UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

A PSYCHOMETRIC ANALYSIS OF THE STATISTICS CONCEPT INVENTORY

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

By

ANDREA STONE
Norman, Oklahoma
2006

Downloaded from SHAREOK. Persistent Handle URL: https://hdl.handle.net/11244/1013

UMI Number: 3208004



UMI Microform 3208004

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

A PSYCHOMETRIC ANALYSIS OF THE STATISTICS CONCEPT INVENTORY

A Dissertation APPROVED FOR THE
DEPARTMENT OF MATHEMATICS

BY

Teri Jo Murphy, Ph.D.

Marilyn Breen, Ph.D.

Kevin Grasse, Ph.D.

Curtis McKnight, Ph.D.

Robert Terry, Ph.D.

© Copyright by ANDREA STONE 2006
All Rights Reserved.

ACKNOWLEDGEMENTS

The researcher wishes to acknowledge the support provided by a grant from the National Science Foundation (DUE-0206977). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

TABLE OF CONTENTS

Chapter 1: Literature 1
1.1 Force Concept Inventory 5
1.2 Other Concept Inventories-Related Efforts 13
1.2.1 Determining and Interpreting Resistive Electric Circuit Concepts Test (DIRECT) 14
1.2.2 Test of Understanding Graphs in Kinematics (TUG-K) 15
1.2.3 Statics Concept Inventory 17
1.2.4 Force and Motion Conceptual Evaluation (FMCE) 19
1.2.5 Signals and Systems Concept Inventory (SSCI) 19
1.2.6 Conceptual Survey of Electricity and Magnetism (CSEM) 22
1.2.7 Thermal and Transport Science Concept Inventory 25
1.2.8 Wave Concept Inventory 26
1.2.9 Dynamics Concept Inventory 27
1.2.10 Fluid Mechanics Concept Inventory (FMCI) 29
1.2.11 Chemistry Concept Inventory 29
1.2.12 Heat Transfer Concept Inventory (HTCI) 31
1.2.13 Materials Concept Inventory (MCI) 32
1.2.14 Other Concept Inventories 33
1.2.15 Common Themes 37
1.3 Statistics Education Research 39
1.3.1 Probabilistic Thinking/General Reasoning Frameworks 39
1.3.2 Averages/Measures of Central Tendency 46
1.3.3 Sampling Distributions 51
1.4 Other Instruments for Statistics Assessment 54
1.5 Test Theory Background 60
1.5.1 Classical Test Theory Model 61
1.5.2 Reliability 62
1.5.3 Factor Analytic Model 66
1.5.4 Item Response Theory 68
1.5.5 Validity 75
1.5.6 Summary 78
Chapter 2: A Classical Test Theory Perspective 79
2.1 Development of the Statistics Concept Inventory 79
2.2 Participants and Data Collection 89
2.3 Results 94
2.3.1 Posttest Scores 94
2.3.2 Gains 98
2.3.3 Correlation with Final Course Grades 99
2.3.4 Coefficient Alpha 102
Chapter 3: An Item Analysis of the Statistics Concepts Inventory 104
3.1 Item Analysis Tools 104
3.2 Statistics Concept Inventory: Annotated Version 110
Chapter 4 An Item Response Theory Perspective 163
4.1 The Data Set 163

4.2 The Two Parameter Logistic Model (2PL) 166
4.2.1 Question Comparisons 168
4.2.2 The Test as a Whole 178
4.3 The Nominal Response Model 180
Chapter 5 Discussion and Directions for Future Research 192
5.1 Scoring 192
5.1.1 Obtaining Ability Estimates 193
5.1.2 Gains 196
5.2 Model-Data Fit 198
5.2.1 Unidimensionality Assumption 199
5.2.2 Minimal Guessing Assumption 203
5.2.3 Non-speeded Assumption 204
5.2.4 Model Features and Behavior 205
5.3 Further analysis 214
5.3.1 Factor Analysis 214
5.3.2 Investigation of Test Bias 215
5.3.3 Confidence Analysis 217
5.4 Future Revisions 218
5.5 Reliability and Validity 220
5.6 Conclusions 224
References 225
Appendix A 236
Appendix B 256

LIST OF TABLES

Table 1-1: A portion of the item classification for the Force Concept Inventory (Hestenes et al., 1992). 6
Table 1-2: Summary of FCI pre and post outcome data, based on Halloun and Hestenes (1985b). 8
Table 1-3: A portion of the taxonomy of misconceptions developed from the Force Concept Inventory (Hestenes et al., 1992). 9
Table 1-4: Summary of results from the Determining and Interpreting Resistive Electric Circuit Concepts Test (DIRECT) (Engelhardt & Beichner, 2004). 14
Table 1-5: A comparison of the percentage of students in each quartile on the Statics Concept Inventory and the final exam score received at one site. 18
Table 1-6: Comparison of normalized gains by course type for the SSCI and the FCI (Wage et al., 2005). 21
Table 1-7: Baseline pre and post test data for CSEM (Maloney et al., 2001). 24
Table 1-8: WCI score results for each class (Roedel et al., 1998). 26
Table 1-9: Dynamics Concept Inventory Results 28
Table 1-10: Chemistry Concept Inventory results summary (Krause et al., 2004). 30
Table 1-11: Concept Matrix for the Heat Transfer Concept Inventory 32
Table 1-12: Summary of Concept Inventories. Those being developed as part of the the Foundation Coalition are marked with a * 34
Table 1-13: Frequency of common probabilistic misconceptions found by Fischbein and Schnarch (1997) 44
Table 1-14: Meta-analysis results from Sedlmeier and Gigerenzer (1997) comparing two forms of questions concerning sample size. 46
Table 1-15: Levels of understanding of variation (Watson et al., 2003). 60
Table 2-1: Results of the Instructor Survey of Statistics Topics, ordered by average ranking. The median ranking was 2.62. Topics were ranked from 1 (not at all important) to 4 (very important). 82
Table 2-2: Item classification for the SCI. 87
Table 2-3: Taxonomy of errors and misconceptions identified by the SCI. 88
Table 2-4: Participating courses and their prerequisites. 90
Table 2-5: SCI Summary Statistics by Course and by Semester 95
Table 2-6: ANOVA summary table. 97
Table 2-7: Gains and normalized gains for classes in which both pre- and posttest data were available. 98
Table 2-8: Correlations of SCI posttest score with final course grades. 99
Table 2-9: Coefficient alpha for each semester and for individual courses. 103
Table 3-1: Alpha-if-item-deleted values and rankings from spring 2005 post test data. 108
Table 3-2: Annotation for SCI question 1 113
Table 3-3: Annotation for SCI question 2 114
Table 3-4: Annotation for SCI question 3 116
Table 3-5: Annotation for SCI question 4 117
Table 3-6: Annotation for SCI question 5 118
Table 3-7: Annotation for SCI question 6 120
Table 3-8: Annotation for SCI question 7 122

Table 3-9: Annotation for SCI question 8..... 123
Table 3-10: Annotation for SCI question 9..... 124
Table 3-11: Annotation for SCI question 10..... 125
Table 3-12: Annotation for SCI question 11..... 126
Table 3-13: Annotation for SCI question 12..... 128
Table 3-14: Annotation for SCI question 13..... 130
Table 3-15: Annotation for SCI question 14..... 132
Table 3-16: Annotation for SCI question 15..... 133
Table 3-17: Annotation for SCI question 16..... 134
Table 3-18: Annotation for SCI question 17..... 135
Table 3-19: Annotation for SCI question 18..... 137
Table 3-20: Annotation for SCI question 19..... 138
Table 3-21: Annotation for SCI question 20..... 139
Table 3-22: Annotation for SCI question 21..... 140
Table 3-23: Annotation for SCI question 22..... 141
Table 3-24: Annotation for SCI question 23..... 142
Table 3-25: Annotation for SCI question 24..... 144
Table 3-26: Annotation for SCI question 25..... 146
Table 3-27: Annotation for SCI question 26..... 147
Table 3-28: Annotation for SCI question 27..... 148
Table 3-29: Annotation for SCI question 28..... 150
Table 3-30: Annotation for SCI question 29..... 151
Table 3-31: Annotation for SCI question 30..... 153
Table 3-32: Annotation for SCI question 31..... 154
Table 3-33: Annotation for SCI question 32..... 155
Table 3-34: Annotation for SCI question 33..... 156
Table 3-35: Annotation for SCI question 34..... 157
Table 3-36: Annotation for SCI question 35..... 158
Table 3-37: Annotation for SCI question 36..... 159
Table 3-38: Annotation for SCI question 37..... 161
Table 3-39: Annotation for SCI question 38..... 162
Table 4-1: Historical matrix for data selection. Data from the shaded areas were included in the analysis..... 165
Table 4-2: Item Statistics and Parameter Estimates..... 167
Table 4-3: Nominal response model item parameter estimates..... 182

LIST OF FIGURES

Figure 1-1: Average normalized gains on the FCI for traditionally taught courses and IE courses. The white bars indicate the fraction of 14 traditional courses and the black bars the fraction of 48 IE courses. (Richard R. Hake, 1998). 11
Figure 1-2: Model of the probability of a correct response to item X for a given Θ under the linear factor analytic model. Note that at the ends of the distribution, the probabilities become impossible. 67
Figure 1-3: Example of an Item Characteristic Curve (ICC). The threshold parameter β is the value of Θ for which the probability of a correct response is 0.5. 69
Figure 1-4: 1PL item characteristic curves for different values of the threshold parameter β. 70
Figure 1-5: 2PL item characteristic curves for different values of a, β=0 for all curves. . 71
Figure 1-6: 3PL item characteristic curve with *a*=1.5, *β*=0, and *c*=0.1.. 72
Figure 1-7: Item characteristic curve and its associated item information function. 74
Figure 2-1: Box Plots of SCI posttest scores by semester, (median represented by -, mean represented by x). 96
Figure 2-2: Box Plots of SCI posttest scores by course, (median represented by -, mean represented by x). 96
Figure 2-3: Results of the Tukey test on the means for course, presented as lines. 97
Figure 2-4: Distribution of course grades and SCI posttest scores by quartile. 101
Figure 3-1: A guide to reading the tables used in the annotated version of the SCI. A portion of a sample table is shown. 111
Figure 4-1: Item characteristic and information curves for items P1 and P1a. 169
Figure 4-2: Item Characteristic Curves for items P2 and P2a. 170
Figure 4-3: Item characteristic and information curves for questions P7 and P7a. 172
Figure 4-4: Item characteristic and information curves for items I4 and I4a. 173
Figure 4-5: Item characteristic and information curves for items I10 and I10a 175
Figure 4-6: Item characteristic and information cureves for items I11 and I11a. 176
Figure 4-7: Item characteristic and information curves for items D8 and D8a. 177
Figure 4-8: Distribution of threshold (*β*), parameters for SCI items. 178
Figure 4-9: Total test information and standard error curves in the logit metric. 180
Figure 4-10: ICC for item I3 from the 2PL model. 184
Figure 4-11: Response curves for item I3 from the nominal response model. 185
Figure 4-12: Response curves for item P1 and P1a for the nominal response model. ... 186
Figure 4-13: Response curves for item G2 for the nominal response model. 188
Figure 4-14: Response curves for item D9 for the nominal response model. 189
Figure 4-15: Response curves for item G6 for the nominal response model. 191
Figure 5-1: Proportion correct vs. theta for the SCI post test data from the summer 2005 administration. The data is superimposed over the test characteristic function which shows the expected proportion correct for each point along the theta distribution. 195
Figure 5-2: Pre and posttest scores for the summer 2005 administration. The top plot shows the observed scores as correct percentages. The bottom plot shows the scores as theta estimates. The line represents no change in the scores. Points above the line demonstrated a positive gain from pre to posttest. Those below represent negative gains. 198

Figure 5-3: Plot of the 18 largest eigenvalues of the inter-item tetrachoric correlation matrix. 200
Figure 5-4: A comparison of threshold (β) estimates obtained from the whole test and from subgroups of items divided by content area. 201
Figure 5-5: A comparison of slope (α) estimates obtained from the whole test and from subgroups of items divided by content area. 202
Figure 5-6: The percent of students whose total scores were at or below the 10th percentile who answered correctly or incorrectly on the 10 most difficult SCI questions. The data is from the spring 2005 post test. The number of students at or below the 10th percentile was 22. 204
Figure 5-7: A comparison of slope (α) parameter estimates obtained for two randomly selected subgroups of examinees. Items lying farthest from the line of equality are labeled with their master number. The correlation coefficient is also shown. 206
Figure 5-8: A comparison of threshold (β) parameter estimates for two randomly selected subgroups of examinees. Items lying farthest from the line of equality are labeled with their master number. The correlation coefficient is also shown. 207
Figure 5-9: Comparison of ability estimates based on even numbered items and odd numbered items only. 208
Figure 5-10: Comparison of ability parameter estimates based on subsets of the easiest and hardest items for data from summer 2005 209
Figure 5-11 Observed and expected proportion correct for item D10 (#29) based on the 2PL model. The line represents the item characteristic curve (and the expected proportion correct) and the data points are the observed proportion correct within the interval. The interval widths are 0.6; the midpoint of the interval is used as the observed theta value. The corresponding residual plot is shown below. 211
Figure 5-12: Item D10 observed proportion correct versus predicted proportion correct based on 1PL model. The corresponding residuals are also shown. 212
Figure 5-13: Observed and predicted proportion correct for item D6. 213
Figure 5-14: Observed and predicted proportion correct for item P2. 213
Figure 5-15: Frequency of confidence ranking for each response to question P1 (#1). Confidence was ranked on a scale from 1 ("Not confident at all") to 4 ("Very confident"). 218

ABSTRACT

The Statistics Concept Inventory (SCI) is one of several concept inventories currently being developed in a variety of engineering disciplines following the success of the Force Concept Inventory (FCI). The direction of the current reform movement in statistics education (as well as other science, engineering, and mathematics fields) is toward an emphasis on conceptual learning instead of focusing on procedural and computational skills. These new curricular goals have given rise to new assessment needs. The SCI is a multiple choice instrument modeled after the FCI which aims to assess conceptual understanding of fundamental statistics concepts. Development of the SCI began in 2002. An overview of the development process is presented here along with baseline performance data from a variety of university level statistics courses. SCI data is analyzed from a classical test theory perspective and from an item response theory (IRT) perspective using the two parameter logistic model and the nominal response model.

Posttest SCI results have been consistently low, between 40% and 50% correct; pretest to posttest gains have been minimal. These outcomes are consistent with concept inventory findings in other disciplines. As part of the ongoing development process, individual item analysis has been conducted including item discrimination, distribution of answers, and item correlation with the total score. Comments from student focus groups have also been used during the revision process. These detailed findings are presented as an annotated version of the SCI. Potential areas of confusion or possible misconceptions can be identified.

A clearer picture of student understanding emerges when the item analyses are combined with analyses obtained using IRT methods. In particular, the nominal response

model appears to be able to shed light on persistent misconceptions versus those that seem to diminish with instruction. Additionally, IRT methods can be utilized during the revision process to compare question versions, help make decisions which increase reliability, and make the revision process more efficient. Item characteristic curves for each question and for each response are presented. Results indicate that these methods should be very useful for revising and interpreting concept inventories, as well as having pedagogical implications.

Chapter 1: Literature

The Statistics Concept Inventory (SCI) is one of several concept inventories currently being developed in a variety of engineering disciplines (Evans, Gray, Krause, Martin, Midkiff, Notaros, Pavelich, Rancour, Rhoads, Steif, Streveler & Wage 2003a, 2005). Statistics is an increasingly important topic in many disciplines and is receiving increased attention in the K-12 curriculum (National Council of Teachers of Mathematics 2000). Within engineering, statistics is recognized as an important component of the engineering curriculum and is explicitly included in the ABET accreditation criteria (Engineering Accreditation Commission 2003).

Enrollment in undergraduate statistics has been rapidly increasing over the last ten years (Loftsgaarden & Watkins 1998, Schaeffer & Stasny 2004). During this same time, the reform movement in statistics education has been gaining momentum. The direction of the current reform movement is toward an emphasis on conceptual learning instead of focusing on procedural and computational skills (Cobb 1993, Gal & Garfield 1997b, Moore 1997, Garfield, Hogg, Schau & Whittinghill 2002, Ben-Zvi & Garfield 2004).

Gal and Garfield (1997b) outlined eight instructional goals for statistics education to help prepare students to understand and be able to use statistical information and data in an increasingly information dense society. These goals are to have students:

1. Understand the big ideas that underlie statistical inquiry. These ideas include:
    - The existence of variation
    - The need to describe populations by collecting data

- The need to reduce raw data by noting trends and main features through summaries and displays of the data
- The need to study samples instead of populations and to infer from samples to populations
- The logic behind related sampling processes
- The notion of error in measurement and inference, and the need to find ways to estimate and control errors
- The need to identify causal processes or factors
- The logic behind methods (such as experiments) for determining causal processes

2. Understand the method of statistical investigations. This includes study planning, data planning, data collecting and organizing, data analysis, interpretation of results, conclusions and implications.
3. Become proficient in procedural skills.
4. Understand the relationship between the mathematical parts (raw data, graphs, summary stats, etc) and how changes in data affect these.
5. Understand probability and chance where the emphasis is on an informal grasp of probability and an understanding of the commonly used language.
6. Develop interpretive skills and statistical literacy in order to become effective users of statistical information and be able to critically analyze and question it.
7. Develop the ability to communicate well and use statistical and probability terminology.
8. Develop an appreciation for statistical methods as a tool.

The American Statistical Association has recently endorsed a set of instructional guidelines published in 2005 by the Guidelines for Assessment and Instruction in Statistics Education (GAISE) project. The report includes the following recommendations for statistics education:

1. Emphasize statistical literacy and develop statistical thinking;
2. Use real data;
3. Stress conceptual understanding rather that mere knowledge of procedures;
4. Foster active learning in the classroom;
5. Use technology for developing conceptual understanding and analyzing data;
6. Use assessments to improve and evaluate student learning (Aliaga, Cobb, Cuff, Garfield, Gould, Lock, Moore, Rossman, Stephenson, Utts, Velleman & Witmer 2005)

New curricular goals have given rise to new assessment needs and challenges. A variety of authentic assessment techniques for classroom use are being explored including student portfolios, case studies, concept maps, group projects, and writing assignments. For a more thorough discussion of alternative assessment techniques see Gal and Garfield (1997a) and Garfield and Chance (2000). However, many of these methods are very time and labor intensive. There is still a place for carefully written multiple choice testing.

A multiple choice assessment tool, the Statistics Concept Inventory (SCI), is currently being developed for statistics content. The instrument will be used to assess student understanding of fundamental statistics concepts. The SCI is based on a model from physics education research. The Force Concept Inventory (FCI) developed by Halloun and Hestenes (1985b) was designed specifically to identify the preconceptions of physics students about Newtonian mechanics. The FCI has been instrumental in efforts to improve introductory physics courses (Hake 1987, 1992, Mazur 1997, Hake 1998). Use

of the FCI has indicated that while student preconceptions have a huge impact on their course performance, conventional instruction is only able to make small changes in the these "common-sense" ideas that students bring with them to class (Halloun & Hestenes 1985b). The enormous success of the FCI has been in its ability to identify issues with learning and to provoke a re-thinking of physics instruction.

It is this success that has prompted the writing of SCI. The SCI is still in the development phase, but it is hoped that it can play a similar role in the statistics education reform process. The following chapters will present some background information on the impetus for the development of the SCI and an evaluation and analysis of the instrument from multiple psychometric perspectives.

Chapter one presents relevant literature in three major areas that have impacted the development process for the SCI. First, a review of the FCI and its influence of the physics education reform movement are presented. Other concept inventories that are currently in development by researchers in the engineering sciences are also discussed with a focus on development procedures, outcomes, and common themes. Next, a focused review of statistics education literature is presented. This literature was consulted to develop items for the SCI that would include known misconceptions and common student errors. Finally, an overview of test theory methods is presented to lay the ground work for the analysis to be presented in chapters two, three and four.

Chapter two describes the development process for the SCI from topic selection to revision practices. It also presents baseline performance data including gains and the relationship between SCI scores and course performance. This chapter presents

information about the total test scores from a classical test theory tradition and includes reliability estimates using coefficient alpha.

Chapter three provides a more microscopic view of the SCI. An item by item analysis is presented in the form of an annotated version of the SCI. For each item, classical test theory item analysis statistics are reported including response patterns, discrimination indices, difficulty, and item-total correlations. This section also includes commentary on question behavior, insights gained from focus groups, and references to relevant literature.

Chapter four presents analyses of the SCI using methods from item response theory. Two models are used: the two parameter logistic model and the nominal response model. This chapter explores ways that item response theory techniques can be used to gain additional insight into question behavior and student misconceptions, in addition to being valuable tools for development and evaluation practices.

Chapter five discusses directions for future research and additional ways that item response theory methods can be employed in the future development of the SCI and other concept inventories.

1.1 Force Concept Inventory

In the early-1980's, physics education researchers began to investigate the level of conceptual understanding demonstrated by students in introductory physics classes. The Force Concept Inventory (FCI) was developed by Halloun and Hestenes (Halloun & Hestenes 1985b, Hestenes, Wells & Swackhamer 1992) as a way to quantitatively measure introductory physics students' understanding of mechanics concepts. In particular, the FCI was designed to address students' common sense beliefs about

physical phenomena derived from real world experiences of motion, force, etc. These common sense beliefs are often at odds with Newtonian physics. Because students integrate new physics instruction into their current understanding, if these misconceptions/preconceptions are not directly addressed by instruction, it is often very difficult to change the way students think about and understand physics.

Six general areas of mechanics are covered on the FCI. As defined by the authors, these six dimensions are Kinematics, Newton's First Law, Newton's Second Law, Newton's Third Law, Superposition Principle, and Kinds of Forces. Each dimension is an essential component of the Newtonian Force Concept. These areas are described in more detail and each item in the inventory is classified into one of these six categories in Hestenes, et al. (1992). A portion of the classification is duplicated in Table 1-1. The authors recommend looking at the inventory as a whole for an overall pattern of Newtonian answers rather than giving great weight to individual items.

Table 1-1: A portion of the item classification for the Force Concept Inventory (Hestenes, et al. 1992).

Topic		*Inventory Item*
0. Kinematics		
	Velocity discriminated from position	20E
	Acceleration discriminated from velocity	21D
	Constant acceleration entails	
	parabolic orbit	23D, 24E
	changing speed	25B
1. First Law		
	With no force	4B, (6B), 10B
	Velocity direction constant	26B
	Speed constant	8A, 27A
	With canceling forces	18B, 28C

The FCI is notably different from traditional physics exams and homework assignments, which typically are predominately problem based. The FCI is made up

entirely of multiple choice questions that assess qualitative understanding of fundamental force concepts and that require no problem solving or computation. There are 36 questions on the FCI, each with five item responses. Each question has one correct answer. The nature and construction of the questions are such that the FCI can be given to students who have had no formal physics instruction, as in a pretest situation to assess what specific misconceptions/preconceptions students hold. That is, for example, the questions contain no physics jargon and do not ask about specific laws by name. Furthermore, the questions are considered to be basic, even "trivial" or "simple" by physics professors (Hestenes, et al. 1992 p. 142, Mazur 1997 p. 4).

The items were designed to target specific areas where common sense understanding is known to be markedly different than the Newtonian explanation of motion and its causes. The FCI was developed by first giving the item stems as free response questions. The distractors were then developed from these responses. The FCI has been found to be very reliable based on a very high coefficient alpha (0.86 for pretest, 0.89 for posttest), stability of student answers in test/retest situations and even in test/interview situations. Student interview data indicate that question responses are related to stable beliefs that are resistant to change, even when challenged explicitly. Face and content validity were determined by consulting physics professors and graduate students for input on the questions, by administering the FCI to 31 physics students who received A's in their physics course, and looking for common mistakes that might be attributable to question formulation (Halloun & Hestenes 1985b).

The FCI was administered to 1,500 college physics students and 80 high school physics students along with a mathematics diagnostic test (Halloun & Hestenes 1985b).

Pretest scores were very consistent across the different student populations. Scores on the FCI remained low even after instruction, as can be seen in Table 1-2. The low correlation between the FCI and the mathematics diagnostic test ($\rho = 0.19$) gave evidence that mathematical competency did not ensure success in physics. The pretest was found to be similarly positively correlated with course grade in three university physics courses ($\rho \approx 0.56, p = 0.001$); post test scores were more highly correlated.

