Statics Concept Inventory: Development and Psychometric Analysis. *Journal of Engineering Education*. 33: 363-371.

Steif, P.S., and Hansen, M. 2006. Comparisons Between Performance in a Statics Concept Inventory and Course Examinations. *International Journal of Engineering Education*. (in press) [http://www.me.cmu.edu/people/faculty/steif/educationalresearch.htm, accessed February 8, 2006]

Stone, A. 2006. A Psychometric Analysis of the Statistics Concept Inventory. Dissertation, University of Oklahoma.

Full References

Statistics Concept Inventory

Allen, K. 2004. Explaining Cronbach's Alpha. Available at http://coecs.ou.edu/sci under Publications. A paper is also being prepared with the findings. (below)

Allen, K., T.R. Rhoads, T.J. Murphy, R.A. Terry, and A. Stone. 2006. Reliability in Practice: An Engineer's Perspective. (tentatively accepted for Journal of Engineering Education, pending final revision/submission)

Allen, K., T.R. Rhoads, and R. Terry. 2006. Misconception or Misunderstanding? Assessing Student Confidence of Introductory Statistics Concepts. *Proceedings of the 36th ASEE/IEEE Frontiers in Education Conference*. Draft paper accepted pending revisions.

Allen, K., A. Stone, T.R. Rhoads, and T.J. Murphy. 2004. The Statistics Concept Inventory: Developing a Valid and Reliable Instrument. *Proceedings of the 2004 American Society for Engineering Education Annual Conference and Exposition*. Session 3230.

Allen, K., A.D. Stone, M. Cohenour, T.R. Rhoads, T.J. Murphy, and R.A. Terry. 2005. The Statistics Concept Inventory: A Tool for Measuring Learning in Introductory Statistics. Poster presented at *Joint Statistical Meetings*, Minneapolis.

Allen, K., and R. Terry. 2006. The Statistics Concept Inventory. Presentation at weekly graduate student seminar in Department of Psychology, The University of Oklahoma.

Stone, A. 2006. A Psychometric Analysis of the Statistics Concept Inventory. Dissertation, University of Oklahoma.

Stone, A., K. Allen, T.R. Rhoads, T.J. Murphy, R.L. Shehab, and C. Saha. 2003. The Statistics Concept Inventory: A Pilot Study. *Proceedings of the 33rd ASEE/IEEE Frontiers in Education Conference*. Session T3D-6.

Concept Inventories

Anderson, D.L., K.M. Fisher, and G.J. Norman. 2002. Development and Evaluation of the Conceptual Inventory of Natural Selection. *Journal of Research in Science Teaching*. 39 (10): 952-978.

Beichner, R.J. 1994. Testing student interpretation of kinematics graphs. *American Journal of Physics*. 62 (8): 750-755.

Cole, R.P., D. MacIsaac, and D.M. Cole. 2001. A Comparison of Paper-Based and Web-Based Testing. *Annual Meeting of the American Educational Research Association*. ERIC Document ED453224.

Engelhardt, P.V., and R.J. Beichner. 2004. Students' understanding of direct current resitive electrical circuits. *American Journal of Physics*. 72 (1): 98-115.

Evans, D.L., G.L. Gray, S. Krause, J. Martin, C. Midkiff, B.M. Notaros, M. Pavelich, D. Rancour, T.R. Rhoads, P. Steif, R.A. Streveler, and K. Wage. 2003. Progress On Concept Inventory Assessment Tools. *Proceedings of the 33rd ASEE/IEEE Frontiers in Education Conference*, Session T4G-8.

Gary, G.L., D. Evans, P. Cornwell, F. Costanzo, and B. Self. 2003. Toward a Nationwide Dynamics Concept Inventory Assessment Test. *Proceedings of the 2003 American Society for Engineering Education Annual Conference & Exposition*. Session 1168.