The pretest and mathematics diagnostic test together were better able to predict student performance in physics courses than any other combination of variables, based on a stepwise linear regression (Halloun & Hestenes 1985b). The two variables accounted for 42% of the total variation. The pretests together with all previous math and physics course work accounted for 51% of the total variation. The researchers also considered gender, age, academic major, and background courses in science and mathematics. Differences in these variables did not affect performance. They also found very little gain in performance between the pre and posttests, between 11 and 15% for college and university physics students. They found that student gain was independent of instructor.

Table 1-2: Summary of FCI pre and post outcome data, based on Halloun and Hestenes (1985b).

		Mean (Standard deviation)		
Course Type	*Sample Size*	Pre	Post	*Gain*
University Physics	478	18.5 (5.6)	23 (5.5)	4.5
(calculus based)		51%	64%	12.5%
College Physics	405 (pre),	13.7 (5.1)	19 (5.2)	5.5
(not calculus based)	82 (post)*	38%	52%	15%
High School Physics	49	10.9 (3.5)	17.3 (4.6)	6.4
		30%	48%	18%

* Posttest data not available for two classes.

Further analysis of the FCI exams, along with student interviews, allowed the researchers to classify student common sense beliefs into six major categories (Halloun &

Hestenes 1985a, Hestenes, et al. 1992). Interestingly, student beliefs were consistent with various historical understandings of physics concepts, such as those held by Aristotle and Galileo. The distractors to the items are categorized as well, and the choice of a particular distractor gives evidence of the presence of particular misconceptions. At least as much information is contained in the incorrect responses as in the correct response. These incorrect choices can help to identify commonly held misconceptions. The authors note that the FCI "is not a test of intelligence; it is a probe of belief systems" (Hestenes, et al. 1992 p. 142). A portion of the taxonomy of misconceptions is shown in Table 1-3. As such, the FCI can function as a very useful diagnostic tool. In addition, an entry threshold FCI score of 60% is proposed, below which students' understanding of concepts makes problem solving very difficult. A score of 85% is suggested as the Newtonian Mastery Threshold. Scores above 85% identify students who are confirmed Newtonian thinkers. The authors also propose that good instructional methods should be able to achieve all students receiving a passing grade scoring at least 75% on the FCI (Hestenes, et al. 1992, Hestenes & Halloun 1995).

Table 1-3: A portion of the taxonomy of misconceptions developed from the Force Concept Inventory (Hestenes, et al. 1992).

Misconception	*Inventory Item*
0. Kinematics	
K1. Position-velocity undiscriminated	20A,C,D
K2. Velocity-acceleration undiscriminated	20A; 21B,C
K3. Nonvectorial velocity composition	7C
1. Impetus	
I1. Impetus supplied by "hit"	9B,C; 22B,C,E; 29D
I2. Loss/recovery of original impetus	4D;6C,E;24A;26A,D,E
I3. Impetus dissipation I4. Gradual/delayed impetus build-up	5A,B,C;8C;16C,D;23E,27C,E;29B 6D;8B,D;24D;29E
I5. Circular impetus	4A,D;10A

Since its development, the FCI has been widely used with thousands of students in high school and college introductory physics courses. In Hake's survey study of more than 6500 students in 62 introductory physics classes, FCI results are reported from a wide variety of schools, (both high school and college) and instructors (Hake 1998). FCI data were solicited at colloquia, meetings, and on list-serves. This type of data selection has inherent bias toward more positive outcomes. The objective of this survey was to determine whether the use of interactive engagement methods could substantially improve the effectiveness of introductory physics courses over traditional methods. Hake then defined interactive engagement (IE) methods as "those designed at least in part to promote conceptual understanding through interactive engagement of students in heads-on (always) and hands-on (usually) activities which yield immediate feedback through discussion with peers and/or instructors" (Hake 1998 p. 65). Courses were classified as IE courses if instructors reported making "substantial use of IE methods".

In order to make these cross course comparisons, Hake reports the normalized gain from pretest score $\langle S_{pre} \rangle$ to posttest scores $\langle S_{post} \rangle$. The normalized gain $\langle g \rangle$ is defined as the ratio of actual average gain $\langle G \rangle$ to the maximum possible average gain:

$$\langle g \rangle \equiv \frac{\langle G \rangle}{\langle G \rangle_{\max}} = \frac{\%\langle S_{post} \rangle - \%\langle S_{pre} \rangle}{100 - \%\langle S_{pre} \rangle}$$

where $\langle S_{pre} \rangle$ and $\langle S_{post} \rangle$ are the pre and post class averages. Three levels of gain are distinguished, "High-g" courses are those with $\langle g \rangle \geq 0.7$, "Medium-g" are those with $0.3 \leq \langle g \rangle < 7$ and "Low-g" are those with $\langle g \rangle < 0.3$.

All 14 of the traditional courses fell in the low-g range (n=2084). Eighty-five percent (41 courses, n=3741) of the IE courses fell in the medium-g range. No courses achieved the high-g range. A histogram of the normalized gains is shown in Figure 1-1. In addition, where data were available, posttest scores were found to be significantly and strongly correlated with Mechanics Baseline posttest scores, $r = +0.91$. The Mechanics Baseline test is designed to measure more quantitative, problem solving aspects of mechanics (Hestenes & Wells 1992). This is additional evidence that problem solving skills are not sacrificed with additional attention to conceptual learning; they are instead enhanced as well.

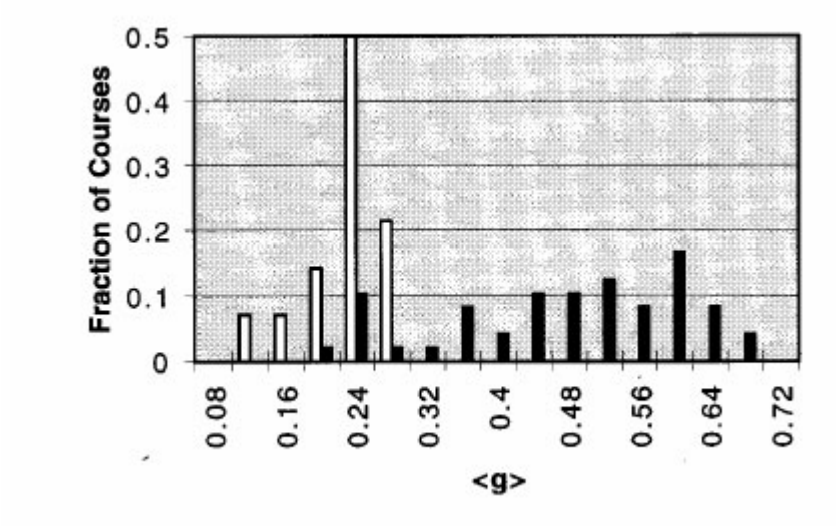


Figure 1-1: Average normalized gains on the FCI for traditionally taught courses and IE courses. The white bars indicate the fraction of 14 traditional courses and the black bars the fraction of 48 IE courses. (Hake 1998).

The results of the FCI have been instrumental in developing and evaluating new methods of physics instruction. One of the more far reaching instructional innovations set in motion by the FCI is called Peer Instruction developed by Mazur (1997). A revised version of the FCI is included in Peer Instruction: A User's Manual (Mazur 1997). This method utilizes pre class reading assignments and in class, small-group student

discussion of ConcepTests. These are single multiple choice questions that address physics concepts. Students are encouraged to convince their peers, thus requiring them to explain their reasoning. Responses are gathered in class and instructional decisions are made based on the outcome (i.e. move forward or revisit material). With this approach, along with changes in examination practices to include concept oriented questions in addition to problem solving questions, student focus is redirected to physics principles and concepts and less on algorithmic problem solving and memorization. Implementation of the Peer Instruction method doubled the normalized gain (gain/maximum possible gain) on the FCI (Crouch & Mazur 2001). As the method has been refined, further improvements have been seen. The Peer Instruction method has also increased problem solving ability as measured by the Mechanics Baseline Test (Hestenes & Wells 1992) and traditional problem based exams

Another example is the Socratic Dialogue Inducing Labs which are designed to introduce cognitive conflict through the use of simple mechanics labs that contradict students' common sense beliefs (Hake 1992). Then students are helped to construct new notions of physics that are consistent with Newtonian mechanics through dialogue with their peers and instructor. Other examples of innovations to physics instruction can be found in Knight (2002), Hake (1987), Halloun and Hestenes (1987), and Tobias and Hake (1988). Most of these methods are built around an active learning environment and a concentration on developing conceptual understanding in addition to problem solving ability.

In addition to starting an energetic movement in physics pedagogical research and instructional innovation, the FCI has sparked an ongoing dialogue about teaching and

learning among science, mathematics, and engineering communities. There has also been tremendous interest in the idea of concept inventories for many different disciplines, especially in the engineering sciences. Some of these are briefly discussed in the next section.

1.2 Other Concept Inventories-Related Efforts

Nineteen concept inventories being developed for engineering sciences are in various stages of development and are presented here. Six are independent efforts. The remaining thirteen are part of a coordinated effort by The Foundation Coalition. This group is a National Science Foundation funded engineering coalition established to help bring about systemic renewal for the engineering educational community. Many of their efforts focus on the first two years of engineering education, the foundational years. Major ideas guiding the work of the Foundation Coalition include: active and cooperative learning; increasing the participation of women and underrepresented minorities; student teams; technology enabled learning; continuous improvement through assessment, evaluation, and feedback; curriculum integration; and curricular change resistance and leadership (Foundation Coalition 2005). As part of this work, in 2000, the Foundation Coalition initiated a program of development for concept inventories for the engineering sciences (Evans, Gray, Krause, Martin, Midkiff, Notaros, Pavelich, Rancour, Rhoads, Steif, Streveler & Wage 2003b). Thirteen concept inventories are either in development or ready for use. The Foundation Coalition concept inventories are discussed in sections 1.2.5 -1.2.14 .

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

The Determining and Interpreting Resistive Electric Circuit Concepts Test (DIRECT) is designed for high school and university audiences (Engelhardt & Beichner 2004). Examination of textbooks and conversations with instructors were used to develop topic lists; then independent experts were consulted for additional input. Multiple questions were drafted for each topic and presented to students in an open-ended format in order to construct distractors. In addition, known misconceptions from research were incorporated into distractors. The first multiple choice version was given to 1135 high school and university students. After initial results were analyzed, a second version was tested with 692 students. The questions had been modified so that each had five response alternatives which resulted in a somewhat more quantitative version. The results from each version are shown in Table 1-4.

Table 1-4: Summary of results from the Determining and Interpreting Resistive Electric Circuit Concepts Test (DIRECT) (Engelhardt & Beichner 2004).

Version	*N*	*Mean*	*Range*	*Reliability (KR20*)*
1.0	1135	48%	14%-97%	0.71
High School	454	41%	14%-90%	
University	681	52%	21%-97%	
1.1	692	41%	3.4%-90%	0.70
High School	251	36%	3.4%-76%	
University	441	44%	10%-90%	

*Kuder-Richardson equation 20, equivalent to coefficient alpha when items are dichotomously scored

The instrument contained 29 items and students were given approximately 30 minutes to complete it. Item analysis was conducted, including difficulty, point-biserial correlations and discrimination. Content validity was addressed by using the expert panel to ascertain adequate topic coverage. Construct validity was evaluated with a factor

analysis: eight factors were found for the first version (1.0) and eleven factors were found for the second version (1.1). In addition, student interviews were conducted to determine that students were correctly interpreting the questions and that results on the inventory replicated results from previous research.

ANOVA and t-test methods were used to look for differences in performance on the inventory for different groups. Differences were found between university high school students and between males and females, with males receiving higher scores, using fewer misconceptions, and expressing more confidence in their answers. Also, differences were found between courses where a traditional lecture format was used compared to courses where alternative formats were used, including active learning and one course which used a new textbook emphasizing the microscopic aspect of circuit phenomena. Students in the alternative course formats outperformed students in both algebra and calculus based traditional courses.

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

Beichner (2004) has developed an instrument to assess understanding of the commonly used kinematics graphs in introductory physics (for example position vs. time, velocity vs. time, etc.). Eight objectives were identified after examining test banks and text books, and conducting informal interviews with physics instructors. The list was reduced to seven after eliminating one that students typically experience little difficulty with (going between a point on a graph to a coordinate pair and back). Three items were written for each objective, for a total of 21 items. Questions focused on kinematics graph interpretation. Distractors were written to incorporate known difficulties. Some questions

were initially posed to students in an open ended format and frequently appearing mistakes were also turned into distractors.

To establish content validity, after initial piloting and revisions, the test was given to 15 instructors from high school through university level. They were asked to complete the test, make comments and criticisms, and match the questions to objectives. A group of 165 junior and senior high school students and 57 college students also took the exam. All had received traditional kinematics instruction. Following the exam, they participated in a two hour laboratory activity in kinematics. Within a week of the laboratory activity, they took a second parallel version of the exam, in which questions had been slightly modified (i.e. graph scales shifted, shapes changed slightly). Correlation between the two versions of the exam was found to be 0.79. There was a significant increase in the scores from pre to post laboratory. This is cited as further evidence of validity, since the only instruction students had received was related to kinematics graphs.

A final version of the exam was given after instruction to 524 college and high school students. The mean score was 40% with a standard deviation of 22%. The reliability was estimated from this administration to be $KR20 = 0.83$. Test and item analyses were also carried out, including calculating the standard error of the mean (0.01%), Ferguson's Delta (0.98), point-biserial correlations, item discrimination, and item difficulty.

Calculus based physics students scored higher than students in algebra based courses. Males scored significantly higher than females. No differences were found between high school and college students. Specific difficulties and misconceptions are identified in Beichner (2004). He notes that the first step in addressing these difficulties

is “for teachers to become aware of the problem. Knowing that students cannot use graphs as ‘fluently’ as they should means that in-class discussions of kinematics situations and variables cannot start by simply referring to their graphs” (p. 755).

1.2.3 Statics Concept Inventory

The development of the Statics Concept Inventory began with classifying the central concepts of statics into four clusters (Steif 2004). The questions on the inventory focus on major conceptual tasks and distractors represent distinct errors. Distractors were selected based on input from experienced instructors and students’ solutions to statics problems. Two groups of student solutions were analyzed, one from the beginning of a statics course in which students had some limited prior exposure to the subject matter and one from students at the end of a completed course in statics. The two sets were used in order to include both naïve errors and errors that persist after instruction.

The inventory consists of 27 questions covering 8 topics (Steif, Dollar & Dantzler 2005). Most of the questions require no computation. While some questions involve minimal computations, distractors to these questions represent correct computational answers based on incorrect conceptual reasoning. Typical errors for groups of questions have been identified (Steif 2004).

During 2003-2004 academic year, data were collected from 245 students from 5 universities (Steif, et al. 2005). Based on these data, the reliability estimate was found to be $\alpha = 0.89$. Confirmatory factor analysis was conducted using an 8 factor model, and a good model-data fit was found (Goodness of Fit Index = 0.90). Several questions were identified as adversely affecting the model-data fit and were modified for the 2004-2005 administration.

Posttest means on the inventory ranged from 14%-20% depending on the site. The higher average score (20%) was obtained in the course taught by the author of the instrument, who acknowledges his teaching is influenced by his work on the inventory. However, the author does not specifically describe his teaching style (e.g., more traditional, IE, concept oriented…).

Analysis of data from 100 students in fall 2004 found strong correlations between scores on the Statics Concept Inventory and average course examination scores or final examination scores. Comparisons were made between course exam scores and inventory scores by quartile. These tables show a clear trend between higher post test inventory scores and higher exam scores. An example of this is reproduced in Table 1-5. Also, a finer analysis of student course exams showed that students who committed certain errors in their course examinations were found to have lower related sub scores on the inventory.

Table 1-5: A comparison of the percentage of students in each quartile on the Statics Concept Inventory and the final exam score received at one site.

	Quartile 1	*Quartile 2*	*Quartile 3*	*Quartile 4*
A (n=9)	0%	11%	44%	44%
B (n=8)	0%	25%	25%	50%
C (n=10)	50%	20%	0%	30%
D (n=2)	50%	50%	0%	0%
F (n=9)	33%	44%	22%	0%

The authors also define the Inventory-Exam Discrepancy (IED) for each student by:

$$\text{IED} = \frac{\text{Inventory Score} * \text{Class Exam Mean}}{\text{Class Inventory Mean}} - \text{Exam Score}$$

.

IED has a mean value of 0 and is positive for students who did well on the inventory

compared to their exam score (relative to the class average on each measure). Highly correlated exam and inventory scores result in a narrow range of IED. The range of IED for the three classes varied widely along with the correlations between scores.

1.2.4 Force and Motion Conceptual Evaluation (FMCE)

The Force and Motion Conceptual Evaluation (FMCE) covers Newtonian force and motion concepts, similar to the FCI (Thornton & Sokoloff 1998). However, the FMCE is more focused in its coverage, covering fewer topics than the FCI with each topic addressed in multiple contexts. There is also more emphasis placed on graphical representations (Saul 1998). Some of the questions are identical to those on the FCI, the rest were developed through student interviews and student answers to open ended questions. The FMCE was developed to evaluate new instructional materials and is intended to be more of a diagnostic instrument for making decisions about individual students.

1.2.5 Signals and Systems Concept Inventory (SSCI)

The recently introduced Signals and Systems Concept Inventory (SSCI) has two versions (continuous-time and discrete-time) available (Wage, Buck, Wright & Welch 2005). The instruments each consist of 25 multiple choice questions, which require little or no computation. The problem stems contain little if any quantitative data so students cannot rely on memorized computational routines. The distractors contain known student misconceptions.

The FCI was once again a catalyst in the development of the SSCI. During the initial development phase, 30 questions (in the continuous-time (CT) format) were presented along with five possible responses and the option to fill in an additional

answer. This option was presented in hopes of eliciting common misconceptions that may have been overlooked by the authors. Few students, however, chose this option. As with the FCI, the questions were intended to be conceptual in nature. The discrete-time (DT) version was constructed by writing discrete analogs to each question on the CT version and by drafting additional questions specific to DT signals and systems.

In spring 2001, data were collected from 129 undergraduate and graduate students from two schools for the CT version. Data for the DT version were collected the following semester, with 188 undergraduate and graduate students participating from three schools. This initial administration indicated that the test was too long and too difficult. Students had difficulty completing the instrument within the allotted one-hour time limit. The mean score for the CT version was 29%. The instruments were revised following this administration. The key changes that were made were: the addition of questions that addressed the mathematical background required in the signals and systems course, item analysis to eliminate the least common distractor, and a reduction in the total number of questions to 25.

Using the revised instrument, data were collected from approximately 600 students from 4 universities. Many of the students were encouraged to give their best effort on the SSCI with offers of performance based incentives such as bonus points. The authors are not specific about how many of the students were offered these incentives. Following Hake's (1998) comparisons with FCI data, the courses were divided into two categories: traditional lecture format and interactive engagement (IE) format. Fifteen traditional courses and five IE courses were included. Normalized gains for the two types of courses were calculated from pre- and posttest scores and were found to be remarkably

similar to those obtained with the FCI (Table 1-6). The normalized gain $\langle g \rangle$ is defined as

$$\langle g \rangle = \frac{post - pre}{100 - pre}$$

where pre and post are the average scores for each course using only students who took both the pre and post tests. The normalized gain can be interpreted as the fraction of concepts that students learn during the course that they did not know prior to the course. These results lend further support for the use of interactive engagement methods to support and promote conceptual understanding.

Table 1-6: Comparison of normalized gains by course type for the SSCI and the FCI (Wage, et al. 2005).

	Traditional Courses	*Interactive Engagement (IE) Courses*
SSCI	$\langle g \rangle = 0.20 \pm 0.07$ $n = 15$	$\langle g \rangle = 0.37 \pm 0.06$ $n = 5$
FCI	$\langle g \rangle = 0.23 \pm 0.04$ $n = 14$	$\langle g \rangle = 0.48 \pm 0.14$ $n = 48$

Correlation analysis was also conducted with pre- and posttest scores, gains, course grades, GPAs, and prerequisite course grades. The authors recommend the use of this type of analysis with the SSCI to aid in making curricular decisions within departments and as an accreditation tool. Strong positive correlations between SSCI scores and signals and systems course grades would provide evidence that the course was promoting conceptual understanding. Similarly, positive correlations between gains and prerequisite courses would identify courses that are most critical to later success in signals and systems courses. In addition to these types of decisions, the authors recommend the use of the SSCI for use in the ABET accreditation process, for

researching the effects of teaching methods and to help in identifying persistent student misconceptions.

The authors do not report any reliability analysis for the instruments, nor do they address validity issues.

1.2.6 Conceptual Survey of Electricity and Magnetism (CSEM)

Maloney, O'Kuma, Hieggelke, and Alan published their 32 questions Conceptual Survey of Electricity and Magnetism (CSEM) in 2001. They report on data collected from over 5000 students at 30 different institutions, including high school and undergraduate introductory physics students (at both 2 and 4 year colleges), as well as some graduate students. The authors consider the instrument to be more of a broad survey of concepts within the domain of electricity and magnetism as compared to the more focused FCI.

In developing the CSEM, the authors had to consider several issues inherent in the subject domain that were less of a problem in the FCI. Electricity and magnetism encompass a much larger domain than that covered by the FCI. There is more widespread unfamiliarity with concepts, phenomena and language (particularly in a pretest situation) among students who have had less exposure and life experience with the phenomena. Additionally, domain concepts rely heavily on other physics concepts such as force, energy and motion. The authors also note that there is minimal research available on alternative student conceptions about topics in this domain area. As such, the CSEM is a balance that combines questions about basic phenomena with questions about the formalism of the discipline, questions posed in everyday language and questions posed in

formal physics terms. “It is a combination of a test of alternative conceptions and knowledge” (2001 p. S19).

Development of the CSEM began with two separate tests, one for electricity and one for magnetism. A list of major topics to be included and the original problem sets were developed by a group of physics professors at a two-year college physics workshop. Original versions of these two tests were piloted during the 1995-1996 academic year. Revisions to the questions were made based on analysis of this data along with results from open-ended versions of some of the questions, and a second version was administered during 1996-1997. After this, a single test was constructed from subsets of questions from the two tests. Revisions continued to be made based on response analysis, student explanations of their responses, and feedback from instructors who used the instrument.

Baseline data for pre- and posttest scores are summarized in Table 1-7. Pretest scores are very low; however, examination of responses indicates that students are not answering randomly. Overall, posttest scores are also low and described as unexpected and disappointing. These include data from two high school classes, which were included with the college data since the scores on both the pre- and posttests were not significantly different between the high school and college courses. In addition to the results given from regular introductory physics courses, results from an honors calculus based engineering physics course are shown. This course is described as an interactive engagement course and significantly higher pre- and posttest scores are observed. While it is not explicitly stated by the authors that all of the other courses surveyed were taught in a traditional manner, it is assumed that this is the case. While the authors note that the

better performance is to be expected of the honors students, in light of the results seen with the FCI (Hake 1998) and the SSCI (Wage, et al. 2005), it is likely that the IE instructional approach may also have played a part in the higher posttest scores. Scores are also shown for a small mixed group of advanced undergraduate and graduate students. The authors point to this progression of scores as evidence that the instrument is measuring learning in this domain.

Other evidence of validity was obtained by surveying 42 two-year college physics professors. They were asked to rank each question for its reasonableness and appropriateness on a 1-5 scale. Separate rankings were obtained for use in algebra based courses and calculus based courses. All questions were ranked as highly reasonable and appropriate for both audiences. Reliability of the instrument was assessed using equation KR20 and found to be approximately 0.75 for post test data.

Table 1-7: Baseline pre and post test data for CSEM (Maloney, et al. 2001).

	Pre		*Post*	
Course	$\bar{x}$ % (sd %)	n	$\bar{x}$ % (sd %)	n
Algebra based	25 (8)	273	44 (13)	262
Calculus based	31 (10)	1213	47 (16)	1030
Honors, Calculus based*	41 (21)	99	69 (14)	145
Majors and Graduate Students			70 (17)	24
Physics Professors attending Two Year College Workshop project sessions			77 (**)	95

*Interactive Engagement Course
**not reported

Individual item response patterns are given combined for all courses in the two categories: algebra based and calculus based. These results combine two and four year colleges, high school students, and graduate students. Misconceptions were identified where possible and topics for further research were highlighted. In addition to response patterns, item analysis included calculation of difficulty and discrimination. Item

difficulty (the percent of students who answered correctly) ranged from 0.1 to just over 0.8. Item discrimination ranged from 0.1 to 0.55, with all but four questions having discrimination values over 0.2.

Principal component analysis was also carried out. Eleven factors were identified, the largest of which accounted for only 16% of the variance. The authors note that factor structure could be improved by adding more questions so that more questions focused on individual topics, however, this would increase the length of time required to take the test, making this modification unfeasible.

The authors suggest that the instrument be used to provide an estimate of student learning in key areas within the domain of interest and hope that results can help guide research into student common sense conceptions.

1.2.7 Thermal and Transport Science Concept Inventory

A Delphi study involving 30 faculty experts and textbook authors was used to identify important concepts (Olds, Streveler & Miller 2004). The experts were asked to rate the concepts on a preliminary list of 28 on how well students understand the concepts and how important the concept is for them to know. From this, 10 critical items were selected.

Sample questions were written and compensated student volunteers were asked to think aloud as they answered the questions. Students were interviewed during this process. Based on these interviews, the questions were revised for clarity and distractors were developed, using the exact language from the interviews when possible. Development of the instrument is continuing. Field testing to assess reliability is currently in progress.

1.2.8 Wave Concept Inventory

A similar assessment tool, the Wave Concepts Inventory (WCI) has been developed to assess student understanding of wave phenomena (Roedel, El-Ghazaly, Rhoads & El-Sharawy 1998). The WCI uses a multiple choice format but allows for more than one correct answer. There are 20 questions with 34 correct answers. The use of multiple correct answers is unique to this concept inventory. Students with a deeper understanding of wave phenomena should be able to recognize more of the correct answers.

The WCI was used to evaluate a new integrated engineering course at the Arizona State University. The course was designed to employ active learning strategies and technology. Students were administered the WCI as a pre- and posttest in both the integrated course and in the traditional course. There was a significant increase between the pretest and posttest scores for the integrated class, but not for the traditional class. Scores and gains for each class are summarized in Table 1-8. As noted by the authors, the teachers for the integrated course were the authors of the WCI. This could have been a factor in the increased performance.

Table 1-8: WCI score results for each class (Roedel, et al. 1998).

Course	**Pretest**	**Posttest**	**Gain**
	Mean % (Standard deviation %)		
Integrated	35 (8)	45 (8)	10 (4) $p = 0.0001$
Traditional	31 (8)	35 (10)	3 (10) $p = 0.077$

$* H_0 : \bar{x} = 0$

Correlation analysis was conducted using pre- and posttest scores, final course grades, and gains. Pretest scores were negatively correlated with gains and positively

correlated with posttest scores, course grades were not significantly correlated with any other variable.

1.2.9 Dynamics Concept Inventory

The impetus for development of the Dynamics Concept Inventory was the FCI and its successful use in stimulating and evaluating instructional innovation. The developers used a modified Delphi process to construct the list of potential topics for inclusion in the instrument. Twenty-five dynamics instructors were asked to "describe the concepts in 2D rigid body dynamics that your students find difficult to understand" and to provide a description of common misunderstandings (Gray, Costanzo, Evans, Cornwell, Self & Lane 2005 p. 2). These responses were combined and then the instructors were asked to rank the 24 different concepts on their importance and degree of difficulty. From these 24 concepts, 10 were selected to be included on the 30 question instrument. This allowed for specific concepts to be addressed by multiple questions. Questions were then drafted and revised by the development team.

The items were administered to students as open ended questions in focus groups in order to develop and refine the response sets. Focus groups were also conducted with multiple choice versions of the questions and student interviews were then held to further understand how students were interpreting and reasoning through the questions. After additional similar revisions, 11 concepts were selected to be included on the exam. Four questions from the FCI were also included.

During 2003 and 2004, the DCI was given at a large public university (LPU) and a small private university (SPU). The SPU instructors made use of concept quizzes in class. The course at LPU was taught with a traditional lecture format during 2003, but in

2004 a clicker system was instituted and multiple choice, concept oriented questions were asked during class, similar to the Peer Instruction model (Mazur 1997). These non-traditional instruction methods resulted in much higher posttest scores than the traditional lecture format, consistent with results in other disciplines. A summary of the results is shown in Table 1-9.

Table 1-9: Dynamics Concept Inventory Results

Course (Instruction type)	*n*	*Mean % (sd %)*	*Coefficient α*
LPU 2003 (Traditional)			
Post	147	32.1 (15.0)	0.640
LPU 2004 (IE)			
Pre	441	30.6 (14.2)	0.719
Post	310	55.7 (19.3)	0.837
SPU 2003 (Concept focused)			
Pre	172	34.9	—
Post	166	63.9 (16.8)	0.730

Reliability estimates using Cronbach's alpha were found for each administration and are also shown in Table 1-9. The authors address the issue of test validity by referring to the test construction process and comments from instructors who have used the instrument. Using the results from the DCI, specific misconceptions have been identified.

The current version of the DCI contains 29 questions and covers 11 concepts. It can be accessed from http://www.esm.psu.edu/dci/. Students should be allowed 30 minutes to complete the instrument. The authors recommend the use of the DCI for evaluating curricular innovations and to measure conceptual gains.

1.2.10 Fluid Mechanics Concept Inventory (FMCI)

The Fluid Mechanics Concept Inventory (FMCI) is still being developed (Martin, Mitchell & Newell 2004). In constructing the instrument, an initial concept list was developed by faculty, then questions were drafted for each concept. Questions were written avoiding computation and including graphical and visual representations of the concepts. A list of special topics was also constructed for use in various disciplines. Three principle areas and twenty-five topics were included.

Students were involved in this initial development stage. Students who had completed a fluid mechanics course were asked to review their textbook and notes and then to construct a list of "10 concepts they were certain of and that they felt were important and a list of 10 concepts they were uncertain of and felt were not important" (Martin, Mitchell & Newell 2003 p. T3D-24). These concepts were compared to the concept list generated by faculty. Students were then videotaped discussing their list and a set of questions developed by the faculty.

The initial version of the FMCI included 27 questions and covered ten basic concepts (Martin, et al. 2004). Reliability assessment has been limited to examining point-biserial correlations for individual questions. Following this assessment, six questions were left unchanged, five deleted, and the remainder revised.

1.2.11 Chemistry Concept Inventory

The Chemistry Concept Inventory is designed to cover chemistry concepts that overlap with subsequent engineering courses (Krause, Birk, Bauer, Jenkins & Pavelich 2004). Two inventories are being produced, for Chemistry I and Chemistry II. Once the topics were selected, a literature search was conducted to identify known misconceptions.

Then three questions were written for each topic to be included. The questions were written to be conceptual, not computational. A total of 61 questions were originally drafted for the two inventories.

The questions were given initially to students during their weekly quizzes after the information had been covered during lecture. Therefore, the questions were not given initially as a single instrument. However, the data were combined and analyzed as if it were a single instrument. Item discrimination, difficulty, and alpha-if-item-deleted were determined for each question along with coefficient alphas for the entire question sets. These initial results are summarized in Table 1-10. Based on this analysis of initial data, questions were eliminated so that 10 questions remained on each version of the exam.

Version B was then given during the summer of 2003. See Table 1-10. Student interviews were conducted with 11 students following this administration. Information from these interviews was used to clarify question wording and verify student misconceptions. Revisions were made again and the third version, C, was used during the fall of 2003.

Table 1-10: Chemistry Concept Inventory results summary (Krause, et al. 2004).

		Mean%		*Coefficient α*
CCI-I Version:	N	Pre	Post	Post
A	326		49.1%*	0.7883
B	42	27.4%	53.0%	0.7135
C	845	24.7%	44.5%	0.6803
CCI-II Version:				
A	158		59.8%*	0.7855
B	42	35.9%	54.7%*	0.4188
C	136	33.6%		48.1%

*Questions given in weekly quizzes throughout the semester and combined for a post test score.

Normalized gains were computed for the data from version C. Low normalized gains were seen in all courses, with average values between 0.2 and 0.3. Significantly

higher gains were seen in Chemistry I classes in which instructors utilized a more non-traditional instructional format including student group work during the lecture period. While many other factors may influence student gains, in light of other research discussed here, instructional format was likely a contributor. Correlations between inventory posttest scores and classroom averages were also calculated. In all cases, the correlations were significant and near 0.6.

1.2.12 Heat Transfer Concept Inventory (HTCI)

Unlike the majority of other concept inventories, the development of the Heat Transfer Concept Inventory (HTCI) began with student groups to develop topic lists. Small groups of students at two universities were hired to participate in the project. They were asked to generate lists of concepts that they felt were important and that they were sure of, and a list of important topics that they were not sure of. In focus groups, students were asked to discuss many of these topics and the conversations were videotaped. It was clear from this work that students had very vague and fuzzy understandings of most concepts in heat transfer and that what was understood was unconnected to other concepts in the course. The authors remark that students were "deeply confused at a fundamental level" and that faculty were "very surprised" at how poorly basic concepts were understood (Jacobi, Martin, Mitchell & Newell 2003).

Following this work with students, faculty generated a list of important concepts. These concepts were divided into 4 main areas. Using these areas, a concept matrix was constructed which included the basic modes of heat transfer and general levels of understanding, see Table 1-11 (Jacobi, Martin, Mitchell & Newell 2004). Questions were generated following the concept matrix and an initial version of the inventory was piloted

as a pretest with 42 students. The reliability coefficient for this data was $\alpha = 0.6$. Validity issues were addressed by conducting item analysis including calculating item difficulty and correlation coefficients. Students were also divided into quintiles based on their total inventory scores, and plots constructed for each item of quintile vs. percent correct on item.

Table 1-11: Concept Matrix for the Heat Transfer Concept Inventory

	Levels of Understanding		
Concept Areas	Physical Intuition	Mechanistic and Physical Description	Mathematical Models
Conduction Convection Radiation Control Volumes, energy balances	Recognition of the mode of heat transfer, basic characteristics	Relationships between heat concepts	Basic Laws governing Heat Transfer

1.2.13 Materials Concept Inventory (MCI)

The topic list for inclusion in the Materials Concept Inventory (MCI) was identified using course textbooks and syllabi (Krause, Decker & Griffin 2003). Topics were divided into expected prior knowledge that students should bring into a materials course (mainly chemistry and geometry concepts) and new course content knowledge students would be expected to learn during the materials course. A total of 30 questions were included, two on geometry topics, eight on chemistry topics, and 20 on new content. Initial distractors were written by faculty. Student generated distractors were elicited through student interviews and weekly, open-ended "intuition quizzes" that were given during lectures. These distractors were incorporated into the inventory.

An initial version of the inventory was used in 2002. Misconceptions were noted in the results and classified as prior (noted on the pretest), persistent (present on both pre- and posttests), or spontaneously generated (found on the posttest only). Limited gains of 15-20% were observed in most classes; however, one class which used some active learning strategies had an average gain of 38%. Developers hope that use of the MCI will help generate debate and change in the teaching of materials science.

1.2.14 Other Concept Inventories

Additional concept inventories in early stages of development include:

- *Electromagnetics Concept Inventory* which is composed of three exams: EMCI-Fields, EMCI-Waves, and EMCI-Fields and Waves. Questions focus on core content material and are mostly non-computational. The EMCI-Fields and EMCI-Waves are each 23 question multiple choice instruments designed for the first and second semesters respectively of a two-semester course. The EMCI-Fields and Waves is an integrated 25 question instrument designed to be used in a one semester course. (Foundation Coalition 2005)
- *Computer Engineering Concept Inventory* (CPECI) (Foundation Coalition 2005)
- *Electronics Concept Inventory* (ECI) {Foundation Coalition, 2005 #88
- *Thermodynamics Concept Inventory* {Foundation Coalition, 2005 #88}
- *Strength of Materials Concept Inventory* (SoMCI) (Foundation Coalition 2005)
- *Device Concept Inventory* (DVI) is a 50 question web based multiple choice instrument. (Skromme 2005)

A summary of the concept inventories presented here can be found in Table 1-12.

Table 1-12: Summary of Concept Inventories. Those being developed as part of the the Foundation Coalition are marked with a *.

Instrument	*# of Items*	*# of Topics or Sub-tests*	*Reliability*	*Validity*	*Scores Pre/Post*
Chemistry Concept Inventory I & II (CCI-I CCI-II) (Krause, et al. 2004)*	I: 20 II: 20	3 topics, 7 subtopics 3 topics, 10 subtopics	α=0.6803 α=0.5957 (post)	Significant positive correlations with course averages (ρ > 0.5). Student interviews confirm presence of misconceptions.	$\bar{x}$ = 24.7% / $\bar{x}$ = 44.5% $\bar{x}$ = 33.6% / $\bar{x}$ = 48.1%
Computer Engineering Concept Inventory (CPECI) (Foundation Coalition 2005)*		3 topic areas			
Conceptual Survey of Electricity and Magnetism (CSEM) (Maloney, et al. 2001)*	32	7 topic areas	$KR20 \approx 0.75$, (post)	Questions ranked by professors for reasonableness and appropriateness.	$\bar{x} \approx 27\%$ / $\bar{x} \approx 45\%$
Determining and Interpreting Resistive Electirc Circuit Concepts Test (DIRECT) (Engelhardt & Beichner 2004)	29		$KR20 = 0.7$	Expert panel consulted to assure adequate and appropriate topic coverage. Factor analysis conducted. Student interviews used to ascertain student interpretation of questions.	/ $\bar{x}$ = 36% high school / $\bar{x}$ = 44% university
Device Concept Inventory (DVI) (Skromme 2005)	50 web based				
Dynamics Concept Inventory (DCI) (Gray, et al. 2005)*	30	11 topics	$\alpha \approx 0.72$ (pre) $\alpha \approx 0.64 - 0.84$ (post)	Modified Delphi process used to select topic lists, student focus groups used to develop distractors and revise questions.	$\bar{x} \approx 32\%$ / $\bar{x} \approx 52\%$
Electromagnetics Concept Inventory (EMCI) Three versions available (Notaros 2002)*	Fields:23 Waves: 23 Fields and Waves: 25				

Table 1-12 continued:

Instrument	*# of Items*	*# of Topics or Sub-tests*	*Reliability*	*Validity*	*Scores Pre/Post*
Fluid Mechanics Concept Inventory (FMCI) (Martin, et al. 2004)*	27	10 topics	Limited to item analysis with point-biserial correlation at this time.	Topic list developed by faculty, compared to lists developed by students. Students were video taped discussing topics and questions.	
Force Concept Inventory (FCI) (Halloun & Hestenes 1985b)	36	6 topic areas	$\alpha = 0.86$ (pre) $\alpha = 0.89$ (post) Stable answers in test/retest and test/interview situations.	Professors and graduate students consulted for input on questions. Tests from physics students who had received A's were examined to look for mistakes that could be due to question formulation.	$\bar{x} = 51\% / \bar{x} = 64\%$ (calculus based) $\bar{x} = 38\% / \bar{x} = 52\%$ (algebra based)
Force and Motion Conceptual Evaluation (FMCE) (Thornton & Sokoloff 1998)	43				
Heat Transfer Concept Inventory (HTCI) (Jacobi, et al. 2004)*	30	12 topics	$\alpha = 0.6$ (pre)	Item analysis including difficulty, correlations, and graphical plots of item percent correct by quintiles.	$\bar{x} = 45\% /$
Materials Concept Inventory (MCI) (Krause, et al. 2003)*	30	Prior knowledge and New Content			
Signals and Systems Concept Inventory (SSCI) Discrete (DT) and Continuous (CT) time versions (Wage, et al. 2005)*	25	5 Subtests (CT) 6 Subtests (DT)			$\bar{x} \approx 40 / \bar{x} \approx 53$

Table 1-12 continued:

Instrument	*# of Items*	*# of Topics or Sub-tests*	*Reliability*	*Validity*	*Scores Pre/Post*
Statics Concept Inventory (SMCI) (Steif 2004, Steif, et al. 2005)	27	8 topics	$\alpha = 0.89$	Predictive validity assessed by comparing inventory performance to course exam scores. Construct validity demonstrated with confirmatory factor analysis.	/ x = 52% – 77%
Strength of Materials Concept Inventory (SoMCI) (Foundation Coalition 2005)*		7 topic areas			
Test of Understanding Graphs in Kinematics (TUG-K) (Beichner 2004)	21	7 topic areas	$KR20 = 0.83$ (post)	15 instructors were asked to complete the test and make comments and criticisms. Significant increase in test/retest scores from pre to post kinematics laboratory activity.	/ $\bar{x}$ = 40%
Thermal and Transport Science Concept Inventory (Olds, et al. 2004)		10 topics			
Thermodynamics Concept Inventory (Foundation Coalition 2005)*	30	6 topic areas			
Wave Concept Inventory (WCI) (Roedel, et al. 1998)*	20 (34 possible correct answers)				$\bar{x} \approx 32\%$ / $\bar{x} \approx 38\%$

1.2.15 Common Themes

Though these concept inventories are being developed for a wide variety of disciplines, there are many common themes among them.

- Topics to be included are often determined by groups of experienced instructors in the field, either informally or more formally using Delphi methods. In some cases students have been enlisted to help identify important topics. Often topics are chosen not only for importance, but also for being often misunderstood or difficult to teach.
- The concept inventories that have been devised are all multiple choice instruments, most of which have only a single correct answer per item. Most are only available in a pencil and paper format, but some are being developed for online administration.
- The questions are concept focused and are mostly or entirely non-computational. This is critical to the design of the concept inventories. The questions are not intended to be answerable using memorized computational skills, equations or algorithms.
- Effective item distractors are gleaned from experienced teachers and from students themselves. Distractors can be found in remarks from student interviews, mistakes made on open ended versions of inventory questions, or from student answers on other assessments such as in class quizzes and exams. Also, prior research on student misconceptions and errors has been used to generate response sets.
- Multiple iterations of testing and revising are used to improve question clarity and common psychometric measures such as reliability and discrimination.
- Questions on the concept inventories frequently appear to instructors to be easy or trivial. Instructors are frequently surprised by the low scores demonstrated on the

concept inventories. Together these two factors can be very motivating and effective in creating instructional change.

- In every subject area in which comparisons have been made, traditionally taught courses show consistently lower concept inventory scores and gains than those taught using interactive engagement and/or concept focused approaches.
- The majority of the concept inventories are attempting to cover a much broader topic area than the FCI. This can make it more difficult to achieve adequate topic coverage and to interpret results.
- Principal component factor analysis yields a large number of factors, often explaining small amounts of variation.

Issues noted by Martin, et al (2003) in developing the Fluid Mechanics Concept Inventory but which apply to most concept inventory assessments include:

- Engineering courses involve both understanding the concepts and using the concepts to solve problems. How should the development of these skills be assessed?
- There are marked differences between how students understand concepts and how instructors assume students understand the material.
- There are very wide gaps in the use of language between students and instructors. Instructors use technical terms in order to be precise in their meaning. Students often have only vague understandings of these terms and are uncomfortable using them. More often they use everyday terminology and associate a variety of meanings to both technical and non-technical terms. It may be helpful to include student based descriptions in concept inventories.

- Students often miss subtleties that instructors feel are very important. How should these be included within the concept inventories?

These common themes and issues will be very important and helpful to the development of any concept inventory, including one for introductory statistics. In order to make use of known misconceptions and difficulties related to statistics concepts, a review of the statistics education literature was undertaken. The most applicable results are presented next.

1.3 Statistics Education Research

1.3.1 Probabilistic Thinking/General Reasoning Frameworks

An article by Kahneman, Slovic, and Tversky (1982) details several common, informal reasoning frameworks that people use when thinking about probabilities of events , including representativeness, availability, and adjustment and anchoring. People using the representative heuristic judge the probability of the occurrence of an event or sample based on the "degree to which it is (i) similar in essential properties to its parent population; and (ii) reflects the salient features of the process by which it is generated" (Kahneman & Tversky 1982 p. 33). The first element would include, for instance, the proportion of the population having a certain characteristic or the mean value for some characteristic. The second would include randomness or consistency. Consider for example, the following sequences of five coin tosses: HHHHH, HTHTH, HTHHT. The third sequence would be judged to be the most probable sequence because it has roughly the same number of heads and tails and has the appearance of randomness, when in fact each sequence is equally likely.

Important conclusions drawn from the many examples presented include that naive students think that sample size is irrelevant in making determinations between two samples (that is, one is as likely to obtain 70% heads in 10 tosses of a coin as in 100 or 1000 tosses) and that they expect that a process will be represented both globally and locally within a sample/event. This type of thinking gives rise to common misconceptions such as the gamblers fallacy and the belief in the "law of small numbers, which asserts that the law of large numbers applies to small numbers as well" (Tversky & Kahneman 1982a p. 25).

Bar-Hillel (1982) studied variations on the "Maternity Ward" question posed by Kahneman and Tversky (1972):

> A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50% of all babies are boys. The exact percentage of baby boys, however, varies from day to day. Sometimes it may be higher than 50%, sometimes lower.
>
> For a period of 1 year, each hospital recorded the days on which (more/less) than 60% of the babies born were boys. Which hospital do you think recorded more such days? (Bar-Hillel 1982 p. 81)

The proportion of boys born was varied from 60% to 100%. The results found by Kahneman and Tversky were replicated when the proportion was 60%; the most common answer was the same for both hospitals. But when the proportion exceeded 70%, the most common answer was the smaller hospital. This suggests that sample size becomes relevant to students once the sample result is perceived as not representative.

The availability heuristic is used when people judge the frequency or probability of events "by the ease with which instances or occurrences can be brought to mind" (Tversky & Kahneman 1982b p. 11). Since availability is affected by many factors, this heuristic leads to predictable biases and errors. Familiarity, recent exposure, and ease of

imagining examples all can impact the perception of frequency and can cause over- or underestimation of events.

The adjustment and anchoring heuristic is used when people make estimates of frequencies or probabilities by making adjustments to some initial, anchor value. This leads to estimates which are biased toward the anchor value. One important application of this is in the evaluation of compound events, people tend to overestimate the probability of conjunctive events (event A and event B) and underestimate the probability of disjunctive events (event A or event B).

Konold has offered an additional framework for understanding reasoning about probabilities which he has termed the *outcome approach* (Konold 1989, Konold, Pollatsek, Well, Lohmeier & Lipson 1993, Konold 1995). Using this approach, students do not think of probabilities in terms of distributions in a sample or population, but in terms of predicting the result of a single trial. When asked to choose the most likely sequence of heads and tails from the sequences discussed above, students reasoning from this perspective would choose "all sequences are equally likely" because, as each sequence could occur, they cannot rule a particular sequence out. Within this framework, students translate probability statements into "yes/no decisions" (Konold 1995, p.3). This is illustrated with a weather forecast problem. Students perceive probabilities greater than 50% as a yes prediction; it will rain. Probabilities less than 50% are seen as a no prediction; it won't rain. Probabilities equal to 50% are seen as simply a lack of information, an inability to predict the outcome.

An important result of the Konold et al. (1993) study was that when they changed the wording of the coin tossing question slightly, from 'which sequence is most likely' to

'which sequence is least likely', many of the students who answered the first question correctly, answered the second incorrectly. It seemed that the students suddenly switched their framework of reasoning from the outcome approach to the representativeness heuristic.

The problem of identifying student misconceptions and preconceptions is broader than identifying a general framework from which students reason. From interviews with students, Konold et al. (1993) found that students would not only apply different reasoning frameworks to slightly different problems, but often would switch between frameworks on the same problem as they worked their way through it. They also found that in addition to the general frameworks of reasoning, students relied on maxim-like beliefs to make decisions such as "the coin has no memory", "heads and tails occur about equally often", and "outcomes of coin flipping are unpredictable" (p. 408). These beliefs can be contradicting. The belief that heads and tails should occur equally often could lead a person to expect that tails are more likely after a sequence of several heads and contradicts the belief that the coin has no memory and that the outcomes are unpredictable.

Important conclusions from this work are that student reasoning in statistics and probability is very complex, based on multiple beliefs and frameworks of reasoning. These are often conflicting and put to use almost simultaneously on the same problem. Additionally, great care must be exercised when trying to assess student understanding based on their response to multiple-choice questions. In Konold's research, many of the students were able to correctly answer the first question even though they were not using

correct probabilistic reasoning, and were then subsequently unable to correctly answer the second very similar question.

Fischbein and Schnarch (1997) studied the evolution of several of these misconceptions across multiple age groups, from grade 5 to college age. While no consistent pattern emerged among the misconceptions, the prevalence of the misconceptions at all age groups is notable. Twenty students from each grade (5th, 7th, 9th and 11th) and 18 undergraduate prospective mathematics teachers were given a 7 item, multiple-choice questionnaire. None of the students had received any probability instruction. Each question targeted a specific misconception. The misconceptions and their frequency are summarized in Table 1-13.