Hake, R. 1998. Interactive-engagement vs. traditional methods: A six-thousand student survey of mechanics test data for introductory physics courses. *American Journal of Physics*. 6 (1): 64-75.

Halloun, I. and D. Hestenes. 1985. The initial knowledge state of college physics students. *American Journal of Physics*. 53 (11): 1043-1055.

Heller, P., and D. Huffman. 1995. Interpreting the Force Concept Inventory: A Reply to Hestenes and Halloun. *The Physics Teacher*. 33 (November): 503-511.

Hestenes, D., and I. Halloun. 1995. Interpreting the Force Concept Inventory: A Response to March 1995 Critique by Huffman and Heller. *The Physics Teacher*. 33 (November): 502-506.

Hestenes, D., and M. Wells. 1992. A Mechanics Baseline Test. *The Physics Teacher*. 30 (March): 159-166.

Hestenes, D., M. Wells, and G. Swackhamer. 1992. Force Concept Inventory. *The Physics Teacher*. 30 (March): 141-158.

Huffman, D., and P. Heller. 1995. What Does the Force Concept Inventory Actually Measure?. *The Physics Teacher*. 33 (March): 138-143.

Jacobi, A., J. Martin, J. Mitchell, and T. Newell, 2003. A Concept Inventory for Heat Transfer. *Proceedings of the 33rd ASEE/IEEE Frontiers in Education Conference*. Session T3D-6.

Jordan, W., H. Cardenas, and C.B. O'Neal. 2005. Using a Materials Concept Inventory to Assess an Introductory Materials Class: Potentials and Problems. *Proceedings of the 2005 American Society for Engineering Education Annual Conference and Exposition*. Session 1064.

Kim, E., and Pak, S.-J. 2000. Students do not overcome conceptual difficulties after solving 1000 traditional problems. *American Journal of Physics*. 70 (7): 759-765.

Krause, S., J.C. Decker, and R. Griffin. 2003. Using a Materials Concept Inventory to Assess Conceptual Gain in Introductory Materials Engineering Courses. *Proceedings of the 33rd ASEE/IEEE Frontiers in Education Conference*. Session T3D-6.

Krause, S., J. Birk, R. Bauer, B. Jenkins, M.J. Pavelich. 2004a. Development, Testing, and Application of a Chemistry Concept Inventory. *Proceedings of the 34th ASEE/IEEE Frontiers in Education Conference*. Session T1G-1.

Krause, S., Tasooji, A., and Griffin, R. 2004b. Origins of Misconceptions in a Materials Concept Inventory From Student Focus Groups. *Proceedings of the 2004 American Society for Engineering Education Annual Conference & Exposition.*

Libarkin, J.C., and S.W. Anderson. 2005. Assessment of Learning in Entry-Level Geoscience Courses: Results from the Geoscience Concept Inventory. *Journal of Geoscience Education*. 53 (4, September): 394-401.

Maloney, D.P., T.L. O'Kuma, C.J. Hieggelke, and A.V. Heuvelen. 2000. Surveying students' conceptual knowledge of electricity and magnetism. *American Journal of Physics*. 69 (7) : S12 – S23.

Martin, J., J. Mitchell, and T. Newell. 2003. Development of a Concept Inventory for Fluid Mechanics. *Proceedings of the 33rd ASEE/IEEE Frontiers in Education Conference*.Session T3D-6.

Martin, J. K., J. Mitchell, and T. Newell. 2004. Analysis of Reliability of the Fluid Mechanics Concept Inventory. *Proceedings of the 34th ASEE/IEEE Frontiers in Education Conference*. Session T1A-1.

Michel, H., J. Jackson, P. Fortier, and H.Liu. Computer Engineering Concept Inventory. http://www.foundationcoalition.org/home/keycomponents/concept/computer.html, Accessed April 17, 2006.

Midkiff, K.C., T.A. Litzinger, and D.L. Evans. 2001. Development of Engineering Thermodynamics Concept Inventory Instruments. *Proceedings of the 31st ASEE/IEEE Frontiers in Education Conference*. Session F2A-3.