Table 1-13: Frequency of common probabilistic misconceptions found by Fischbein and Schnarch (1997)

Misconception	*Question Context*	*% demonstrating main misconception 5th, 7th, 9th, 11th, Undergraduate*
Representativeness	Choosing most likely winner of two lottery tickets: (1,2,3,4,5,6) and (39,1,17,33,8,27)	70, 55, 35, 35, 22—Second ticket more likely to win
Negative and Positive Recency Effects (Gambler's Fallacy)	Coin tossing, what is the most likely outcome of a fourth flip following 3 heads	35, 35, 20,10, 0—Tails more likely 0, 5, 0, 0, 6—Heads more likely
Compound and Simple Events	Two dice are rolled simultaneously, which is the most likely outcome 6-6 or 5-6	70, 70, 75, 75, 78—Both outcomes are equally likely
Conjunction Fallacy	Dan is described as an aspiring doctor…Which is more likely? Dan is a student or Dan is a student of the medical school?	85, 70, 80, 40, 44—Dan is student of the medical school
Effect of Sample Size	A. Maternity Ward problem as in Kahneman and Tversky (1972) B. Likelihood of getting 2 out of 3 heads compared to getting 200/300 heads when flipping coins	A. 10, 30,70, 80, 89—Equal in the two hospitals B.30, 45, 60, 75, 44—Equal in both samples
Availability Heuristic	Number of possibilities for 2 vs. 8 member committees chosen from 10 people	10, 20, 65, 85, 72— 2 member committees greater than 8 member committees
Effect of Time Axis (Falk Fallacy): An event cannot act retroactively on its cause	2 white and 2 black marbles are placed in a box. A: a white marble is drawn, which is more likely for second draw? B: One marble is drawn without looking, second marble drawn is white. Which is more likely outcome for the first draw?	5, 30, 35, 70, 44—answered A correctly and B incorrectly

In a meta-analysis of studies which considered whether or not sample size was taken into account, Sedlmeier and Gigerenzer (1997) attempt to explain apparently conflicting evidence about human intuition and the effect of sample size. They propose that two types of tasks have been studied: tasks that ask about frequency distributions and tasks that ask about sampling distributions. The "Maternity Ward" question presented above (Kahneman & Tversky 1972) is an example of a sampling distribution question. A similar question can be posed in the context of a frequency distribution: which hospital is more likely to have 60% of boys born on a single day? In the sampling distribution form, the required knowledge is that smaller samples have greater variability than larger samples. The frequency form of the question tests the understanding of the "empirical law of large numbers which … states that a proportion from a larger sample is a more accurate estimator of the population proportion than one from a smaller sample" (Sedlmeier & Gigerenzer 1997 p. 37).

The first part of the analysis compared two sets of studies. In these studies, students were asked the maternity ward question or its equivalent in either the frequency or sampling distribution form. Three response alternatives were given: the smaller sample, the larger sample, or no difference. Six studies asked the frequency form and 29 the sampling form. Only one study in the sets asked both forms of the question. Students did much better on the frequency form of the question, and there was almost no overlap in the distribution of the results. See Table 1-14. In the second part of the analysis, other studies that asked frequency distribution questions that had no sampling distribution analog were considered. In general, these studies found that students took sample size into account.

Table 1-14: Meta-analysis results from Sedlmeier and Gigerenzer (1997) comparing two forms of questions concerning sample size.

	Frequency Distribution Form	*Sampling Distribution Form*
Number of Studies	6	29
Range of % Correct	56-87%	7-59%
Median of % Correct	76%	33%

1.3.2 Averages/Measures of Central Tendency

Most of the research that has been carried out about student understanding of averages has been with school age children, so it is not clear how much can be generalized to an undergraduate population. However, if these conceptions are present among middle and especially high school students, there is a good chance that some students will bring these conceptions with them to the undergraduate classroom. In addition, much of this research has been concentrated on describing the development of student understanding. These descriptions generally paint a picture of an incomplete understanding or multiple levels of understanding with upper levels being more complete or complex. In general, errors of understanding or specific misconceptions are not documented.

One study, conducted by Watson and Moritz (1999), included a wide range of K-12 students in Australia (n=2250), including a relatively large group of 11th graders (n=164). Students were asked to complete four multiple choice and short answer questions as part of a larger questionnaire on data and chance. Responses were analyzed both quantitatively and using NUD*IST™ language analysis software and categorized as representing one of four levels of reasoning. They documented a more complete and complex understanding of average and the three concepts of mean, median, and mode among older students. However, even among 11th graders, the majority of students had

difficulty distinguishing between the three concepts in an applied situation. They also found that students tend to think about an average value as a middle value or most common value rather than a representative value.

Mokros and Russell (1995) interviewed twenty-one 4th, 6th, and 8th graders using open ended questions about the concept of average. In two of the questions, the students were asked to construct a distribution of data which would have a specified average value. One question asked them to interpret a distribution of data, and one involved weighted averages. The questions were posed in a context that was familiar to students, such as allowances or prices of chips. This form of questioning is very different from questions typically seen that deal with average values and was quite challenging for many of the students.

The 45 minute individual interviews were video-taped and transcribed. Analysis of the data resulted in five categories of approaches to problem solving, described briefly here:

- Average as mode: Used the mode most often to address problems, thought of in terms of "most" but not as representative of the data as a whole (most commonly found among younger students).
- Average as algorithm: Thought of average in terms of a process to be carried out, unable to interpret solutions, often confused concepts of data, total, and average.
- Average as reasonable: Viewed the average as a way to think about data, judged reasonableness based on life experiences, viewed average not as a single number but as an estimate that may take on several values.

- Average as midpoint: believe that mean and middle are essentially the same, midpoint viewed alternatively as median, middle of range, middle of X-axis. Symmetry figured prominently in reasoning. Had difficulty interpreting or constructing non-symmetrical distributions.
- Average as mathematical point of balance: Looked for a point of balance to represent the total data, had an understanding of the different concepts of data, total, and average, were able to work from average to total and average to data and total to data.

The students interviewed predominately used only one style of approach and could be classified in one of these five groups. One misconception identified among students thinking with the balance point approach is that they believed the data on each side of the average must sum to the same total, focusing on the value of the data point rather than the distance from the mean. The authors note that they have seen this type of reasoning among teachers as well. One other important point made by the authors is that the average was only seen as a representative of the data set once the data set was conceived of as an entity itself, not only a collection of individual values.

Seven properties of the concept of average were identified and studied in 8, 10, 12, and 14 year olds as part of a developmental study conducted by Strauss and Bichler (1988). The seven properties are:

A. The average is located within the range of the data values
B. The sum of the deviations from the mean is zero
C. The average is influenced by values other than the mean
D. The average can take on a value not contained in the data set

E. The average can be a non-integer value that has no physical counterpart (e.g. the average number of discrete objects)

F. When calculating an average, any values of zero in the data set must be taken into account

G. The average is a representative of the data set

Twenty students in each age group from Israel were interviewed individually. The students were presented with 32 tasks focusing on the 7 properties. The tasks differed in whether they asked about continuous or discontinuous events, and in how they were presented: in story form, in a concrete form (i.e. with physical items to manipulate as they worked), or in a numerical form. The tasks were all very similar to the following example: "One day children in a class brought books for their class library. They passed out all of the books the children brought, and it turned out that each child got two books. Does this mean that someone originally brought 2 books? Why do you think so?" (Strauss & Bichler 1988 p. 70)

ANOVA with repeated measures was performed with four age groups, six properties (property E omitted), two media (story or concrete), and two quantity types (continuous or discontinuous) as the main effects. Significant main effects were age and property. Older students outperformed younger students at each age group, and properties A,C, D were easier than B, F, G. 14 year-olds performed well for most properties, but had the most trouble with properties B ($\bar{x} = 1.31, sd = 0.36$) F ($\bar{x} = 1.61, sd = 0.47$), and G ($\bar{x} = 1.63, sd = 0.46$). The maximum score for each property was 2. For the three more difficult properties, the most common justification given for incorrect answers were:

Property B: the problem could not be solved because of a lack of information about the individual data points or lack of information about the total sum of the data;

Property F: zeros did not need to be taken into account since they did not change the sum, did not have to be considered when added or subtracted; and

Property G: many different reasons were given, did not give the impression that they understood the average as a representative of the group.

Zawojewski and Shaughnessy (Zawojewski & Shaughnessy 2000) identified several possible misconceptions based on 7th and 11th grade data from the National Assessment of Educational Progress. These misconceptions indicate a belief in the superiority of the mean over the median. This includes ideas that the median is not a representative value or that the mean is a more precise value. Students do not seem to understand the relative advantages of each or when one might be more appropriate. The data also indicated that students have difficulty distinguishing between the three measures: mean, median, and mode.

Pollatsek, Lima, and Well (1981) interviewed undergraduate students and found that many had only an understanding of how to compute a simple mean and lacked any conceptual understanding of the mean. In particular, students had difficulty understanding and making use of a weighted mean. Asked to find the overall mean of two groups of unequal size given only the mean for each group, many students responded by either finding the simple mean of the two averages, adding the two means and dividing by the total number of the combined groups (resulting in a smaller number than either mean), or by stating that they could not find the overall mean without knowing what the individual data values were.

Misconceptions identified by Garfield (2002) include believing that the average is the most common number, confusing the mean with the median, believing that groups should always be compared based on the differences in their averages, and having a formulaic understanding of the computation of the average without regard for outliers. Herman (1997) considered the distribution of answers to multiple choice items on classroom examination questions from 101 undergraduate students. His results support these same ideas; he also notes that students confuse measures of central tendency with measures of spread.

1.3.3 Sampling Distributions

The effect of representativeness is evident in student understanding of sampling distributions. Kahneman, et al.(1982) asked subjects to produce sampling distributions for three sample sizes (n=10, 100, and 1000) for each of three scenarios (e.g. heights of men). In this task the resulting distributions were indistinguishable for each sample size. In addition, the samples were flatter than would be expected for the correct distribution sample size n=10.

Sedlmeier and Gigrerenzer (1997) suggest that when asked to construct sampling distributions students actually construct frequency distributions. They asked one group of 55 participants to construct frequency distributions for heights of men for sample sizes of n=20 and n=200 and another group of 56 participants to construct sampling distributions for the same. The median distributions for all four cases were identical. They also cite Well, Pollatsek, and Boyce (1990) who found that half of their subjects who had incorrectly completed a sampling distribution task recalled it as a frequency distribution task. Only 3 of 21 participants recalled it as a sampling task.

Chance, Delmas, and Garfield (2004) echo these findings. They provide an overview of a five stage study as part of an ongoing program of research on sampling distributions. The classroom based research investigated the impact of computer software tools on student understanding of sampling distributions and used both quantitative assessment data and interview data. The learning, teaching, and assessment tools were developed and tested for college level introductory statistics courses. A portion of the diagnostic instrument was included as an appendix to the paper.

The first two stages of the study focused on the use of the simulation software. Students experimented with changing the shape of the parent population and the sample size and examined the changes to the resulting sampling distribution. The focus was on developing an understanding of the Central Limit Theorem for the sample mean. Student responses were categorized and used to make improvements to the software and the learning activities. Improvement to student understanding was observed but misconceptions persisted for many of the students.

In order to determine if inadequate understanding of prerequisite ideas was part of the problem, the third stage of the study involved a conceptual analysis based on observations by the researchers, input from colleagues, and student performance on assessment items. This series of analyses produced detailed lists of prerequisite knowledge for understanding sampling distributions, what students should know about sampling distributions, what they should be able to do with their knowledge, and common misconceptions. The misconceptions identified are: students

- believe sampling distribution should look like the population (for sample size $n>1$).

- think sampling distribution should look more like the population distribution as the sample size increases (generalizes expectations for a single sample of observed values to a sampling distribution).
- predict that sampling distributions for small and large sample sizes have the same variability.
- believe sampling distributions for large samples have more variability.
- do not understand that a sampling distribution is a distribution of sample statistics.
- confuse one sample (real data) with all possible samples (in distribution) or potential samples.
- pay attention to the wrong things, for example heights of histogram bars.
- think the mean of a positive skewed distribution will be greater than the mean of the sampling distribution for samples taken from this population (Chance, et al. 2004 p. 302).

The fourth stage of the study resulted in the creation of a developmental model for student understanding of sampling distributions. Students enrolled in a graduate level introductory statistics course were interviewed individually and asked open ended questions. The developmental model has five levels that describe student reasoning: idiosyncratic reasoning, verbal reasoning, transitional reasoning, procedural reasoning, and integrated process reasoning.

The fifth stage of the study was conducted to validate the model. A nine item, multiple choice, diagnostic instrument was developed and administered to 105 undergraduates enrolled in introductory statistics. Nine senior statistics majors also completed the instrument. Answers to the questions were variable and often inconsistent.

Students fared worse on the graphical items than on the non-graphical items. Only 49% made consistent choices between the two types of problems.

Nine of these students consented to participate in interviews. Four additional students were chosen from a master's level introductory statistics course to increase the variety of levels of statistical reasoning in the interview pool. The data from the assessment instrument and the interviews did not support the idea that students develop linearly through the levels of the developmental model, but that development occurs along several dimensions.

Saldanha and Thompson (2002) put forth two ways of conceiving of samples: an additive conception of sample and a multiplicative conception of sample. In the first, samples are simply seen as subsets of the population and multiple samples are simply multiple subsets. Students with this view often confused the number of samples with the number sampled in a resampling process. A multiplicative view sees a sample more as a quasi-proportional version of the population. This view is cognizant of other possible outcomes of the sampling process and requires moving between multiple levels of reasoning: the sample at hand, its relationship to other possible samples, and how it represents the population. In a teaching experiment conducted with 11th and 12th grade students enrolled in a non-AP statistics course, students were found to fall on a spectrum between these conceptions of sample. Most of the students, however, fell much toward the additive side. The authors suggest targeting these ideas in instruction.

1.4 Other Instruments for Statistics Assessment

The Statistical Reasoning Assessment (SRA) is a 20 question multiple choice instrument designed to assess the student reasoning behind the correct and incorrect

choices selected (Garfield 2003). Each response option includes a rationale statement and students are asked to select the response that best matches their own thinking. The instrument is scored on 16 scales: 8 categories for correct reasoning and 8 categories for incorrect reasoning. The scaled scores range from 0 to 2 in each category. The scores may be summed to get total scores for correct and incorrect reasoning.

Questions on the instrument address reasoning in the following areas:

- Understanding and computing probabilities
- Averages
- Independence (in the context of coin flipping only)
- Sampling Variability (one maternity ward style question, one comparing two groups of equal size)
- Correlation vs. Causation
- Two way tables
- Sample size

Specific misconceptions assessed are:

- Representativeness
- Outcome approach
- Law of Small Numbers
- Correlation implies Causation
- Equiprobability bias (events are viewed as equally likely; in the context of dice throws only, one question sequential and three on simultaneous dice throws)
- Groups must be the same size in order to compare them
- Good samples must represent a high percentage of the population

- Misconceptions involving averages, including confusing mean and median and failure to consider outliers (Garfield 2003 p. 27)

The SRA does not address any aspect of statistical inference, issues of graphical interpretation of data, or specific measures of center and spread.

Total correct scores (not the scaled sub scores) were found to have very low correlations with other course outcome measures such as final scores, project scores, and quiz totals. The items were found to have low inter-item correlations. Specific reliability coefficients are not reported. Instead, test-retest reliability was checked using the scaled sub scores. Thirty-two students enrolled in an assessment course for pre-service teachers were given the SRA and retested one week later. The test/re-test reliability is reported to be 0.70 for the correct reasoning total score and 0.75 for the incorrect reasoning total score. It is not reported what kind of statistics training these students had, nor what the overall performance of this group was on the SRA.

The SRA was used in a study comparing college students from the United States (n=267) and Taiwan (n=245) at the end of an introductory business statistics course. The two groups had very similar outcomes on the 16 scale scores. Both groups had the lowest correct reasoning scores in the areas of probability and sampling variability and the highest incorrect reasoning scores in the areas of equiprobability bias. ANOVA was carried out on the total correct and incorrect reasoning scores to check for differences due to gender and country. For correct reasoning, students from Taiwan scored significantly higher than those from the United States. No significant difference was found for gender or for the interaction of gender and country. Incorrect reasoning scores were significantly different for gender and country effect. Males had lower misconception scores than

females in both countries, and the United States students had higher misconception scores. The interaction was not significant.

The possible score ranges are not reported for the total correct and incorrect reasoning scores, so it is difficult to interpret how well the students did on the instrument overall. The mean scores for correct reasoning were around 21. If this is out of 40 possible points (2 points per item) then this would be approximately 50%. The mean scale score for the United States students for correct reasoning was 1.14 out of 2.

The SRA is unique in that it provides two distinct scores, a correct reasoning score and a misconceptions score, as opposed to the correct only scoring that is typical of most assessment instruments. The limitations of the SRA, as pointed out by the author, are that the content coverage is a small subset of the statistics curriculum and the instrument has not been demonstrated to have high reliability. Test-retest reliability of 0.70 for total correct answer scores and 0.75 for incorrect reasoning scores are reported.

The Quantitative Reasoning Quotient (QRQ) is a revision of the SRA (Sundre 2003) that consists of the same questions edited into a format that is easily scored by machine. Whereas the SRA presents some questions that ask students to select as many responses as they agree with, on the QRQ this style of question was converted to multiple questions and the students were asked to agree or disagree with the statements. The resulting QRQ is a 40 item instrument. The QRQ scoring method was slightly modified as well to have 11 scales for correct reasoning and 15 scales for incorrect reasoning.

The modifications made for the QRQ were done in an effort to increase the reliability of the instrument by increasing the number of questions (by forced response to each option), to increase the ease of scoring making it easier to administer to larger

groups of students, to capture more information by scoring more of the responses, and to enlarge the set of reasoning strategies assessed. As noted, however, the questions posed are essentially the same. So while a few more strategies and misconceptions are specifically scored, the subset of statistics concepts addressed is not significantly greater.

The QRQ was administered to 804 sophomore students in the spring of 2002. Students were randomly selected based on their student ID as part of a campus wide "Assessment Day". It should be noted that students may or may not have had any statistical training. The reliability was found to be 0.62. Minor revisions were made to the QRQ and it was administered to 1,083 incoming freshmen in the fall 2002. Again, the students were randomly selected to take the assessment as part of a required orientation, the students had had no college training in statistics and any possible prior statistics training was unknown. The reliability for this administration was found to be 0.55. The drop in alpha is attributed to increased random error of the inexperienced students.

A survey of items to measure understanding of variation for K-12 students was developed by Watson, Kelly, Callingham, and Shaughnessy (2003). Questions in the survey focused on "sampling variation, displaying variation, chance variation, describing/measuring variation, and sources of variation (explanations, inferences)" (Watson, et al. 2003 p. 3). Based on pilot results with 58 4th and 10th graders, a core set of questions was chosen for use with 3rd graders, with other questions added for each successive age group. Five component areas were identified in constructing the questionnaire: basic chance, graph and table reading, variation in chance, variation in data/graphs, and variation in sampling. The questions were a mixture of open-ended and multiple choice items, with explanations requested for multiple choice items. The revised,

16 question (some multi-part) instrument was administered to 746 students in grades 3 (n=177), 5 (n=183), 7 (n=189), and 9 (n=197) in Australia. A coding scheme was developed and 44 sub-parts were coded. Some items were coded right/wrong and some were coded with a hierarchical scheme.

The scoring rubric and reported results focus on an incremental understanding of variability rather than on incorrect interpretations. For example, students were asked to fill in a table to predict the number of times each face of a die would turn up if it were thrown 60 times. The answers were coded from 0 to 4 in the following manner:

0 - Inappropriate response: Included answers which did not add to 60, had a single number greater than 21, or misinterpretation of the question

1 - Answers summed to 60 but had idiosyncratic reasoning for the variation

2 - Answers reflected strict probabilistic outcomes or with unusual variation but with reasoning that reflected some understanding of the context

3 - Too wide or too narrow variation, but appropriate reasoning

4 - Appropriate variation and reasoning

Selections of the questions are described along with the scoring rubrics and example answers are given.

The data were analyzed using a one-parameter item response model (Rasch model). The authors provide a variable map with ability plotted on the left side of the logit scale and item difficulty for each response code on the right. The authors note that the 5 components they had identified were satisfactorily distributed along the scale and that the item difficulty distribution matched that of the student ability. From this they conclude that the scale can be used to measure student achievement on each of the sub-

components as well as overall understanding, and that the scale was able to measure along the full range of student ability.

Further, the authors identified four levels of increasing understanding of variation and their corresponding threshold values. These are summarized in Table 1-15.

Table 1-15: Levels of understanding of variation (Watson, et al. 2003).

Level of Understanding	*Description*
Level 1: Prerequisites for Variation	Exhibits limited skills in reading tables and graphs, very limited recognition of variation, likely to use personal stories to justify responses.
Level 2: Partial Recognition of Variation	Responses do not indicate understanding of chance or variation, focuses on patterns when interpreting graphs, has difficulty expressing the meanings of terminology (e.g. sample, random, variation).
Level 3: Applications of Variation	Exhibits improved graph reading skills, focuses on some appropriate aspects of concepts while overlooking or being misled by others, gives more structured definitions to important terms.
Level 4: Critical Aspects of Variation	Summarizes graphical information in appropriate ways, acknowledges variation, demonstrates sophisticated understanding of key terms, identifies bias, acknowledges the role of chance in variation, and integrates different components of the concepts.

1.5 Test Theory Background

Several important measurement models have been developed over the last century. The analysis of the Statistics Concept Inventory presented in this dissertation makes use of multiple models, including Classical Test Theory, Factor Analysis, and Item Response Theory. This section will provide a brief overview of the models and some background material for the analyses to follow.

1.5.1 Classical Test Theory Model

Under the Classical Test Theory (CTT) model (sometimes called the true score model), a measure or test score Y is a function of two random and independent components: the true score Θ and measurement error ε. The random variable Θ is continuous and is assumed to represent the latent trait that is being measured. Under CTT, the test must be unidimensional, i.e., it only measures one construct. In the SCI case, this construct is conceptual understanding of statistics. Under this model, $Y = \Theta + \varepsilon$. The measure Y is the total score on the instrument comprised of parallel test items, X_i: $Y = \sum_{i=1}^{k} X_i$. Items are parallel if they have equal means, variances, and correlations with any and all other variables. For an individual, the true score is assumed to be constant. Since Θ can not be directly observed, the measure Y gives an estimate of this value. The measurement error is assumed to have a mean of zero, so that true score Θ is equal to the expected value of the measure, Y.

The simplicity of the model has made it widely applicable and a large body of test theory has been built up around it. Despite the simplicity of the model, CTT has led to the development of many important psychometric measures including estimation methods for reliability, standard error of measurement, item difficulty and item discrimination. There are important limitations that should be considered, however.

CTT provides information at the whole test level, not the individual item level, which limits the conclusion that can be drawn about individual test items or groups of items.

The machinery of CTT that is used in the development and evaluation of tests (such as the reliability estimates, item difficulties, item discrimination, etc.) are sample

dependent. They will vary for different samples from the population. Therefore, generalizations that can be made are limited to populations that are very similar to the sample from which the statistics were derived.

CTT relies heavily on the concepts of parallel items and parallel forms which in practice are difficult to achieve.

There are no provisions in the model to allow for differences in sensitivity, measurement error, or reliability at different points along the Θ distribution (e.g., the test works equally well at low, middle, and high ability levels) (Hambleton & Swaminathan 1985).

To overcome these limitations, other test models have been developed. Before addressing these models, however, we will look at one of the key ideas from classical test theory, reliability.

1.5.2 Reliability

The reliability of an instrument is defined as the amount of the total test variation that is attributable to the variation in the true score vs. how much is due to measurement error. In this sense, reliability gives us an idea of how reproducible the measure is. If the reliability is high, there is little measurement error impacting the results. The reliability of a measure Y, $\text{Rel}(Y)$, can be defined in several equivalent ways:

$$\text{Rel}(Y) = \rho^2(Y, \Theta)$$

$$= \rho(Y, Y_j)$$

$$= \frac{\text{Var}(\Theta)}{\text{Var}(Y)} = \frac{\text{Var}(\Theta)}{\text{Var}(\Theta) + \text{Var}(\varepsilon)}$$

where Θ represents the true score, ρ the correlation of two items, Var the variance, and Y, Y_j two parallel measures (McDonald 1999). As defined by the first relationship, reliability is a measure of how well the observed test score correlates with the true score. A good test would of course need to be highly correlated with what it claims to be measuring. The second relationship defines reliability as an estimate of the average correlation of the test with all possible other parallel tests (Nunnally 1967). The last relationship in the equation above shows the relationship between error variance and true score variance.

There are a variety of methods for estimating reliability; the most recognized are test-retest methods, parallel forms, and internal analysis. Test-retest methods involve administering the same form of a test to a group of examinees twice, with a lapse of time in between. The scores from the two administrations are then correlated. Parallel forms methods require constructing two alternative, non-overlapping, parallel forms of the test and administering both to the same set of examinees. The scores on the two forms are correlated and this is used to estimate the reliability. These two methods are difficult to implement in practice due to carry-over and learning effects and difficulties constructing alternate forms and insuring that they are in fact parallel.

The third method, internal analysis, looks at the relationship between the individual items on the test. While a variety of methods have been proposed, the most commonly used are the Kuder-Richardson formula 20 (KR-20) (Kuder & Richardson 1937) and its generalization, coefficient alpha α (Guttman 1945, Cronbach 1951). For dichotomous items, the two are equivalent.

Coefficient alpha, α, is a widely used index that estimates the reliability of an instrument since true scores are not known. Coefficient alpha can be calculated for an instrument $Y = \sum_{i=1}^{k} X_i$ where X_i are parallel items, that is $Var(X_i) = Var(X_j)$, and $Cov(X_i, X_j)$ are equal for all items. Then alpha can be defined and interpreted in multiple ways. The standard definition of alpha is given by:

$$\alpha = \left(\frac{k}{k-1}\right)\left(\frac{\text{Var}(Y) - \sum \text{Var}(X_i)}{\text{Var}(Y)}\right) = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum \text{Var}(X_i)}{\text{Var}(Y)}\right) \qquad (2.)$$

Using covariance algebra to expand $\text{Var}(Y)$, we have

$$Var(Y) = \text{Var}\left(\sum_{i=1}^{k} X_i\right) = \sum_{i=1}^{k} \text{Var}(X_i) + 2\sum_{i=j+1}^{k}\sum_{j=1}^{k} \text{Cov}(X_i, X_j) \qquad (3.)$$

Substituting this into equation 2, we have the equivalent form

$$\alpha = \left(\frac{k}{k-1}\right)\left(\frac{2\sum_{i=j+1}^{k}\sum_{j=1}^{k} \text{Cov}(X_i, X_j)}{\sum_{i=1}^{k} \text{Var}(X_i) + 2\sum_{i=j+1}^{k}\sum_{j=1}^{k} \text{Cov}(X_i, X_j)}\right) \qquad (4.)$$

An alternative version of alpha is given by

$$\alpha = \frac{k\overline{\rho}(X_i, X_j)}{1 + (k-1)\overline{\rho}(X_{i,} X_j)} \qquad (5.)$$

where $\overline{\rho}(X_i, X_j)$ is the average correlation between the item pairs X_i, X_j. In each version, we can see from equations 4 and 5 that if the test items are not correlated, alpha will be small. If the items X_i, X_j are independent, then the $\text{Cov}(X_i, X_j)$ and subsequently their correlation will be zero. Therefore, alpha gives us a sense of how "dependent" the items are as a group, with higher dependence or correlation resulting in larger values of alpha.

From the perspective of the whole test, we want there to be dependence among the items because they are functioning to give us a measure of some construct to which we believe they are all related. It is this dependence, this relationship, between the items that makes it reasonable to look at a total score *Y*.

We can also see from equations 4 and 5 that alpha can be made large by either strong inter-item correlations or by weaker correlations among many items (large *k*). If there are strong inter-item correlations, then the items are behaving in much the same way with little influence of random error, therefore giving us a reliable, reproducible measure. If the items have weak inter-item correlations, then if there are enough of them, the aggregate of information "strengthens the signal" so that the influence of random/measurement error is reduced.

Ideally, coefficient alpha should be used under the classical test theory model, where the test is unidimensional and comprised of parallel items. Alpha is also reasonable to use with tests that are "essentially unidimensional"—that is, the items share a general factor but may be subdivided into groups which share additional commonalities. Since the strict ideal cases are hard to achieve in practice, this case is noteworthy. Less ideal cases will result in lower estimates of reliability.

Ideally, values of coefficient α will be as close to 1 as possible. In practice, values of 0.8 or more are usually considered adequate {Nunnally, 1967 #143}. It should also be noted that coefficient α gives a lower bound for test reliability. Tests which are not strictly unidimensional and tests which have items that have unequal factor loadings will always have a coefficient α which is strictly less than the true reliability of the test. Other measures of test reliability that can be used include coefficient omega, ω, which is

calculated using the item factor loadings and test information which is a concept from item response theory. These estimation methods will be discussed later. Both of these measures are more computationally intensive than alpha and though they give better estimates of reliability are not widely used at this time.

1.5.3 Factor Analytic Model

The Factor Analytic Model extends the classical test theory model. In its simplest form, the single factor model, the test is again assumed to measure a single latent trait, Θ, which is referred to as a factor. The advantage of the single factor model over the CTT model is that it allows each item on the test to vary in its difficulty and in its ability to measure the underlying factor. The model takes the form $Y = \sum_{i=1}^{k} X_i$ where each item is modeled by $X_i = \mu + \lambda_i \Theta + \varepsilon_i$. The coefficient λ_i is called the factor loading for the item and measures how well the item measures the latent trait. The intercept μ allows for each item to have a different difficulty level and ε_i is the random error component specific to item X_i (McDonald 1999). The factor analytic model can also be expanded to more complex multiple factor models.

Under the factor analytic model, another estimate for reliability can be obtained. This estimate is called coefficient omega ω, and it is defined by

$$\omega = \frac{\left(\sum \lambda_i\right)^2}{\left(\sum \lambda_i\right)^2 + \sum \mathrm{Var}(\varepsilon_i)} \quad .$$

The $\mathrm{Var}(\varepsilon_i)$ can be estimated from the item variance once the factor loadings have been obtained by the relationship $\mathrm{Var}(X_i) = \lambda_i^2 + \mathrm{Var}(\varepsilon_i)$ (McDonald 1999). Coefficient ω is derived by beginning with the definition of reliability as the ratio of true score variance to

total test variance. When test items are parallel as in the CTT model so that all the item covariances are equal and the factor loadings are equal, then coefficient ω is identical to coefficient α. Otherwise, it is strictly larger. One advantage to coefficient ω is that it does not assume that the test is unidimensional.

There are also shortcomings with factor analytic models. Since the models are linear, they generate impossible probabilities (less than zero or greater than 1) at the extremes of the Θ distribution, see Figure 1-2. Furthermore, the assumptions of the model include that the unique item variances $\mathrm{Var}(\varepsilon_i)$ are independent of Θ and that the measurement error estimate is constant for all values of Θ. Both assumptions are not true for dichotomous items (McDonald 1999).

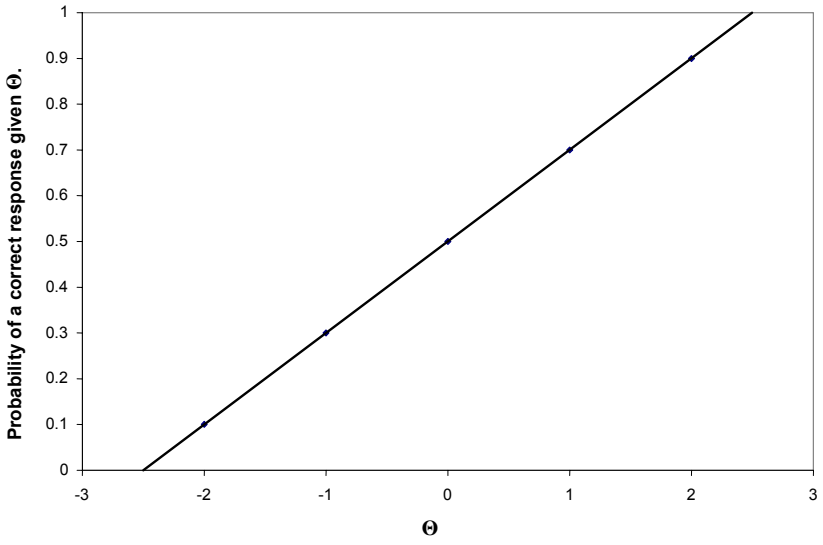


Figure 1-2: Model of the probability of a correct response to item X for a given Θ under the linear factor analytic model. Note that at the ends of the distribution, the probabilities become impossible.

1.5.4 Item Response Theory

The third method of test modeling is called the item response theory (IRT). IRT methods model the probabilities of a correct response using nonlinear models. The basic problem remains the same. There exists a latent trait, Θ, which the test is trying to measure. The trait is, as usual, unobservable and the items on the test are used to estimate Θ. By using nonlinear equations to model the item response functions, we can obtain functions that asymptotically approach 1 for high values of Θ and asymptotically approach 0 for low values of theta (Figure 1-3). Though there is no prescribed function that must be used, there are three models that are typically used.

For each model, the relationship between the latent trait and the observed examinee responses to test items is modeled by a logistic function. The focus of an IRT analysis is on the pattern of responses to the individual test items for each examinee, as opposed to the total test score. The item response patterns are used to determine a set of parameters for each item. These parameters then determine the shape of the item's item characteristic curve, which models the probability that an examinee with a given ability level will answer the item correctly, $P(X_i = 1 \mid \Theta)$, see Figure 1-3. The three models that are commonly in use are the one-, two-, and three parameter logistic models, referred to as 1PL, 2PL, and 3PL models respectively.

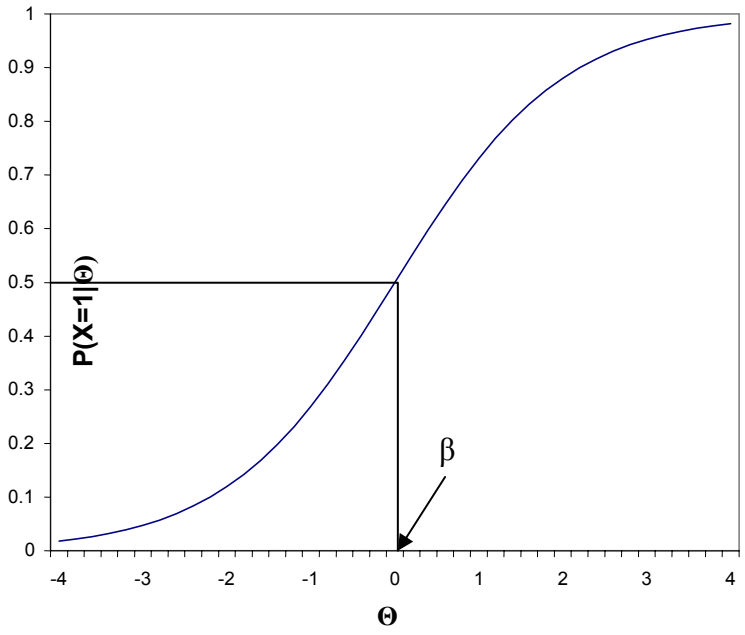


Figure 1-3: Example of an Item Characteristic Curve (ICC). The threshold parameter β is the value of Θ for which the probability of a correct response is 0.5.

In 1PL model, also known as the Rasch model, the probability of a correct response is modeled by the function

$$P(X_i = 1 \mid \Theta) = \frac{\exp(\Theta - \beta_i)}{1 + \exp(\Theta - \beta_i)}$$

where the parameter β is called the threshold parameter. β is equal to the value of Θ for which the probability of a correct response is 0.5, that is $P(X_i = 1 \mid \Theta = \beta_i) = 0.5$. The threshold parameter measures the difficulty of the item. Different ICCs are shown in Figure 1-4 for varying levels of the threshold parameter β. Items are assumed to have equal discrimination and little effects from guessing. These assumptions are restrictive and items that meet them are difficult to construct.

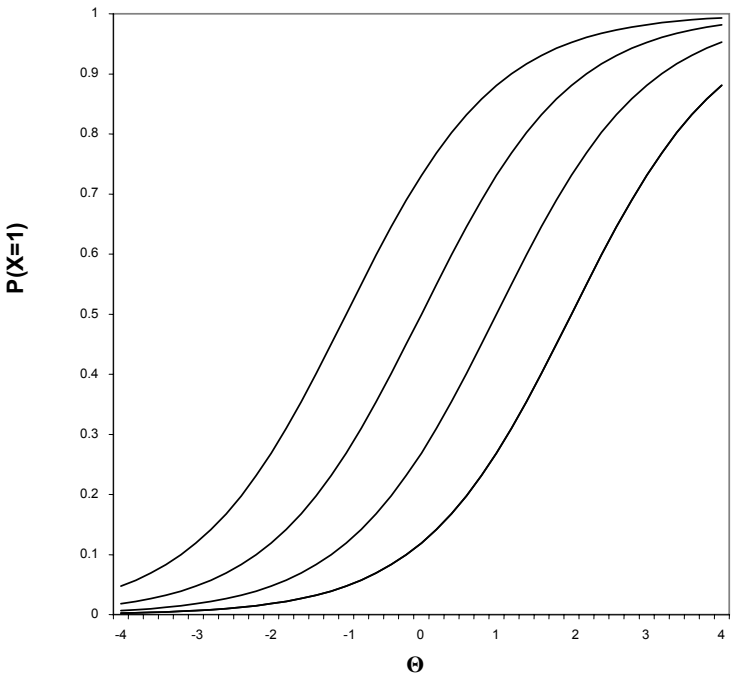


Figure 1-4: 1PL item characteristic curves for different values of the threshold parameter β.

The 2PL model adds an additional parameter, a, which is a discrimination parameter. The model takes the form

$$P(X_i = 1 \mid \Theta) = \frac{\exp[a_i(\Theta - \beta_i)]}{1 + \exp[a_i(\Theta - \beta_i)]}$$

where a_i is the value of the slope of the curve at the point Θ = β. The two parameters allow the items to differ in difficulty and discrimination, the ability of the item to differentiate between ability levels. Items which have high a_i values have steep slopes, so that once the threshold ability level is past, the probability of a correct response increases sharply. For lower a_i values, the curves and likewise the probabilities increase gradually, as in Figure 1-5. Steeply increasing curves are more desirable because if a respondent answers a question correctly, then we can be more confident that their ability level is greater than Θ = β. Questions with lower slopes result in more error in the ability estimations.

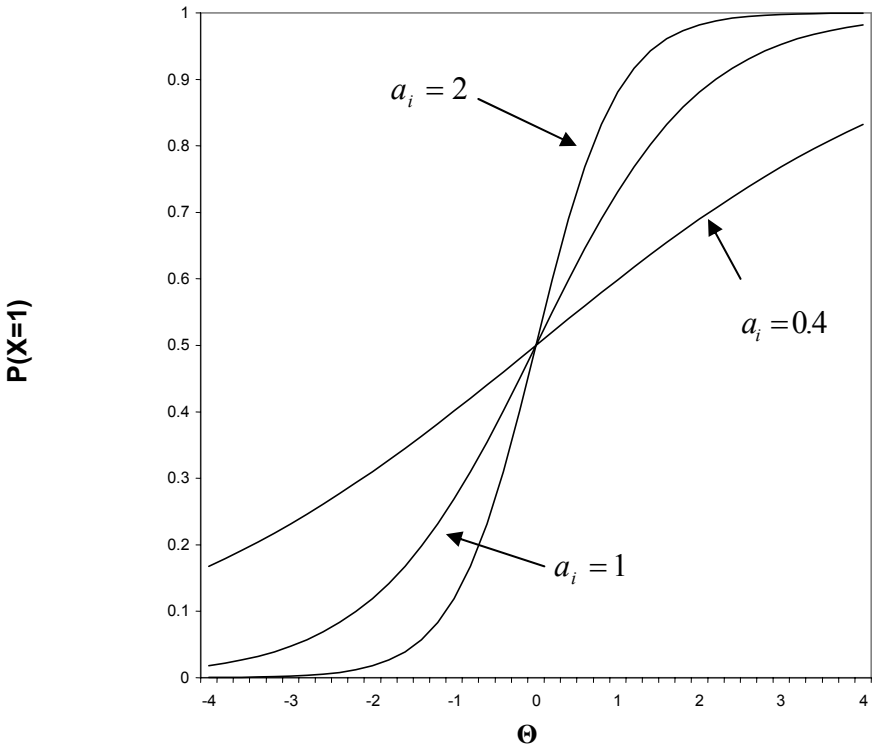


Figure 1-5: 2PL item characteristic curves for different values of a, β=0 for all curves.

The 3PL model adds one more parameter, c_i, which sets the lower asymptote of the curve. This is the probability that lower ability students will answer the question correctly and takes into account the effects of guessing. This parameter is referred to as a pseudo-guessing parameter. Pseudo because the probability is often lower than what would result from purely random guessing due to the attractiveness of some of the item distractors.

The 3PL model takes the form

$$P(X_i = 1 \mid \Theta) = c_i + (1 - c_i)\frac{\exp[a_i(\Theta - \beta_i)]}{1 + \exp[a_i(\Theta - \beta_i)]}.$$

When a guessing parameter is included, the threshold parameter is the value of Θ for

which the probability of answering correctly is equal to $\frac{1+c_i}{2}$. It is clear that the 1PL and 2PL models are special cases of the 3PL model.

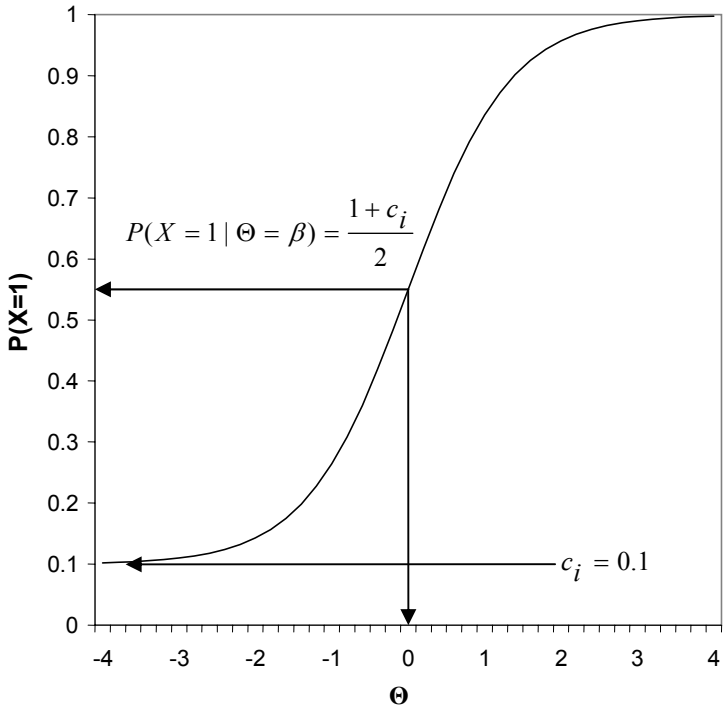


Figure 1-6: 3PL item characteristic curve with *a*=1.5, *β*=0, and *c*=0.1..

The parameter estimates are made using marginal maximum likelihood estimation procedures (Hambleton, Swaminathan & Rogers 1991). Under the IRT model, the probability of a correct response depends on the ability and the item parameters, all of which are unknown. What is known is the response pattern for each person. These response patterns are used to select values of the item parameters that maximize the likelihood of obtaining those response patterns. Once the item parameters are known, ability estimates can be obtained for each individual.

The assumptions of the IRT models are that the test is unidimensional; there is only one trait that accounts for the test performance. In practice this assumption is

considered to be met if there is a single dominant trait that influences the item responses, this is the trait that is measured by the test. The second assumption is that of local independence. This requires that an examinee's response to one item is independent of their response to another item, once ability has been taken into consideration. Essentially, this means that questions should not give clues to other questions, build on previous questions, etc.

There are several major advantages that IRT provides over CTT and factor analytic models. Assuming that the model fits the data, the parameter estimates are not sample dependent. Furthermore, estimates of examinee ability are also independent of the specific items chosen. The model also allows the measurement error to vary across the ability distribution. These advantages allow for the construction of shorter, more reliable tests, the possibility of adaptive testing, and tests that can be more tailored to specific needs (for example to distinguish between examinees at a narrow part of the ability distribution). It also provides better methods for test equating and detecting test bias.

Despite all the advantages of IRT, there are still important disadvantages. The model assumptions are more restrictive than for the other test models reviewed here. The estimation procedures are much more difficult to employ: they require many computer intensive calculations and special software that is expensive, not widely available, and not particularly easy to use. In addition, large data sets are required in order to estimate the item parameters.

IRT provides another tool for estimation of the measurement error, and thus the reliability of a test. This is the concept of item information. For each item on the test, the item information function, $I_i(\Theta)$, is constructed, where

$$I_i(\Theta) = \frac{\left[P_i'(\Theta)\right]^2}{P_i(\Theta)(1 - P_i(\Theta))}$$

(Hambleton, et al. 1991). The item information function describes how well the item estimates the unobserved ability. The item information function is highest when Θ is near the threshold level β, when the discrimination parameter a is large, and when the pseudo-guessing parameter c approaches zero see (Figure 1-7).

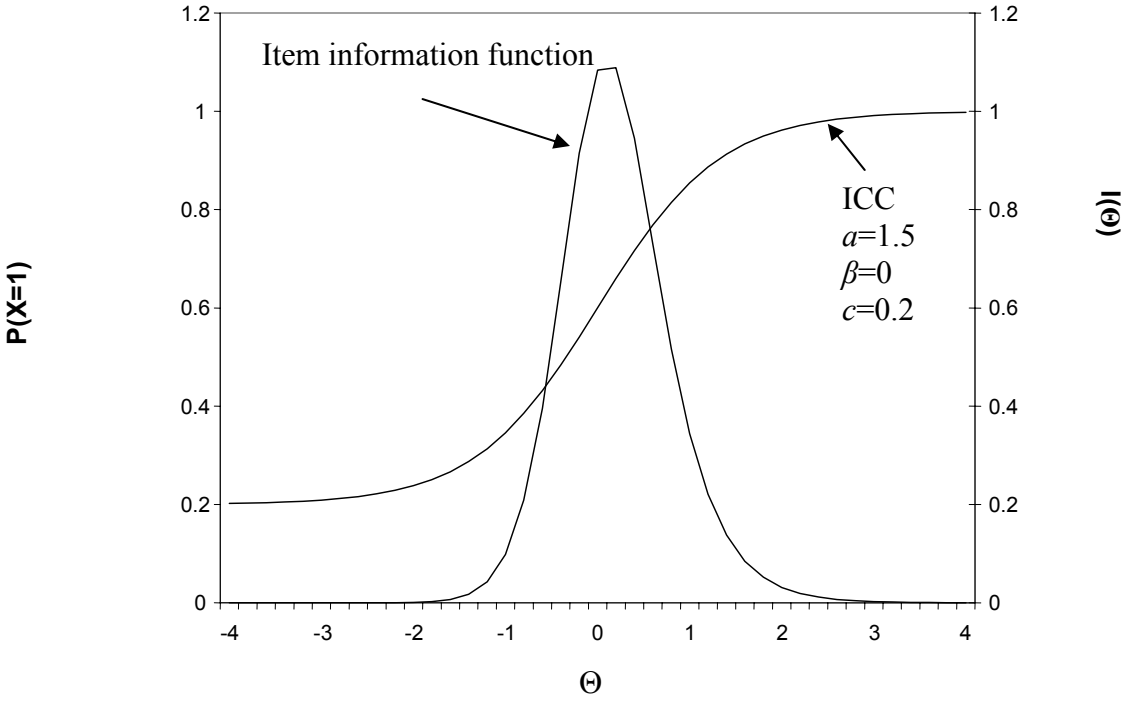


Figure 1-7: Item characteristic curve and its associated item information function.

The test information function, $I(\Theta)$ is simply the sum of the individual item information functions, $I(\Theta) = \sum I_i(\Theta)$ and describes the information provided by the test as a whole over the ability distribution. The test information function is used to define the standard error of measurement,

$$SE(\Theta) = \frac{1}{\sqrt{I(\Theta)}}$$

which estimates the precision with which ability is estimated. This estimate can be used as a measure of reliability of the test. Since the standard error is a function of Θ, it varies across the ability distribution.

Other IRT models are available as well, including those that deal with data that are not dichotomous (multiple response models) and even models for multidimensional data.