Morgan, J., and J. Richardson. Strength of Materials (SOM) Concept Inventory. http://www.foundationcoalition.org/home/keycomponents/concept/strength.html, Available in either pdf or ppt, Accessed September 8, 2005.

Notaros, B. Electromagnetics Concept Inventory. http://www.foundationcoalition.org/home/keycomponents/concept/electromagnetics.html, Accessed April 17, 2006.

Olds, B.M., R. Streveler, R.L. Miller, and M.A. Nelson. 2004. Preliminary Results from the Development of a Concept Inventory in Thermal and Transport Science. *Proceedings of the 2004 American Society for Engineering Education Annual Conference & Exposition*. Session 3230.

Pavelich, M., B. Jenkins, J. Birk, R. Bauer, and S. Krause. 2004. Development of a Chemistry Concept Inventory for Use in Chemistry, Materials and other Engineering Courses. (Draft). *Proceedings of the 2004 American Society for Engineering Education Annual Conference & Exposition*.

Richardson, J., P. Steif, J. Morgan, and J. Dantzler. 2003. Development Of A Concept Inventory For Strength Of Materials. *Proceedings of 33rd ASEE/IEEE Frontiers in Education Conference*. Session T3D-29.

Rhoads, T.R., and R.J. Roedel. 1999. The Wave Concept Inventory - A Cognitive Instrument Based on Bloom's Taxonomy. *Proceedings of the 29th ASEE/IEEE Frontiers in Education Conference*. Session 13c1. Paper 13c1-14.

Roedel, R.J., S. El-Ghazaly, T.R. Rhoads, and E. El-Sharawy. 1998. The Wave Concepts Inventory – An Assessment Tool for Courses in Electromagnetic Engineering. *Frontiers In Education Conference Proceedings 1998*.

Simoni, M.F., M.E. Herniter, and B.A. Ferguson. 2004. Concepts to Questions: Creating an Electronics Concept Inventory Exam. *Proceedings of the 2004 American Society for Engineering Education Annual Conference & Exposition*. Session 1793.

Steif, P. 2003. Comparison Between Performance on a Concept Inventory and Solving Multifaceted Problems. *Proceedings of the 33rd ASEE/IEEE Frontiers in Education Conference*. Session T3D-6.

Steif, P. 2004. Initial Data From A Statics Concept Inventory. *Proceedings of the 2004 American Society for Engineering Education Annual Conference & Exposition*. Session 1368.

Steif, P.S., and J.A. Dantzler. 2005. A Statics Concept Inventory: Development and Psychometric Analysis. *Journal of Engineering Education*. 33: 363-371.

Steif, P.S., and M. Hansen. 2006. Comparisons Between Performance in a Statics Concept Inventory and Course Examinations. *International Journal of Engineering Education*. (in press) [http://www.me.cmu.edu/people/faculty/steif/educationalresearch.htm, accessed February 8, 2006]

Thornton, R.K., and D.R. Sokoloff. 1998. Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation and the Evaluation of Active Learning Laboratory. *American Journal of Physics*. 66(4): 338-352.

Voska, K. W., and H.W. Heikkinen, 2000. Identification and Analysis of Student Conceptions Used to Solve Chemical Equilibrium Problems. *Journal of Research in Science Teaching*. 37: 160-176.

Wage, K.E., J.R. Buck, T.B. Welch, and C.H.G. Wright. 2002. Testing and Validation of the Signals and Systems Concept Inventory. *Proceedings of the 2nd IEEE Signal Processing Education Workshop*. Session 4.6: pp. 1-6.

Wage, K.E., J.R. Buck, C.H.G. Wright, and T.B. Welch. 2005. The Signals and Systems Concept Inventory. *IEEE Transactions on Education*. 48 (3): 448-461.

Albert, J.H. 2003. College Students' Conceptions of Probability. *The American Statistician*. 57 (1): 37-45.

Austin, J.D. 1974. An Experimental Study of the Effects of Three Instructional Methods in Basic Probability and Statistics. *Journal for Research in Mathematics Education*. 5 (3, May): 146-154.

Baloğlu, M. 2003. Individual differences in statistics anxiety among college students. *Personality and Individual Differences*. 34 (5) : 855-865.

Bar-Hillel, M. 1974. Similarity and Probability. *Organizational Behavior and Human Performance*. 11: 277-282.

Fong, G.T., Krantz, D.H., and Nisbett, R.E. 1986. The Effects of Statistical Training on Thinking about Everyday Problems. *Cognitive Psychology*. 18: 253-292.

Fong, G.T., and Nisbett, R.E. 1991. Immediate and Delayed Transfer of Training Effects in Statistical Reasoning. *Journal of Experimental Psychology: General*. 120 (1): 34-45.

Garfield, J. and A. Ahlgren. 1988. Difficulties in Learning Basic Concepts in Probability and Statistics: Implications for Research. *Journal for Research in Mathematics Education*. 19 (1): 44-63.

Hirsch, L.S., and A.M O'Donnell, 2001. Representativeness in statistical reasoning: Identifying and assessing misconceptions. *Journal of Statistics Education, 9*(2).

Kahneman, D. and A. Tversky. 1972. Subjective Probability: A Judgment of Representativeness. *Cognitive Psychology*. 3 (3): 430-453.

Kahneman, D., P. Slovic, and A. Tversky, eds. 1982. Judgement under uncertainty: Heuristics and biases. Cambridge University Press: Cambridge.

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

Konold, C., 1995. Issues in Assessing Conceptual Understanding in Probability and Statistics. *Journal of Statistics Education*. 3 (1), online.

Konold, C., A. Pollatsek, A. Well, J. Lohmeier, and A. Lipson. 1993. Inconsistencies in Students' Reasoning About Probability. *Journal for Research in Mathematics Education*. 24 (5): 392-414.

Mevarech, Z.R. 1983. A Deep Structure Model of Students' Statistical Misconceptions. *Educational Studies in Mathematics*. 14: 415-429.

Murtonen, M., and E. Lehtinen. 2003. Difficulties Experienced by Education and Sociology Students in Quantitative Methods Courses. *Studies in Higher Education*. 28 (2): 171-185.

Piaget, J., and B. Inhelder. 1951. La genèse de l'idée de hazard chez l'enfant (in English: The Origin of the Idea of Chance in Children). translated by Leake, Burrell, Fishbein and published by W.W. Norton & Company, Inc.: New York, 1975.

Ploger, D., and M. Wilson. 1991. Statistical Reasoning: What Is the Role of Inferential Rule Training? Comment on Fong and Nisbett. *Journal of Experimental Psychology: General*. 120 (2): 213-214.

Pollatsek, A., S. Lima, and A.D. Well. 1981. Concept or Computation: Students' Understanding of the Mean. *Educational Studies in Mathematics*. 12: 191-204.

Pollatsek, A., C.E. Konold, A.D. Well, and S.D. Lima. 1984. Beliefs underlying random sampling. *Memory & Cognition*. 12 (4): 395-401.

Schacht, S., and B.J. Stewart. 1990. What's Funny about Statistics? A Technique for Reducing Student Anxiety. *Teaching Sociology*. 18 (1, Jan.): 52-56.

Simon, J., D. Atkinson, and C. Shevokas. 1976. Probability and Statistics: Experimental Results of a radically different teaching methods. *American Mathematical Monthly*. 83: 733-739.

Rhoads, T.R., and N.F. Hubele. 2000. Student Attitudes Toward Statistics Before and After a Computer-Integrated Introductory Statistics Course. *IEEE Transactions on Education*. 43 (2, June): 182-187.