1.5.5 Validity

One other issue to consider before moving on is that of test validity. Validity is an important though somewhat murky concept in the test development process. Messick (1989 p. 13) defines validity as "the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores." Validity is a property of test scores and the inferences and decisions that are made based on them. Historically different types of validity have been proposed. These have been described as:

> Content validity is evaluated by showing how well the content of the test samples the class of situations or subject matter about which conclusions are to be drawn.
>
> Criterion-related validity is evaluated by comparing the test scores with one or more external variables (called criteria) considered to provide a direct measure of the characteristic or behavior in question.
>
> Predictive validity indicates the extent to which an individual's future level on the criterion is predicted from prior test performance.
>
> Concurrent validity indicates the extent to which the test scores estimate an individual's present standing on the criterion.
>
> Construct validity is evaluated by investigating what qualities a test measures, that is by determining the degree to which certain explanatory concepts or constructs account for performance on the test. (Messick 1989 p. 16)

However, Loevinger and Messick both argue that construct validity encompasses all other types of validity (Loevinger 1957, Messick 1989). They describe three components of construct validity: a substantive component, a structural component, and an external component. They make the case that a variety of evidence should be amassed to substantiate any validity claims.

To claim that an instrument has substantive validity, evidence must be presented to show that the items included in the instrument are consistent with the construct that the instrument intends to measure. Unlike traditional content validity, the substantive component of construct validity goes beyond making this claim based on the fact that the items were derived from a domain space clearly specified in advance and judged by experts to be representatives of the domain space. Instead, items should initially be drawn from a more broadly defined domain space (that should include competing constructs), the items should be used with a sample from the population of interest, and then the item set narrowed down based on empirical evidence from this administration. This analysis should include whether the individual items behave in ways consistent with the construct, consistent with one another, and consistent with the test format and what is known about objective testing.

Item response analysis should include individual analysis of keyed and distractor responses and factor groups analysis. The item should be included if the evidence shows it is a good question in terms of general objective test theory (i.e. the distractors are well written, it keys on the construct and not general test-taking skills, etc.), and that it is better explained by the construct in question rather than competing alternative constructs (i.e. it groups together better with the items believed to measure the intended construct

rather than those believed to measure the alternative). Once the item set has been selected, the set should be re-analyzed to check that the specific domain space is still well represented. In this manner, each question has a justified presence on the exam that is much more than an asserted belief that it would be a measure of the construct.

The structural component of construct validity is complementary to the substantive component. Since the substantive component relies heavily on the analysis of test scores, it is important for the scoring methods to be in tune with the construct being measured. Questions such as whether the score should be reported as a single score, sub scores only, or both sub scores and composite scores should be answered by the behavior of the underlying construct theory. For example, is the construct strictly unidimensional? Additionally, appropriate scoring measures should be employed based on whether the instrument is intended to be a normative or criterion measure. The scoring model, i.e. the use of a cumulative scoring model where item responses are summed or a class model where item responses result in a classification for the individual, should be justified by the construct theory. A disconnect between the construct theory and the scoring routines, score reporting, and score interpretation has serious implications for the validity of the instrument.

The external component of construct validity encompasses how the test interacts with other measures (both test and non-test behaviors) as predicted by construct theory. Do the test scores as reported by the scoring model correlate in predictable ways with other methods of measuring and behavioral indicators of the same construct and with measures of alternative constructs?

The three components of validity are highly intertwined in that each cannot stand alone and each draws strength and credibility when the other two are well specified and supported by evidence. They differ in distinct ways however. The substantive component focuses on the test at the item level. Is each item that is included well conceived and justified by the theoretical construct? Do the items work together as a whole to represent the domain space? The structural component focuses on how the measurement is structured, reported, and interpreted. Does the scoring and measurement model make sense beyond the individual items and is it consistent with the underlying theory? The external component looks outside of the test to make sure that the test behaves in predictable ways in relation to other variables.

As validity is a complex construct, there is no single measure to point to when trying to establish the validity of tests, test scores, and most importantly the inferences that are drawn from them. Instead, evidence of validity must be collected from a variety of sources. Establishing validity is an ongoing process.

1.5.6 Summary

The literature reviewed here is focused on three areas that have been critical in the development and evaluation of the Statistics Concept Inventory. Examining other concept inventories that have been or are being developed provides a roadmap that has helped to guide the construction process by identifying good practices, common characteristics, typical baseline data, and tools that have been successfully used in the construction and evaluation process of similar instruments.

Chapter 2: A Classical Test Theory Perspective

This chapter will provide an overview of the Statistics Concept Inventory (SCI) from the prospective of the whole test, following a classical test theory tradition. A brief history of how the SCI was developed is presented including topic selection, question development, and revision practices. A classical test theory analysis follows including baseline data for pretest and posttest scores along with normalized gains and a reliability analysis based on coefficient alpha.

2.1 Development of the Statistics Concept Inventory

Development of the SCI began in fall 2002. The goal of the project was to write a multiple choice instrument that was non-computational and would assess understanding of statistics concepts and that would help identify common student misconceptions. The SCI was to be modeled after the Force Concept Inventory (FCI) for introductory physics (Halloun & Hestenes 1985); however, several important distinctions between the two disciplines had to be considered.

The FCI covers only the mechanics portion of a first semester physics course, typically about the first half of the semester. These very important concepts form an integrated and well organized system for understanding motion and force. While these concepts stand as an important base for much of the following physics content, they also delimit a natural segment of the physics curriculum. The FCI is also designed so that anyone can understand and complete the instrument, even with no formal physics

training. This is possible because everyone has first hand, real world experience with the content matter. That is, everyone has experience with things falling, moving, colliding, and moving at different speeds. Everyone has developed an intuitive understanding of these phenomena, whether these conceptions are accurate in the Newtonian sense or not.

In contrast, much of the content of an introductory statistics course does not easily resonate with everyday experience. P-value, confidence intervals, and hypothesis testing are a few examples that have no tangible counterpart in daily experience, and many of the topics are not easily discussed without some specific jargon. There is also not a distinct and natural breaking point in the curriculum to target a concept inventory toward. As a result, a large set of possible topics must be considered for inclusion in a concept inventory.

The possibility of writing multiple inventories was considered, such as one each for probability, for descriptive statistics, and for inferential statistics. This idea was discarded, however, for several reasons. First, in practice, these areas of statistics usually go hand in hand, one informing the other. An introductory statistics course certainly spans these topics, and the desire to have an instrument(s) that can help evaluate the instructional methods and student learning in an introductory statistics course dictates that the instrument(s) span the topic coverage of the course. From a practical standpoint, administering an assessment instrument takes up valuable class time. It is not likely that instructors would be willing or able to dedicate the time required to administer multiple instruments. Thus, the choice was to develop a single instrument. The tradeoff in making this choice is that not all topics can be covered and that, for many topics, only a single item can be devoted to that topic.

The starting point for selecting topics to include was previous research using a modified Delphi technique. A list of possible topics had been compiled by surveying topics widely addressed in introductory engineering statistics texts. Input from faculty members in the College of Engineering at the University of Oklahoma was sought to identify important statistics concepts and concepts that these instructors felt were difficult to teach. The respondents were asked to rank the importance of statistics topics for their curricular needs on a scale from 1 (not at all important) to 4 (very important), along with the option of "No opinion" if the topic was unfamiliar. Respondents were asked to note any other key topics that they felt were missing. Twenty-three faculty members completed the survey. A summary of the results of this survey is included in Table 2-1. Topics are ranked by their average importance. Additionally, a list of topics covered on the Advanced Placement Statistics Exam served as a further guide to topic coverage (College Entrance Examination Board 2001). Once a topic list had been established, problems and possible answers were drafted incorporating misconceptions that had been identified from research for topics on that existed in the literature (see Section 1.3).

Table 2-1: Results of the Instructor Survey of Statistics Topics, ordered by average ranking. The median ranking was 2.62. Topics were ranked from 1 (not at all important) to 4 (very important).

General Area	Specific Topic	Average Ranking	# of Rankings
Other (write in category, responses varied)	Other	3.75	8
Data Summary & Presentation	Measure of variability	3.68	22
Data Summary & Presentation	Importance of data summary	3.65	23
Linear Regression	Simple linear regression	3.52	21
Continuous Random Variables & Probability Distribution	Normal distribution	3.48	23
Data Summary & Presentation	Methods of displaying data	3.43	23
Continuous Random Variables & Probability Distribution	Continuous uniform distribution	3.32	22
Probability	Interpretation of probability	3.26	23
Discrete Probability Distributions	Poisson distribution	3.14	22
Joint probability Distributions	Covariance and correlation	3.10	21
Linear Regression	Properties of the least squares	3.10	21
Data Summary & Presentation	Frequency dist and histograms	3.09	22
Random Variables	Expected values	3.09	23
Data Summary & Presentation	Time sequence plot	3.00	20
Probability	Independence	3.00	22
Parameter Estimation	The central limit theorem	3.00	19
Probability	Sample space and events	2.95	21
Parameter Estimation	Random sampling	2.95	21
Linear Regression	Correlation	2.95	21
Continuous Random Variables & Probability Distribution	Standardized normal	2.87	23
Discrete Probability Distributions	Binomial distribution	2.86	21
Linear Regression	Use of the regression for prediction	2.86	21
Probability	Conditional probability	2.85	20
Parameter Estimation	Properties of estimators	2.84	19
Probability	Multiplication and total probability rules	2.81	21
Linear Regression	Confidence intervals for the regression	2.81	21
Probability	Axiomatic rules	2.80	20
Random Variables	Linear combinations	2.80	20
Confidence Intervals & Hypothesis Testing	Testing for goodness of fit	2.78	18
Probability	Counting concepts	2.77	22
Random Variables	Functions of random var.	2.76	21
Discrete Probability Distributions	Discrete uniform distribution	2.76	21
Joint probability Distributions	Two discrete random variables	2.75	20
Parameter Estimation	Sampling distribution	2.75	20
Confidence Intervals & Hypothesis Testing	Inference on the mean of a population	2.74	19
Probability	Addition rules	2.72	18
Linear Regression	Assessing the adequacy of regression	2.71	21
Confidence Intervals & Hypothesis Testing	Sample size determination	2.68	19
Linear Regression	Hypothesis tests in regression	2.67	21

Table 2-1 continued.

Probability	Bayes' theorem	2.63	19
Confidence Intervals & Hypothesis Testing	Inference on the var. of a norm	2.63	19
Continuous Random Variables & Probability Distributions	Lognormal distribution	2.62	21
Parameter Estimation	Maximum likelihood estimation	2.60	20
Data Summary & Presentation	Percentiles and quartiles	2.59	22
Confidence Intervals & Hypothesis Testing	Inference on a population proportion	2.59	17
Parameter Estimation	Estimators and their properties	2.58	19
Single factor experiments	Estimation of model parameters	2.56	18
Joint probability Distributions	Multiple discrete random variables	2.52	21
Continuous Random Variables & Probability Distribution	Exponential distribution	2.50	22
Time Series, etc.	The ratio-to-moving-average method	2.50	16
Single factor experiments	Sample size	2.50	18
Joint probability Distributions	Bivariate normal distribution	2.44	16
Time Series, etc.	Exponential smoothing methods	2.44	16
Confidence Intervals & Hypothesis Testing	Inference on means of 2 norm population.	2.41	17
Continuous Random Variables & Probability Distribution	Normal approx.	2.41	22
Random Variables	Moment generating functions	2.38	21
Time Series, etc.	Trend analysis	2.38	16
Time Series, etc.	Seasonally and cyclic behavior	2.38	16
Confidence Intervals & Hypothesis Testing	Paired comparisons	2.35	17
Discrete Probability Distributions	Hypogeometric distribution	2.30	20
Linear Regression	F test of the regression Model	2.29	21
Multi-factor designs	Expected mean squares	2.29	14
Single factor experiments	Model adequacy check	2.28	18
Data Summary & Presentation	Box plots	2.26	19
Discrete Probability Distributions	Geometric and neg. binomial	2.25	20
Confidence Intervals & Hypothesis Testing	Inference on 2 population proportions	2.25	16
Multi-factor designs	Rand complete block design	2.25	16
Confidence Intervals & Hypothesis Testing	Infer on var. of 2 norm populations.	2.24	17
Single factor experiments	Analysis of the fixed effects	2.24	17
Single factor experiments	Non parametric ANOVA	2.24	17
Multi-factor designs	2 factor factorial design	2.24	17
Single factor experiments	Comparison of treatment means	2.22	18
Parameter Estimation	Chebyshev's inequality	2.21	19
Continuous Random Variables & Probability Distribution	Beta distribution	2.14	22
Multi-factor designs	General factorial design	2.12	17
Confidence Intervals & Hypothesis Testing	Contingency table tests	2.07	15
Data Summary & Presentation	Skewness and kurtosis	2.06	18
Single factor experiments	ANACOVA	2.06	18
Continuous Random Variables & Probability Distribution	Weibull distribution	2.00	22
Multi-factor designs	Latin square design	2.00	16
Data Summary & Presentation	Stem-and-leaf diagrams	1.89	18
Multi-factor designs	Factorial design with rand factors	1.88	17
Multi-factor designs	Graeco-latin square design	1.88	16

The instrument was developed with an engineering student population in mind. However, as the content of an introductory statistics course is fairly homogenous across disciplines, we saw no need to limit use of the SCI to engineering courses. To facilitate this wider use, we chose to limit the amount of engineering contexts and jargon within the questions. This decision had the added benefit of reducing the effects of possible confounding based on unfamiliarity with engineering concepts either from those in different branches of engineering or those who have not encountered them yet in their coursework.

The initial set of 32 questions was piloted during the fall 2002 semester in four statistics courses at the University of Oklahoma (Stone, Allen, Rhoads, Murphy, Shehab & Saha 2003). During the spring 2003 semester, small focus groups were conducted with students who had completed the instrument in the fall. Students were asked to comment on why they chose certain answers and how they eliminated others, as well as to point out any areas of confusion. The instrument was revised based on comments from the focus groups and on the distribution of answers to each response alternative. Distracters that were not chosen were rewritten or replaced. Nearly a third of the questions were replaced to make the instrument more closely aligned with the goals of the project.

The revised instrument was given during the summer 2003 to additional sections of introductory statistics and two Research Experience for Undergraduates (REU) groups. The REU groups were unique in that they were not currently receiving formal statistics instruction, though some statistics review was provided. REU students attended two, 2-hour seminars presented on statistics. The students came from a variety of institutions and had wide range of statistics training and educational background. Additional larger focus

groups were conducted with the REU students. These focus groups were well attended and students were candid and enthusiastic in their responses.

The instrument was once again revised based on these comments and item statistics were calculated, including the distribution of answers, difficulty, discrimination, and alpha-if-item-deleted. Effort was also made to eliminate or reconstruct poorly written items. Small changes were made such as italicizing important words in the stems, e.g. least and most to minimize incorrect answers due to inaccurate reading. Each question was evaluated on the basis of seven criteria identified by Gibb (1964); questions with these properties may lead students with good test-taking skills to figure out the answer in the absence of content knowledge:

- Phrase-Repeat: there is an alliterative association between the correct answer and the question stem, for example the correct answer contains a key sound, word, or phrase that is also contained in the question's stem.
- Absurd Relationship: distracters are unrelated to the stem or are clearly not plausible.
- Categorical Exclusive: distracters contain words such as all, never, or every.
- Precise: the correct answer is more precise, clear, or qualified than the incorrect alternatives.
- Length: the correct answer is visually longer than the other responses.
- Grammar: the tense or plurality of the distracters does not match that of the stem, or there is not a match between articles (a, an, the).
- Give-Away: correct answer is given away by another item in the test.

In this manner, we have continued to revise the instrument each semester. Focus groups were again held during summer 2004 with additional REU students. We have revised the

questions to improve their clarity, increase their psychometric properties (discrimination, reliability), and to sharpen the focus on concepts rather than definitions, recall, or problem solving ability. More detail about the revision process can be found in Allen, Stone, Rhoads, and Murphy (2004) and in Chapter 3.

Table 2-2 identifies the topics and their associated items that are included in the current version of the SCI. Table 2-3 provides a taxonomy of some errors and misconceptions that are included on the SCI. It is assumed that selection of these responses gives some evidence for the presence of the misconception.

Table 2-2: Item classification for the SCI.

Topic	*SCI Item*
Descriptive	
Choose best sampling method, stratified random sampling	3f --> 3d
Median	9c
Impact of outliers on descriptive statistics	11b
Weighted mean	12c
Choosing an appropriate measure for the central tendency of a data set, accounting for outliers	15b
Correctly identifying data sets that would be normally distributed	23a
Correctly interpret standard deviation	26c
The standard deviation must always be positive	29d
Compare the variability of different data sets, scenario format	6d
Compare the variability of different distributions in graphical format	30a
Percentiles	8c
Inferential	
Hypothesis Testing- Formulating alternate hypothesis, one tailed	10d
Confidence Intervals, meaning of	17c
Larger samples decrease the width of a confidence interval	35a
p-value, meaning of in hypothesis testing	18c
Properties of t-distribution	19d
Parameter Estimation, interpreting mean and standard deviation of a sample relative to a single observation	20d
Relationship between p-value and sample size	22a
Correctly decide whether to reject the null hypothesis using p-value	32d
Correctly choose which statistical test is appropriate for a given situation	2b,36c
Interpret correlation coefficient	38b
Sampling, identifying potential bias	27b
Probability	
Make a prediction based on available data	5b
Distributions, waiting time, memoryless property	13c
Sequence of independent events are equally likely	16d
Probability laws for independent events	21d
For dependent events, the occurrence of one event changes the probability of the other	31b
Correctly use 68-95-99 rule for normal distribution	33a
Apply the law of large numbers	4a;34c
Marginal probability	1c
Graphical	
Interpret and make comparisons between different graphical representations	7a
Central Limit Theorem, sample means are normally distributed	14a
Identify most likely parent distribution, uniform	25b
Correctly read and interpret a histogram	28b
Estimate correlation coefficient from a scatter plot of data	24c
Correctly interpret changes to correlation coefficients when specific data points are removed from a scatterplot.	37c

Table 2-3: Taxonomy of errors and misconceptions identified by the SCI.

Error/Misconception	*SCI item*
Hypothesis Testing	
Confuse null and alternate hypotheses	10a
Fail to distinguish between one and two tailed situation	10b
The relationship between p-value and significance level not understood	18a
p-value confused with power	18b
p-value interpreted as the probability the null hypothesis is true/false	18d,e
p-value is unrelated to sample size	22c
Incorrectly interpret p-value in deciding whether or not to reject the null hypothesis	32c
Belief that the null/alternate hypotheses can be proven to be true using a statistical test	32a,b
Unable to correctly choose among statistical tests for given situations	2a,c;36a,b,d
Average/Central tendency	
Always add all the numbers and divide by the total numbers summed to determine mean	12a,b
Fail to appreciate the effect of outliers on mean	11c,15a
Believe that the mean cannot be determined unless every data point is known, e.g. in weighted mean or frequency data situations	12d
Confuse mean, median, and mode	9d;15a,c
Believe that the mean is a superior measure	15a
Believe that it is possible for all data points to be below the mean	29b
Spread	
Standard deviation gives information about the symmetry of a distribution	26d
Standard deviation gives information about the location of the data	
Standard deviation can be negative	26a;29a,b,c
Interpret variability as "bumpiness" of a histogram	29a,b,c
Associates or equates variability with randomness	30b,c 6c
Sampling Distributions	
A good sample must contain a large percentage of the population	3c
Sampling distributions should look like the population distribution	14d
Partially applying central limit theorem, failing to center distribution at the population mean	14b
t-distribution has less area in the tails than normal distribution	19c,e
t-distribution not used for small samples	19b
Probability	
Representative Heuristic	4c,d;5a;16a,b,c
Misapply probability laws or use Anchoring and Adjustment Heuristic	21a,b,c
Probabilities of conditional events are equal to individual probabilities	31c
Unable to make predictions about the probability of an event using 68-95-99 rule for a normal distribution	33b,c,d
Belief in the "Law of small numbers"	4c;34a,b
Confidence Intervals	
An X% confidence interval implies that X% of the observations will fall within the limits of the confidence interval	17b
Sample size does not affect the width of confidence intervals	35c
Larger samples increase the width of confidence intervals	35b

2.2 Participants and Data Collection

Participants have been recruited from statistics courses at the University of Oklahoma in the departments of Engineering, Mathematics, Psychology, and Communication. Instructors from other institutions who have expressed an interest in the instrument have also contributed data. The classes were chosen depending on the availability of class time and the permission of the instructor. The courses that have been involved, including a brief description and their prerequisites, are shown in Table 2-4.

The majority of the participants were junior and senior engineering, mathematics, physics, or meteorology majors enrolled in their first or second statistics class. These courses required calculus as a prerequisite. There have also been a few classes of predominately social science majors that are comprised of more freshmen and sophomores. These classes did not require calculus as a prerequisite and in general these students had much less mathematics experience.

Participating students were administered the instrument in class as a pretest when possible during the first two weeks of class and as a posttest during the last two weeks of class. Students completed the instrument along with a short demographic questionnaire and in some cases an attitude survey. Students were given between 35 and 45 minutes to complete the instrument. They were asked to answer each question to the best of their ability. They were told that they would not need calculators, but they were free to use them if they liked.

Table 2-4: Participating courses and their prerequisites. Courses are at the University of Oklahoma except where noted otherwise.

Course	Description	Prerequisites	Fall 2002	Summer 2003	Fall 2003	Spring 2004	Summer 2004	Fall 2004	Spring 2005	Summer 2005
ENGR 3293 *Applied Engineering Statistics* College of Engineering	"Introduction to probability, one and higher dimensional random variates, functions of random variables, expectation, discrete and continuous distributions, sampling and descriptive statistics, parameter estimation, use of statistical packages" (University of Oklahoma 2005) This course is comprised of sophomore, junior and senior engineering majors.	ENGR 1112(Introduction to Engineering) and ENGR 1001(Engineering computing) or COMP 1313 (Computer Programming for Non-majors) or COMP 1323 (Introduction to Computer Programming) and MATH 2433 Calculus/Analytical Geometry III)	X	X	X	X		X	X	X
MATH 4753 *Applied Statistical Methods* Department of Mathematics	"Estimation, hypothesis testing, analysis of variance, regression and correlation, goodness-of-fit, other topics as time permits. Emphasis on applications of statistical methods." (University of Oklahoma 2005) This course is taken by junior, senior and graduate students. Around 50% of the students are engineering majors, 30% are Geo-science majors (typically meteorology), while only around 15% are mathematics majors.	MATH 2123 (Calculus II for Business, Life and Social Sciences) or MATH 2423 (Calculus and Analytical Geometry II)	X	X	X	X		X	X	X
MATH 4773 *Applied Regression Analysis* Department of Mathematics	"The general regression problem of fitting an equation involving a single dependent variable and several independent variables, estimation and tests of regression parameters, residual analysis, selecting the "best" regression equation." (University of Oklahoma 2005) This course is taken by mainly senior and graduate students in mathematics (50%), engineering (25%) and geosciences (20%).	MATH 3333 (Linear Algebra), MATH 4733 (Probability) or MATH 4753 (Applied Statistical Methods) or any statistical probability course at an equivalent level.						X		

Table 2-4 continued.

Course	Description	Prerequisites	Fall 2002	Summer 2003	Fall 2003	Spring 2004	Summer 2004	Fall 2004	Spring 2005	Summer 2005
IE 4553 *Engineering Experimental Design* College of Engineering	Fundamentals of design of experiments. Analysis of variance models for single factor designs with blocking factors and multi-factor designs, including factorial and nested designs. Fixed, random and mixed models. Analysis of covariance models. (University of Oklahoma 2005) This course is taken by junior, senior, and graduate engineering students.	ENGR 3293	X		X					
COMM 2513 *Introduction to Statistics* Department of Communications	"This course introduces statistics with the purpose of providing tools which aid in conducting scientific research. Topics include: measurement, central tendency, variability, normal distribution, probability, correlation, sampling distributions." (University of Oklahoma 2005)This course is predominately taken by freshmen, sophomore, and junior social science and life science majors including many in pre-health disciplines.	High School Algebra	*X*							
PSY 2003 *Understanding Statistics* Department of Psychology	"An introductory applied statistics course which will focus on descriptive and inferential statistical methods. Emphasis will be placed on in-class activities and homework which help the student learn by experience. Topics include measures of central tendency and variability, z-scores, normal distribution, correlation, regression, sampling distributions, hypotheses testing, t-tests and chi-square tests." (University of Oklahoma 2005) This course is comprised of freshman-senior level students in a variety of life science, social science, and pre-health majors.	Math 0123-Algebra							X	

Table 2-4 continued.