Tversky, A., and D. Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science*. 185: 1124-1131. (also available as Chapter 1 in same-titled book; eds. Kahneman, Slovic, Tversky. 1982.)

Tversky, A., and D. Kahneman. 1982. Judgments of and by representativeness. Chapter 1 in Judgment under uncertainty; eds. Kahneman, Slovic, Tversky.

Watts, D.G. 1991. Why Is Introductory Statistics Difficult to Learn? And What Can We Do to Make It Easier? *The American Statistician*. 45 (4, Nov.): 290-291.

Well, A.D., A. Pollatsek, and S.J. Boyce. 1990. Understanding the Effects of Sample Size on the Variability of the Mean. Organizational Behavior and Human Decision Processes. *47: 289-312.*

Armor, D. 1974. "Theta reliability and factor scaling." In Sociological Methodology 1970, H. Costner, Ed., pp.17-50. Jossey-Bass: San Francisco.

Ausubel, D. 1968. Educational Psychology: A Cognitive View. Holt, Reinhart, and Winston: New York.

Brown, F.G. 1983. *Principles of Educational and Psychological Testing*. Holt, Reinhart, and Winston: New York.

Brown, J.B. 1975. The Number of Alternatives for Optimum Test Reliability. *Journal of Educational Measurement*. 12(2): 109-113.

Clariana, R., and P. Wallace. 2002. Paper-based versus computer-based assessment: key factors associated with the test mode effect. *British Journal of Educational Technology*. 33 (5): 593-602.

Cortina, J.M. 1993. What is Coefficient Alpha? An Examination of Theory and Applications. *Journal of Applied Psychology*. 78 (1): 98-104.

Costin, F. 1970. The Optimal Number of Alternatives in Multiple-Choice Achievement Tests: Some Empirical Evidence for a Mathematical Proof. *Educational and Psychological Measurement*. 30: 353-358.

Costin, F. 1972. Three-choice Versues Four-choice Items: Implications for Reliability and Validity of Objective Achievement Tests. *Educational and Psychological Measurement*. 32: 1035-1038.

Cronbach, L.J. 1943. On Estimates of Test Reliability. *Journal of Educational Psychology*. 34: 485-494.

Cronbach, L.J. 1946. A Case Study of the Split-Half Reliability Coefficient. *Journal of Educational Psychology*. 37: 473-480.

Cronbach, L.J. 1947. Test 'Reliability': Its Meaning and Determination. *Psychometrika*. 12 (1): 1-16.

Cronbach, L.J. 1951. Coefficient Alpha and the Internal Structure of Tests. *Psychometrika*. 16 (3): 297-334.

Ebel, R. 1954. Procedures for the Analysis of Classroom Tests. *Educational & Psychological Measurement*. 14: 352-364.

Ebel, R.L. 1969. Expected Reliability as a Function of Choices Per Item. *Educational and Psychological Measurement*. 29: 565-570.

Feldt, L. S. 1969. A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. *Psychometrika*. 34: 363-373.

Flanagan, J.C. 1937. A proposed procedure for increasing the efficiency of objective tests. *Journal of Educational Psychology*. 28: 17-21.

Goldberg, A.J., and J.J. Pedulla. 2002. Performance Differences According to Test Mode and Computer Familiarity on a Practice Graduate Record Exam. *Educational and Psychological Measurement*. 62 (6, December): 1053-1067.

Green, V., and E. Carmines. 1979. "Assessing the Reliability of Linear Composites." In Sociological Methodology 1980, K.F. Schuessler, Ed., pp.160-75. Jossey-Bass: San Francisco.

Guttman, L. 1945. A Basis for Analyzing Test-Retest Reliability. *Psychometrika*. 10 (4): 255-282.

Hambleton, R.K. 1980. Test Score Validity and Standard-Setting Methods. Chapter 4 (pp. 80-123) in Criterion-Referenced Measurement. ed. R.A. Berk. The Johns Hopkins University Press: Baltimore.