Course	Description	Prerequisites	Fall 2002	Summer 2003	Fall 2003	Spring 2004	Summer 2004	Fall 2004	Spring 2005	Summer 2005
PSY 2113 *Research Methods I: Statistics* Department of Psychology	"An introduction to scientific method in psychological research. Topics include: philosophical issues; hypothesis formulation; experimental design; and data collection, organization and interpretation." (University of Oklahoma 2005)	Math 0123-Algebra								X
IE 1071 *Probability and Statistics for Engineers 2* University of Pittsburg, Engineering	Review of joint distributions and estimation; Chi square, t, and F sampling distributions introduced; estimation hypothesis testing; multiple regression; empirical model building; analysis of variance and design of experiments; goodness-of-fit tests and contingency tables; introduction to statistical quality control. (University of Pittsburg 2005) This course is comprised of junior industrial engineering majors.			X	X			X	X	
APMA 311 *Applied Statistics and Probability* University of Virginia, Engineering and Applied Science	Examines variability and its impact on decision-making. Introduces students to basic concepts of probability, such as random variables, probability distribution functions, and the central limit theorem. Based on this foundation, the course then emphasizes applied statistics - covering topics such as descriptive statistics, statistical inference, and regression modeling. (University of Virginia 2005) This course is comprised of sophomore and junior engineering students.	Multivariate Calculus			*X*					

Table 2-4 continued.

Course	Description	Prerequisites	Fall 2002	Summer 2003	Fall 2003	Spring 2004	Summer 2004	Fall 2004	Spring 2005	Summer 2005
MATH 2023 *Elementary Statistics* Northern Oklahoma College	"Descriptive measures, probability, sampling distributions, estimation and hypotheses testing, chi-square, regression and correlation, analysis of variance" (Northern Oklahoma College 2005). This course is taken by a wide variety of students and is offered on multiple campuses and as a distance-learning class, however, the majority of students are sophomores and business majors.	College Algebra			X					
MATH 106 *Elementary Statistics* Monmouth College	"A study of the methods of describing and analyzing data and an introduction to statistical inference with applications. Topics include mean and variance, data displays, normal distribution, correlation and regression and tests of significance for means and proportions." The majority of students in this course are freshmen and sophomore business, sociology, and elementary education majors.	None						X	X	
REU *Research Experience for Undergraduates* University of Oklahoma	Program sponsored by the National Science Foundation to actively involve undergraduate students in research activities. Students must apply to specific REU sites More information is available from the National Science Foundation (2005).	Undergraduates have a variety of background experiences, including varying levels of exposure to statistics training.		X			X			

2.3 Results

2.3.1 Posttest Scores

Since fall 2002, over 1100 students have completed the SCI as a posttest. The summary statistics for each course and semester are shown in Table 2-5. The scores for each semester were normally distributed, except for fall 2002 and fall 2003 in which they were approximately normal. The mean posttest scores have been consistently low ranging from 45-50% each semester after fall 2002. The mean scores by course have ranged from 32 to 51%, with the majority falling between 45 and 50%, as well.

Side-by-side box plots of the posttest scores by semester are shown in Figure 2-1. Following the major revisions after fall 2002, the scores have been very consistent from semester to semester. Figure 2-2 shows side-by-side box plots by course. There is much more variation among the posttest scores by course. ANOVA was conducted to test for the effect of semester, S, and course, C, on the posttest score, Y, using a nested factorial model:

$Y_{ijk} = \mu + S_i + C_j + SC_{ij} + P(SC)_k + \varepsilon_{ijk}$ where μ is the overall mean, SC_{ij} is the interaction between semester and course, $P(SC)_k$ is participants nested in the interaction between semester and course, and ε_{ijk} is random error.

Table 2-5: SCI Summary Statistics by Course and by Semester

		Participants (n)								*SCI Posttest*				*Gender %*	
		Fall 2002	Summer 2003	Fall 2003	Spring 2004	Summer 2004	Fall 2004	Spring 2005	Summer 2005	Total	Mean %	sd %	Range %	F	M
Courses	APMA 311			102						102	51.3	11.6	24-82	36	64
	COMM 2513	65								65	32	9.6	9-56	64	36
	ENGR 3293	36	24	53	31		8	17	10	179	44.3	14.4	9-82	18	82
	IE 1071		38	43			41	49		171	49.4	13.2	16-82	30	70
	IE 4553	30		26						56	38.7	10.4	18-72	41	59
	MATH 106						55	48		103	41.9	10.7	14-70		
	MATH 2023			37						37	31.5	10.1	18-56	64	36
	MATH 4753	39	14	19	63		28	40	36	239	47.7	15	15-80	32	68
	MATH 4773						31			31	50.7	13.6	30-81	48	52
	PSY 2003							106		106	46.2	11.5	24-73	70	30
	PSY 2113								14	14	36.0	9.7	18-53	64	36
	REU		27			16				43	49.5	11.5	30-70	33	67
	Total (n):	170	103	280	94	16	163	260	60	1146	45.5	14.1	9-82	39	61
SCI Posttest	# of items:	32	34	34	35	37	37	38	38						
	Mean %	36.7	49.2	45.5	49.7	49.6	50.5	46.3	45.7	45.5				42.2	47.5
	St Dev %	12	14.4	14.4	13.8	11.5	12.9	13.2	13.3	14.1				13.1	14.4
	Range %	9-75	15-82	18-82	23-80	30-70	22-81	16-82	18-76	9-82				9-79	9-82
	Coefficient Alpha	0.59	0.74	0.75	0.72	0.67	0.67	0.69	0.70						

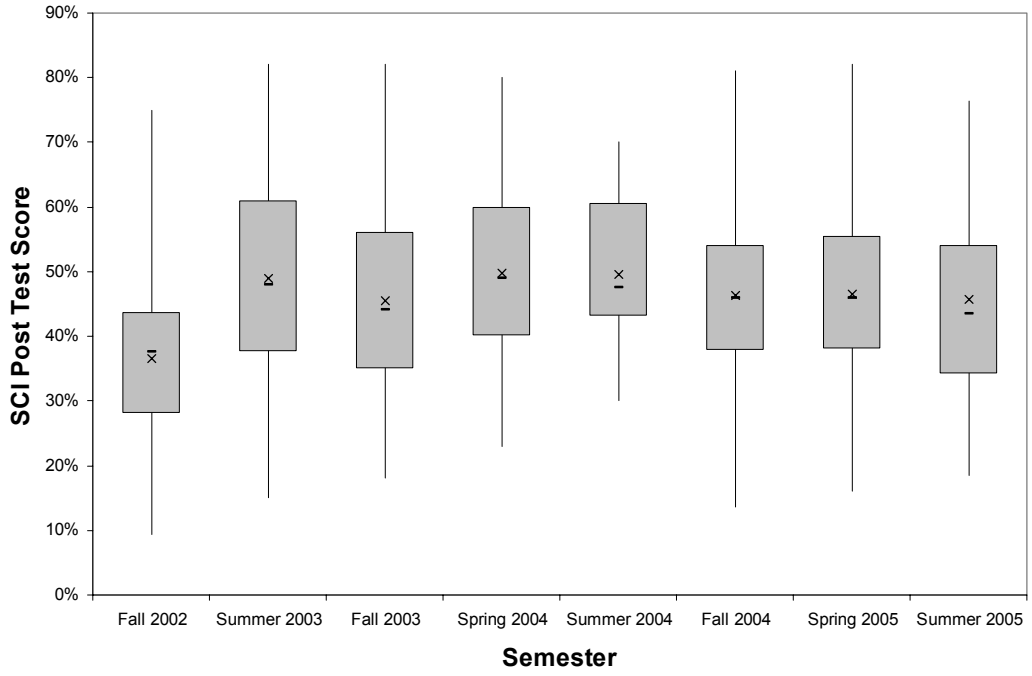


Figure 2-1: Box Plots of SCI posttest scores by semester, (median represented by -, mean represented by x).

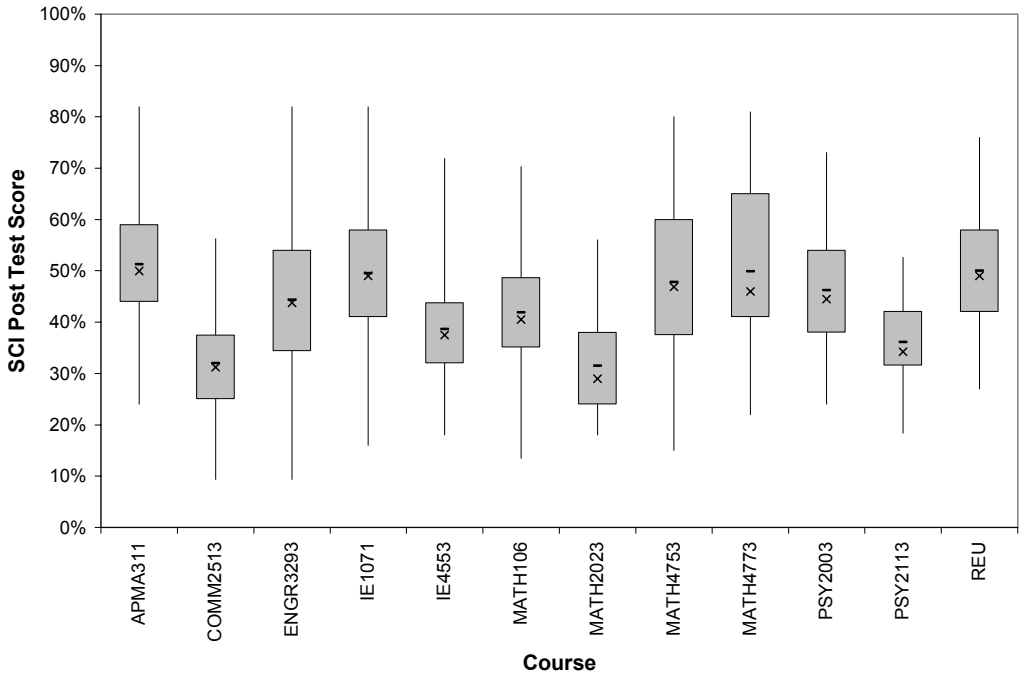


Figure 2-2: Box Plots of SCI posttest scores by course, (median represented by -, mean represented by x).

Both the semester and course effects were significant; the interaction was not. See Table 2-6. The fall 2002 semester was the only semester significantly different. This is most likely due to the significant revisions that took place after the initial piloting of the questions that semester. A Tukey test on the means for course at the $\alpha=0.05$ significance level indicated that many of the groupings overlap, making it difficult to draw firm conclusions, see Figure 2-3. In general, however, courses that serve a non-engineering student population tend to have lower scores. These courses may have different content coverage as well as serving students who are younger, have less mathematics background, and who are usually non-science majors. Any or all of these factors are likely to contribute to the difference in scores.

Table 2-6: ANOVA summary table.

Source	df	Sums of Squares	Mean Square	F Value	Pr > F
Semester	11	0.41341222	0.05905889	3.64	0.0007
Course	7	2.06940097	0.18812736	11.60	<0.0001
Semester x Course	11	0.25787637	0.02344331	1.45	0.1465
Participant (Semester x Course)	1079	17.49193150	0.01621124	.	.
Error	0	.			
Total	1108	20.23262106			

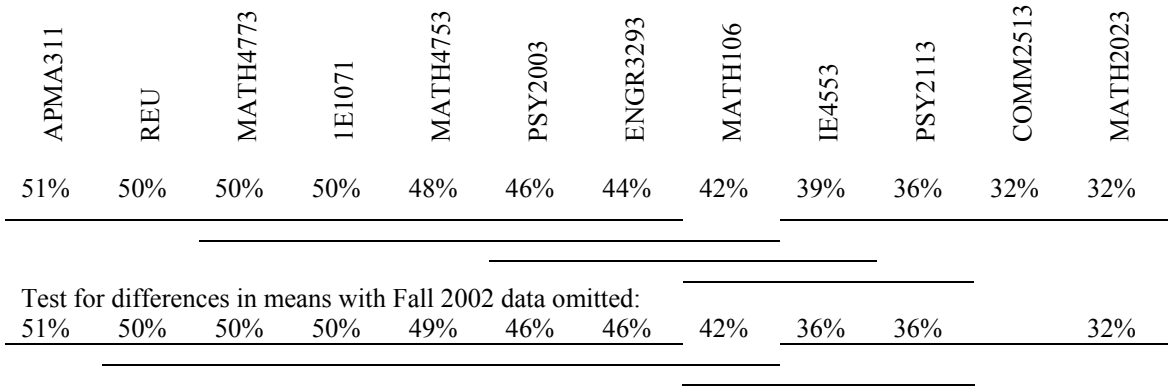


Figure 2-3: Results of the Tukey test for differences in means for course, presented as lines.

2.3.2 Gains

Pretest data were not available for all courses. Using only the observations for which both pre- and posttest data were available, gains and normalized gains were analyzed, where the gain is the posttest score less the pretest score and the normalized gain is the ratio of the gain to the total possible gain. Gains are minimal in most classes, see Table 2-7. The range of normalized gains is consistent with the range found with the FCI in traditionally taught physics classes (Hake 1998), the Signals and Systems Concept Inventory (SSCI) (Wage, Buck, Wright & Welch 2005), the Wave Concept Inventory (WCI) (Roedel, El-Ghazaly, Rhoads & El-Sharawy 1998), and the Materials Concept Inventory (MCI) (Krause, Decker & Griffin 2003).

Table 2-7: Gains and normalized gains for classes in which both pre- and posttest data were available.

Semester	COURSE	N	Average Gain	Normalized Gain	Pretest	Posttest
SU03	ENGR3293	23	13%	20%	35%	48%
	MATH4753	12	15%	25%	39%	54%
F03	APMA311	99	4%	8%	47%	51%
	ENGR3293	47	1%	2%	42%	44%
	IE4553	16	6%	9%	33%	39%
	MATH2023	32	2%	4%	30%	33%
	MATH4753	14	3%	5%	49%	51%
SP04	ENGR3293	29	7%	12%	41%	48%
	MATH4753	59	4%	8%	47%	51%
F04	MATH4753	24	2%	5%	50%	52%
	MATH4773	27	3%	5%	48%	51%
SP05	PSY2003	94	8%	13%	39%	47%
SU05	ENGR3293	7	11%	19%	41%	52%
	MATH4753	34	6%	11%	43%	49%
	PSY2113	12	1%	1%	34%	35%

2.3.3 Correlation with Final Course Grades

To determine how the SCI compares to another external measure of statistics learning, posttest scores were compared to final course grades, (percentage grades, not letter grades). Limited final course grade data are available, and correlations were varied, see Table 2-8. In four of the nine classes for which data were available, significant positive correlations were obtained. In the remaining five classes, correlations were not significant, but note that the sample size was small.

Table 2-8: Correlations of SCI posttest score with final course grades.

Semester	Course	N	Correlation	P-value
Summer 2003	ENGR3293	21	0.59401	0.0045
	MATH4753	12	-0.02491	0.9387
Fall 2003	APMA311	102	0.33401	0.0006
	IE4553	17	0.06523	0.8036
	MATH 2023	31	0.43933	0.0134
	MATH4753	14	-0.06082	0.8364
Spring 2004	MATH4753	60	0.46079	0.0003
Summer 2005	MATH4753	12	0.24296	0.4467
	PSY2113	11	-0.33430	0.3150

Correlations were not expected to be particularly strong because the method of determining the final course grade is not a standardized procedure and can be quite variable from instructor to instructor. In general, course grade would be expected to be a measure of multiple aspects of the course including problem solving ability, writing ability, and possibly even attendance or participation, as well as the conceptual understanding construct targeted by the SCI. Individual instructor grading practices and philosophies can have a large impact on the distribution of grades as well. The distributions of letter grades by SCI quartile for four classes are shown in Figure 2-1. Notice that in general higher grades are associated with higher SCI quartiles, but not

exclusively. A number of "A" students received SCI scores in the lowest quartile for their class.

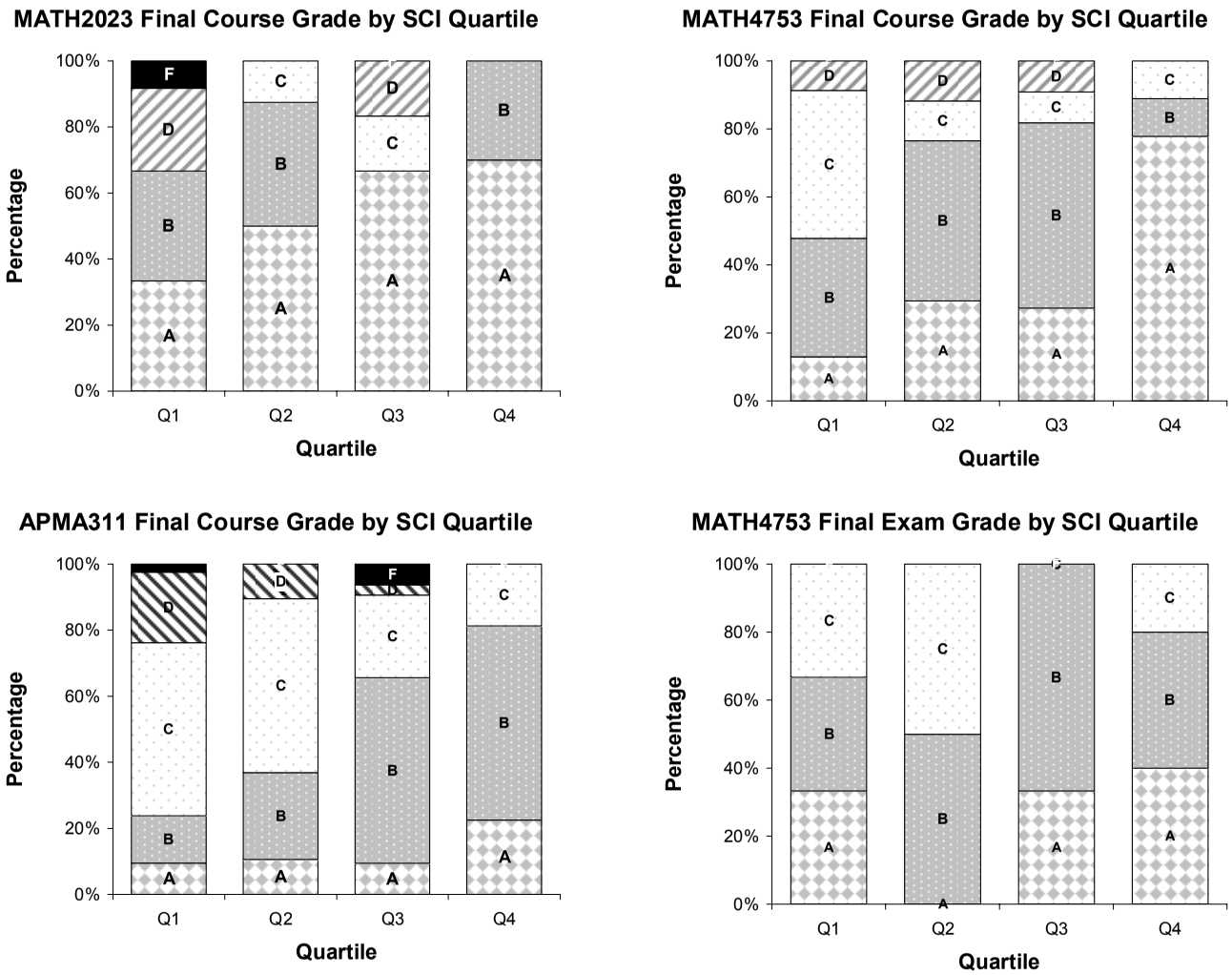


Figure 2-4: Distribution of course grades and SCI posttest scores by quartile.

2.3.4 Coefficient Alpha

The standard reliability measure under the classical test theory model is coefficient alpha. Since coefficient alpha is a whole test measure, it can only be computed for single administrations of the instrument. Coefficient alpha is sample dependent and can vary depending upon the characteristics of the sample. This can make it somewhat difficult to interpret. Coefficient alpha provides a lower bound for the test reliability.

Coefficient alpha values for each semester and also for each individual course are given in Table 2-9. For most semesters in our data set, coefficient alpha was around 0.7, but by class it is quite variable. For some classes, alpha was over 0.8, but for others it was much lower. While there were exceptions, classes that serve a predominately engineering population produced higher coefficient alpha values than those that were comprised of non-engineering majors. The differences in these populations discussed earlier as well as possible differences in topic coverage for these courses may have created more guessing in some non-engineering classes and this contributed to lower reliability estimates.

Table 2-9: Coefficient alpha for each semester and for individual courses.

		Alpha				Alpha	
Semester	Course	Pre	Post	Semester	Course	Pre	Post
Fall 2002			0.59	Spring 2004			0.72
	COMM2513		0.49		ENGR3293	0.69	0.66
	ENGR3293		0.39		MATH4753	0.68	0.75
	MATH4753		0.77		MATH4753	0.69	0.72
	MATH4753		0.71	Summer 2004	REU		0.67
	IE4553		0.37	Fall 2004			0.67
Summer 2003			0.74		ENGR3293	0.59	0.49
	ENGR3293	0.68	0.81		MATH4753	0.71	0.70
	MATH4753	0.68	0.86		MATH4773	0.74	0.73
	MATH4753	0.69	n/a		IE1071		0.62
	REU		0.6	Spring 2005			0.69
	IE1071		0.6		ENGR3293	0.66	0.77
Fall 2003			0.75		MATH4753	0.61	0.77
	ENGR3293	0.69	0.75		MATH4753	0.76	0.71
	MATH4753	0.71	n/a		IE1071		0.78
	MATH4753	0.67	0.72		PSY2003	0.43	0.59
	IE1071	0.7	0.73	Summer 2005			0.70
	APMA311	0.57	0.65		ENGR3293		0.80
	APMA311	0.66	0.58		MATH4753		0.63
	IE4553	0.62	0.56		MATH4753		0.67
	MATH2023	0.2	0.54		PSY2113		0.44

In order to improve the reliability of the SCI, future revisions will need to improve the individual item characteristics such as item discrimination and the correlation of the item score with the total test score. Improving distractor sets and rewriting questions to focus more on concepts and less of definitional understanding will be the strategy for future revisions to try to improve the overall test reliability.

Chapter 3: An Item Analysis of the Statistics Concepts Inventory

3.1 Item Analysis Tools

To guide the development process, the individual test items were analyzed each semester using tools from classical test theory. For each question, the difficulty, discrimination index, correlation with the total score, and alpha-if-item-deleted values were determined. This information, along with the distribution of responses and comments from focus groups, was used to make revisions. These statistics are briefly described in this section.

Item Difficulty: The *item difficulty* ranges from 0 to 1 and is simply the proportion of students who answered the item correctly. Questions with a low item difficulty are harder questions and those with a high item difficulty are easier. While there is no perfect item difficulty to try to achieve, items that are extremely easy or extremely difficult decrease the total variance of the test because they do not distinguish well between students (i.e. nearly all the students will answer correctly or incorrectly).

Discrimination Index: The *discrimination index* is a measure of how well an item separates students who have a high score on the total test from those who have a low score. The discrimination index for an item is calculated by comparing the proportion of students who answered the item correctly in two groups at the extremes of the total score distribution. We define the two groups as:

$$U = \{\text{students whose scores were at or above the 3rd quartile}\}$$

$$L = \{\text{students whose scores were at or below the 1st quartile}\}$$

For each item, the proportion of students who answered correctly is determined for each group: $p(U)$ and $p(L)$. The discrimination index is the difference $D = p(U) - p(L)$ (Kelly 1939).

The discrimination index attains its maximum value of 1 if every student in the upper group answered the question correctly and every student in the lower group answered it incorrectly. The minimum value of -1 is attained if every student in the lower group answered correctly while every student in the upper group answered incorrectly. Questions with a large, positive discrimination index are good, in that the "right" students are answering it correctly. That is, those students who are having trouble with the test as a whole are also having trouble with this question. This gives evidence that the question is measuring the same construct as the whole test and helps to contribute to the reliability of the test.

Questions with a low or negative discrimination index are equally or more difficult for those students in the upper group. These questions may need to be rewritten or reconsidered. It may be that questions with a negative discrimination index are measuring a different construct than the rest of the test. A low discrimination index will also occur with questions that are relatively easy (or hard), in which case most of the students in the lower group are also able to answer correctly (or most of the students in the upper group are also missing it).