Hopkins, K.D., J.C. Stanley, and B.R. Hopkins. 1990. *Educational and Psychological Measurement and Evaluation, 7th Edition*. Prentice Hall: Englewood Cliffs, NJ.

Kelley, T. 1939. The Selection of Upper and Lower Groups for the Validation of Test Items. *Journal of Educational Psychology*. 30: 17-24.

Kline, P. 1986. A Handbook of Test Construction. Methuen & Co. Ltd: New York.

Kline, P. 1993. The Handbook of Psychological Testing. Routledge: London and New York.

Kuder, G.F., and M.W. Richardson. 1937. The Theory of the Estimation of Test Reliability. *Psychometrika*. 2 (3): 151-160.

Lord, F.M. 1977. Optimal Number of Choices per Item – A Comparison of Four Approaches. *Journal of Educational Measurement*. 14(1): 33-38.

Mead, A.D., and F. Drasgow. 1993. Equivalence of Computerized and Paper-and-Pencil Cognitive Ability Tests: A Meta-Analysis. *Psychological Bulletin*. 114 (3): 449-458.

Neuman, G., and R. Baydoun. 1998. Computerization of Paper-and-Pencil Tests: When Are They Equivalent? *Applied Psychological Measurement*. 22 (1, March): 71-83.

Novick, M.R., and C. Lewis. Coefficient Alpha and the Reliability of Composite Measurements. *Psychometrika*. 32 (1): 1-13.

Nunnally, J. 1978. Psychometric Theory. McGraw-Hill: New York.

Oosterhof, A. 1996. Developing and Using Classroom Assessment. Merrill / Prentice Hall: Englewood Cliffs, New Jersey.

Ramos, R.A., and Stern, J. 1973. Item Behavior Associated with Changes in the Number of Alternatives in Multiple Choice Items. *Journal of Educational Measurement*. 10(4): 305-310.

Rebello, N.S., and Zollman, D.A. 2004. The effect of distracters on student performance on the force concept inventory. *American Journal of Physics*. 72 (1, January): 116-125.

Rogers, W.T., and Harley, D. 1999. An Empirical Comparison of Three- and Four-Choice Items and Tests: Susceptibility to Testwisness and Internal Consistency Reliability. *Educational and Psychological Measurement*. 59(2): 234-247.

Rulon, P.J. 1939. A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review*. 9: 99-103.

Spray, J.A., T.A. Ackerman, M.D. Recakse, and J.E. Carlson. 1989. Effect of the Medium of Item Presentation on Examinee Performance and Item Characteristics. *Journal of Educational Measurement*. 26 (3, Autumn): 261-271.

Streiner, D. L. 2003. Starting at the Beginning: An Introduction to Coefficient Alpha and Internal Consistency. *Journal of Personality Assessment*. 80 (1): 99-103.

Thompson, B., and X. Fan, 2003. Confidence Intervals About Score Reliability Coefficients. Chapter 5 in Score Reliability. B. Thompson editor. Sage Publications: Thousand Oaks, CA.

Thorndike, R.L. 1982. Applied Psychometrics. Houghton Mifflin: Boston.

Factor Analysis

Harman, H.H. 1976. Modern Factor Analysis. 3rd ed. The University of Chicago Press: Chicago.

Hoogland, J.J., and A. Boomsma. 1998. "Robustness Studies in Covariance Structural Modeling." *Sociological Methods & Research*. 26 (3 / February): 329-367.

Kaplan, D. 2000. Structural Equation Modeling. Sage Publications: Thousand Oaks, CA.

Loehlin, J.C. 2004. Latent Variable Models. 4th ed. Lawrence Erlbaum Associates: Mahwah, NJ.

Kim, J.O., N. Nie, and S. Verba. 1977. "A note on factor analyzing dichotomous variables: the case of political participation." *Political Methodology*. 4: 39-62.

Kim, J.O., and C.W. Mueller. 1978. Factor Analysis: statistical methods and practical issues. Sage Publications: Beverly Hills.

Marsh, H.W., J.R. Balla, and K.T. Hau. 1996. "An Evaluation of Incremental Fit Indices: A Clarification of Mathematical and Empirical Properties." In Advanced Structural Equation Modeling, G.A. Marcoulides and R.E. Schumacker, Eds., pp. 315-353.

Mulaik, S.A., L.R. James, J. Van Alstine, N. Bennett, S. Lind, and C.D. Stilwell. 1989. "Evaluation of goodness-of-fit indices for structural equation models." *Psychological Bulletin*. 105: 430-445.

Rummel, R.J. 1970. Applied Factor Analysis. Northwestern University Press: Evanston.

Statistics Textbooks

Agresti, A., and B. Finlay. 1997. Statistical Methods for the Social Sciences. 3rd ed. Prentice-Hall, Inc.: Upper Saddle River, NJ.

Gravetter, F.J., and L.B. Wallnau. 1988. Statistics for the Behavioral Sciences. 2nd ed. West Publishing Company: St. Paul.

Johnson, R.A. 1994. Miller & Freund's Probability & Statistics For Engineers. 5th ed. Prentice-Hall, Inc.: Englewood Cliffs, NJ.

Johnson, R.A., and D.W. Wichern. 2002. Applied Multivariate Statistical Analysis. 5th ed. Prentice-Hall: Upper Saddle River, NJ.

Mendenhall, W., and T. Sincich. 1995. Statistics for Engineering and the Sciences. 4th ed. Prentice-Hall, Inc.: Englewood Cliffs, NJ.

Montgomery, D. and G. Runger. 1994. Applied Statistics and Probability for Engineers. Wiley: New York.

Moore, D. 1997. The Active Practice of Statistics. W. H. Freeman and Company: New York.

Neter, J., M.H. Kutner, C.J. Nachtsheim, and W. Wasserman. 1996. Applied Linear Statistical Models. 4th Edition. McGraw-Hill: Boston.

Others

Bureau of Labor Statistics. 2006. "Occupational Outlook Handbook, 2006-07 Edition." [http://www.bls.gov/oco/], Accessed March 30, 2006.

College Board. 2003. Course Description: Statistics. Retrieved December 18, 2003. http://www.collegeboard.com/prod_downloads/ap/students/statistics/ap03_statistics.pdf.

Crouch, C.H., and E. Mazur. 2001. Peer Instruction: Ten years of experience and results. *American Journal of Physics*. 69 (9): 970-977.

Engineering Accreditation Commission. 2003. Criteria For Accrediting Engineering Programs. 2004-2005 Criteria. http://www.abet.org/criteria_eac.html.

Engineering Accreditation Commission. 2006. Criteria For Accrediting Engineering Programs, 2006-2007 Criteria. http://www.abet.org/forms.shtml, Accessed March 16, 2006.

Enzmann, D. 2001. "TetCorr 2.1", Accessed March 31, 2006. [http://www2.jura.uni-hamburg.de/instkrim/kriminologie/Mitarbeiter/Enzmann/Software/Enzmann_Software.html]

Gibb, B. 1964. *Test-Wiseness as Secondary Cue Response*. Dissertation, Stanford University.

Patton, M.Q. 1990. Qualitative Evaluation and Research Methods. 2nd Ed. Sage Publications: Newbury Park, CA.

Schau, C., T.L. Dauphinee, A. Del Vecchio and J.J. Stevens. Surveys of Attitudes Toward Statistics. [http://www.unm.edu/~cshau/downloadsats.pdf, accessed October 2, 2002]

Zeller, R.A., and E.G. Carmines. 1980. Measurement in the social sciences. Cambridge University Press: London.

Do I contradict myself?
Very well then, I contradict myself.
I am large, I contain multitudes.
-- Walt Whitman, *Leaves of Grass* (“Song of Myself”)

finis…