Correlation with the Total Score: As part of the item analysis, for each question, the correlation between the item score and the total score for the remaining items is calculated. The total score for all items is not used because this score includes the individual item score as well and would artificially inflate the correlations. This is particularly noticeable on instruments that have a small number of items. Correlations will typically range from zero to 0.4, with values above 0.2 considered good (Nunnally 1967). While negative correlations are possible, they are not desirable and questions with negative or near zero correlations are candidates for elimination or rewriting. Questions that have higher correlations with the total test score are more discriminating and will contribute to a more reliable test.

Overall Alpha Rank: As discussed previously, coefficient alpha (α) is a commonly used estimate of the reliability of an instrument as a whole. When analyzing the individual items of an instrument, it is possible to gain some sense of how each individual item contributes to the overall test reliability by looking at the *alpha-if-item-deleted* statistic. This is determined by omitting the item from the data set and calculating α for all of the remaining items of the test. This value can then be compared to the overall coefficient alpha for all items. If the alpha-if-item-deleted value is smaller, then removing the item would lower the overall test reliability; therefore in terms of overall test reliability the item is good. If the alpha-if-item-deleted value is larger, this indicates a poor question in terms of overall test reliability because removing the question causes the test reliability to go up. These questions should be examined to see if they can be improved (e.g. by eliminating ambiguous wording or cues within the question, or reframing questions that involve too much guessing or that require recall only).

Both the overall coefficient alpha and the alpha-if-item-deleted statistics will vary from sample to sample. Generally, the difference between the alpha-if-item-deleted values and the overall coefficient alpha is quite small. In order to make better comparisons across semesters, we can look instead at the rank of the alpha-if-item-deleted statistic. By looking at the rankings of items across semesters, we can get a sense of which questions are ranked consistently high and low, and also which questions have rankings that are not consistent. This information can be used to make decisions when editing questions.

Table 3-1 shows an example of the alpha-if-item-deleted values and rankings from the spring 2005 semester post test data, in addition to the item correlations with the total score. The items are shown in rank order to more clearly demonstrate the relationships between the three measures. The coefficient alpha for the entire test was $\alpha = 0.692653$, while the range of alpha-if-item-deleted values was only from 0.670215 to 0.697479. Only seven questions had alpha-if-item-deleted values higher than the overall test alpha. The first column of Table 3-1 is the item number on the SCI. The second column of Table 3-1 is a master number that has been assigned to the item to facilitate tracking the items through the different versions of the instrument and back through each semester. These master numbers include a letter and number. The letter identifies the general topic area: probability, descriptive, inferential, or graphical.

Table 3-1: Alpha-if-item-deleted values and rankings from spring 2005 post test data.

Overall coefficient alpha= 0.692653

Deleted Item		Correlation with Total	Alpha-if-item-deleted	Rank
33	P8	0.43914	0.670215	1
12	D6	0.367124	0.675931	2
29	D10	0.343868	0.676854	3
35	I10A	0.342823	0.676974	4
26	D9	0.351383	0.677007	5
8	D3	0.322555	0.67854	6
31	P7A	0.300719	0.679669	7
20	I6	0.290422	0.680974	8
7	G1	0.286283	0.681551	9
27	I2	0.319493	0.682125	10
34	D5	0.267213	0.682409	11
22	I7	0.242095	0.683803	12
32	I9	0.228325	0.684794	13
23	D8A	0.22757	0.684826	14
30	G6	0.226007	0.685239	15
1	P1	0.221537	0.68524	16
11	D5	0.214865	0.685703	17
13	P4	0.262735	0.686012	18
25	G4	0.208108	0.686164	19
4	P2	0.191836	0.687241	20
9	D4	0.176566	0.68813	21
6	D2	0.164098	0.689193	22
2	I1	0.144577	0.690567	23
15	D7	0.141557	0.690682	24
21	P6	0.13858	0.690854	25
10	I2	0.137835	0.690914	26
24	G3	0.134072	0.691284	27
28	G5	0.129252	0.691306	28
17	D10	0.124778	0.691802	29
5	G6	0.098348	0.692276	30
37	G7	0.10678	0.693089	31
3	D1	0.09163	0.693597	32
14	G2	0.051259	0.694423	33
16	P5	0.067941	0.695775	34
19	I5	0.045453	0.696684	35
18	I4	0.037283	0.697204	36
36	I11	0.039484	0.697479	37

Factor Loading: If a test is assumed to model a single, common attribute (conceptual understanding of statistics, for instance), items on the test can vary in the way that they measure the attribute. Some items may be more discriminating and items may have unique variation due to context, question style, or other individual characteristics. Under the classical test theory model, an examinee's response on the jth item, X_j is modeled by $X_j = \Theta + \varepsilon_j$ where Θ is the examinee's true score for the attribute and ε_j is the random error of the observation. All items are assumed to measure the attribute equally well. Differences among the items on an exam can begin to be accounted for by using the single general factor model: $X_j = \mu_j + \lambda_j \Theta + \varepsilon_j$ where X_j is again an examinee's response to the jth item, λ_j is the *factor loading* and indicates how well the item measures the underlying attribute Θ, ε_j is the random error for the observation, and μ_j is the intercept term that accounts for the individual item difficulty (McDonald 1999).

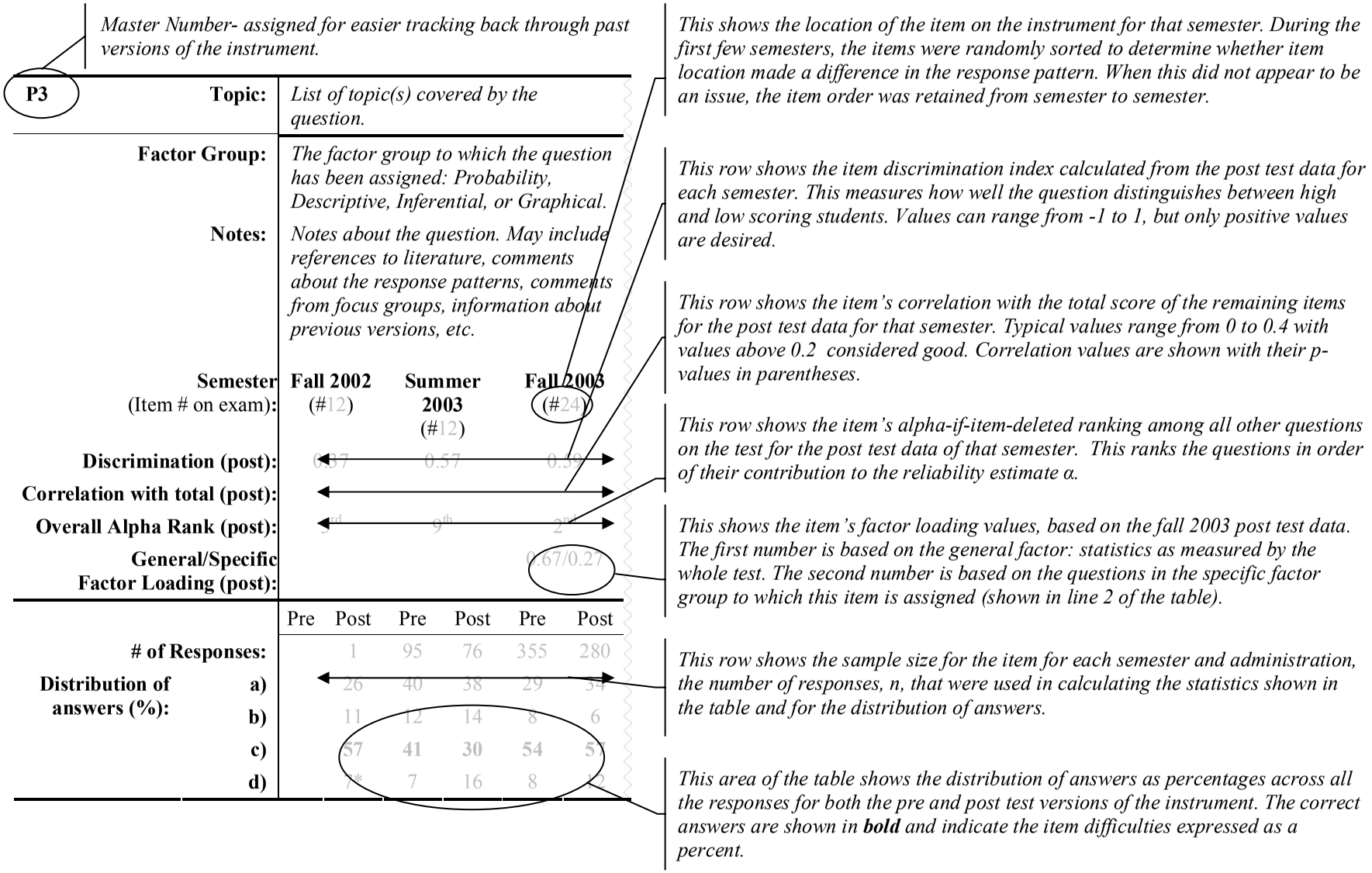
This model can be extended to more complex multiple factor models. For the analysis here, the data were fit to a model consisting of a general factor for which all the items were expected to have some loading and a group factor to which the items were assigned. The general factor is assumed to be a broad statistics factor and the group factors are based on more specific areas of statistics: probability, descriptive statistics, inferential statistics, and graphical. The items were assigned to one of the four factor groups based on their topical content. The factor analysis was carried out with the fall 2003 data using a maximum likelihood, nonlinear factor analysis method with TESTFACT 4.0™ software (Wood 2003).

3.2 Statistics Concept Inventory: Annotated Version

This type of analysis generates a large amount of data for each question. In order to summarize the data and be able to look for trends over time, an annotated version of the SCI is presented here. Each question is presented in the order it appears on the SCI followed by a table that includes the statistics generated from the item analyses and the response distributions for each semester. In addition, comments about the evolution of each question and relevant literature references are included. Figure 3-1 shows a portion of a sample table with brief explanations for interpreting each part.

.

Figure 3-1: A guide to reading the tables used in the annotated version of the SCI. A portion of a sample table is shown.



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 result 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 (Correct)
d) 0.95*0.98
e) 0.02+0.05

Table 3-2: Annotation for SCI question 1.

P1	**Topic:**	Probability, joint probabilities, marginal probabilities, contingency tables, probability rules, Bayes theorem													
	Factor Group:	Probability													
	Notes:	Earlier versions of this question did not contain the contingency table. In general, about half answered correctly on both the pre and post tests. The contingency table was added for the summer and fall 04 (pre) with all of the interior numbers filled in. Also, the data presentation was changed to frequency format instead of a percentage format. Approximately 80% answered correctly on the fall pre-test. The post test was changed to the above version because it was felt that we had given too much information. Option e) was omitted from the Fall 2004 version and from the Spring 2005 pretest as well as some of the post tests; these are designated with a -.													
	Semester	**Fall 2002**		**Summer 2003**		**Fall 2003**		**Spring 2004**		**Fall 2004**		**Spring 2005**		**Summer 2005**	
	(Item # on exam):	(#17)		(#19)		(#33)		(#1)		(#1)		(#1)		(#1)	
	Discrimination (post):	0.30		0.39		0.44		0.09		0.30		0.38		0.23	
	Correlation with total (post):	0.08		0.22		0.30		0.03		0.15		0.22		0.12	
	Overall Alpha Rank (post):	30th		8th		12th		33rd		24th		16th		28th	
	General/Specific Factor Loading (post):					0.44/-0.04									
		Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre‡	Post‡	Pre§	Post§	Pre	Post
	# of Responses:			95	76	355	280	143	94	123	108	105	163/48	73	60
	Distribution of answers (%): **a)**		13	2	7	6	8	8	6	11	8	14	13/0	8	22
	b)		*	*	**	***	***	†	†	7	15	23	18/8	12	13
	c)		**47**	**57**	**51**	**51**	**49**	**50**	**63**	**79**	**67**	**58**	**62/50**	**49**	**43**
	d)		*	*	**	***	***	†	†	2	9	5	6/2	0	2
	e)		28	17	24	26	25	29	26	-	-	-	-/40	27	17

*Alternate responses: 0.05 (9%), 0.05-0.02 (9%).
**Alternate responses (pre%, post%): 0.05 (4%, 3%), (0.05+0.02)*0.98 (19%, 14%).
***Alternate responses (pre%, post%): (0.05+0.02)*0.98 (14%, 18%).
†Alternate responses (pre%, post%): (0.05+0.02)*0.98 (1%, 5%).
‡Note that a different version of the question was given on the post test than on the pre test.
§Option e) was omitted from the pretest and some forms of the post test. The post test results are reported separately for each case.

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 (Correct)
    c) two-sample t test

Table 3-3: Annotation for SCI question 2.

I1	**Topic:**	Experimental design, choosing the appropriate statistical test, paired t-test													
	Factor Group:	Inferential													
	Notes:	Added Summer 2004. Initially included answer d) one-sample t-test which was deleted to eliminate possible ambiguity, the question could be conceived as a one sample t-test on the differences. This question requires the student to distinguish between z and t tests based on the given sample size and then recognize the two samples are not independent.													
	Semester (Item # on exam):	**Fall 2002**		**Summer 2003**		**Fall 2003**		**Spring 2004**		**Fall 2004** (#2)		**Spring 2005** (#2)		**Summer 2005** (#2)	
	Discrimination (post):									0.33		0.30		.41	
	Correlation with total (post):									0.21		0.14		0.16	
	Overall Alpha Rank (post):									18th		23rd		22nd	
	General/Specific Factor Loading:														
		Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
	# of Responses:									123	108	105	212	73	60
	Distribution of answers (%): **a)**									**16**	**19**	**22**	**22**	**18**	**8**
	b)									34	44	51	50	48	45
	c)									29	33	26	28	27	43
	d)									*					

*Additional response: one-sample t-test (18%).

3. In practice, which data collection strategy would be the best way to estimate the *mean household income* in the United States? One should measure the income level of

    a) every individual within the United States
    b) every household within the United States
    c) 1500 randomly selected individuals in the United States
    d) 1500 randomly selected households in the United States
    e) 10 random individuals within each of 150 random US counties
    f) 10 random households within each of 150 random US counties (Correct)

    New version beginning fall 2005:
    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 (Correct)
    d) 1500 is not a large enough sample

Table 3-4: Annotation for SCI question 3.

D1	**Topic:**	Data collection, Sampling, Stratified Random Sampling													
	Factor Group:	Descriptive													
	Notes:	New question for Fall 04 post test. Many distracters were included in this initial version to see what might appeal to students. Clearly, only two are. This question distinguishes between random sampling and stratified random sampling. Response d) of the new version is intended to capture the misconception that good samples must represent a large percentage of the population (Garfield 2003).													
	Semester (Item # on exam)**:**	**Fall 2002**		**Summer 2003**		**Fall 2003**		**Spring 2004**		**Fall 2004** (#3)		**Spring 2005** (#3)		**Summer 2005** (#3)	
	Discrimination (post):									0.22		.25		.48	
	Correlation with total (post):									0.11		0.09		0.23	
	Overall Alpha Rank (post):									27th		32nd		19th	
	General/Specific Factor Loading:														
		Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
	# of Responses:										108	105	212	73	60
	Distribution of answers (%): **a)**										0	0	1	1	0
	b)										6	8	8	16	3
	c)										3	0	4	0	5
	d)										68	42	54	47	57
	e)										0	2	1	0	3
	f)										**23**	**49**	**31**	**36**	**30**

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 (Correct)
   b) Urban
   c) Equally likely
   d) Both are extremely unlikely

Table 3-5: Annotation for SCI question 4.

P2	**Topic:**	Law of large numbers.												
	Factor Group:	Probability												
	Notes:	This question has had good alpha and discrimination values at each administration. This type of question has been well documented in the literature. An adaptation of Kahneman and Tversky's (1972) "Maternity Ward" problem, this version follows the frequency distribution format as described by Sedlmeier and Gigerenzer (1997). However, response (d) was not included in the studies they cite. The percent correct for this question on the SCI is not in the range of those cited for the frequency distribution format (median = 76%, range = 56-87%), but is much more in line with the percent correct found for the sampling distribution format (median = 33%, range = 7-59%). The most common answer is (c), and is evidence of reasoning using the *representative heuristic* and the misconception dubbed "belief in the law of small numbers" (Kahneman, et al. 1972). Students fail to consider the effect of sample size. Students who choose response (d) may also be reasoning according to the representative heuristic but feel that the sample proportion is not representative, therefore unlikely in any setting.												
	Semester (Item # on exam):	**Fall 2002** (#12)		**Summer 2003** (#12)		**Fall 2003** (#24)		**Spring 2004** (#4)		**Fall 2004** (#4)		**Spring 2005** (#4)		**Summer 2005** (#4)	
	Discrimination (post):	0.37		0.57		0.59		0.43		0.52		0.36		0.47	
	Correlation with total (post):	0.23		0.37		0.45		0.28		0.35		0.19		0.27	
	Overall Alpha Rank (post):	3rd		9th		2nd		10th		4th		20th		15th	
	General/Specific Factor Loading:					0.67/0.27									
		Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
	# of Responses:		**1**	**95**	**76**	**355**	**280**	**143**	**94**	**123**	**108**	**105**	**212**	**73**	**60**
	Distribution of answers (%): **a)**		26	40	38	29	34	31	24	36	40	35	35	27	28
	b)		11	12	14	8	6	6	6	5	6	5	6	7	5
	c)		57	41	30	54	57	45	57	46	47	45	46	52	53
	d)		7*	7	16	8	12	17	12	13	**	15	13	14	13

* Alternate response: Both are impossible. ** Response d) was omitted from the Fall 2004 post test.

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 (Correct)
    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

Table 3-6: Annotation for SCI question 5.

P3	**Topic:**	Independence, Probabilities of sequences, inference about population parameters
	Factor Group:	Probability
	Notes:	Changed significantly for Summer 2004. The earlier version had a *fair* coin flipped four times and asked for the most likely outcome of the fifth flip. The distracters were chosen to tap into known misconceptions identified from research. Nearly everyone chose (d) at the pre and post administrations. Focus group discussions indicated that students were answering by rote: if you are flipping a coin, the outcome is always equally likely. So, the question was amended to get away from the typical situation. We would like to test other versions of this question as well that ask them to say what they would do if they were "betting" on the outcome of the next flip. For this new version, 80-90% still answer (d) on the pre-test. Responses (a) and (c) may indicate reasoning by the representative heuristic. However, in both forms of the question (fair coin vs. unknown coin) these responses are not popular despite the large body of evidence of the presence of this type of thinking (Kahneman, Slovic and Tversky 1982). Response (d) may be accounted for by the *outcome approach* in which probabilities are based on the expected response to a single trial, (Konold, Pollatsek, Well, Lohmeier and Lipson 1993). They have found that this reasoning is not invoked when asked about the *least* likely outcome, and we have found this to be true in our work with a similarly worded question. (See question 16.) In comparing the two versions of this question, it seems that the context of the coin flipping is the driving force. In focus group interviews for the original version, students would comment that they keyed on the phrase "equally likely" and that heads and tails are *always* equally likely. With the unknown coin version, students commented that the coin may not be fair, but still chose equally likely because they were taught that that was always the answer. Few students considered all the information given them and were comfortable with breaking away from their conditioned answer.

Semester (Item # on exam):		Fall 2002 (#21)		Summer 2003 (#25)		Fall 2003 (#6)		Spring 2004 (#5)		Fall 2004 (#5)		Spring 2005 (#5)		Summer 2005 (#5)	
Discrimination (post):		0.48		0.24		0.11		0.07		0.15		0.17		0.28	
Correlation with total (post):		0.35		0.27		0.19		0.27		0.12		0.10		0.24	
Overall Alpha Rank (post):		10th		15th		24th		25th		26th		30th		18th	
General/Specific Factor Loading:						0.39/-0.26									
		Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
	# of Responses:			95	76	355	280	143	94	123	108	105	212	73	60
Distribution of answers (%):	**a)**		9	1	4	5	3	1	2	1	3	6	5	3	0
	b)		4	6	3	4	3	4	1	**11**	**12**	**9**	**18**	**12**	**17**
	c)		12	7	4	5	2	6	2	4	1	2	5	1	3
	d)		**74***	**84**	**88***	**86**	**93**	**89**	**95**	83	84	84	72	84	80

*Additional response (pre%, post%): Tails because there have been so many heads, we are due a tail (1%, 2%).

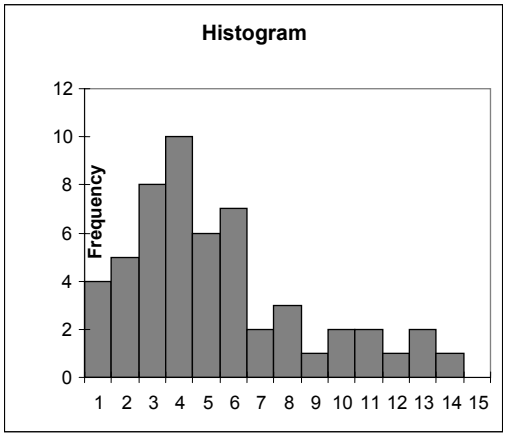
6. An Olympic track team consists of 6 sprinters (2 compete in the 100 meter event, 2 compete in the 200 meter event, and the remaining 2 compete in the 400 meter event). For which of the following samples would you expect to calculate the largest variance?

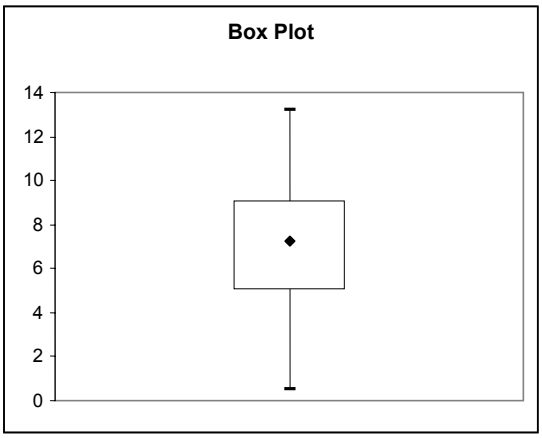
    a) A randomly selected sprinter's running times for 15 trials of the 200 meter event
    b) The track team's (all six members) running times for the 200 meter event
    c) A randomly selected sprinter's running times for 5 trials each of the 100 meter, 200 meter and 400 meter events
    d) The track team's running times for the 100 meter, 200 meter, and 400 meter events, each person running all three events (Correct)

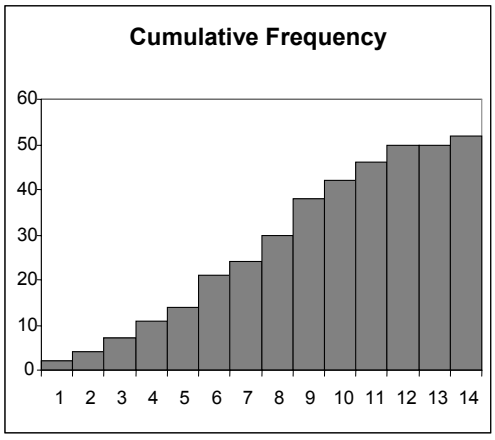
Table 3-7: Annotation for SCI question 6.

D2	**Topic:**	Variance													
	Factor Group:	Descriptive													
	Notes:	This question was added Spring 04 as a companion question to #30. Students seem to have a lot of trouble understanding variance in the graphical context of #30, and we wanted to try to distinguish between that, and a similar non-graphical situation. The Spring 04 version asked for which would have the *least* variation, hence (a) was the correct response. It was changed to *largest* to more closely match the graphical question in Fall 2004. The question was re-worded slightly after the fall 04 pre-test to clarify who was running which events; the phrases "randomly selected", "(all six members)" and "each person running all three events" were added. The phrase randomly selected seems to have made choice (c) more attractive and might imply that some students associate randomness and variability. The proportion of students answering this question correctly is much higher than that of students answering question #30 correctly, typically only 20-30% on the post test. When asked in focus groups how they thought about this problem, most students described thinking about the set of running times as a collection, i.e., many different numbers equals more variability. When specifically asked if they saw similarities between the two questions, they did not conceive of the graphical question in the same way.													
	Semester (Item # on exam):	**Fall 2002**		**Summer 2003**		**Fall 2003**		**Spring 2004** (#6)		**Fall 2004** (#6)		**Spring 2005** (#6)		**Summer 2005** (#6)	
	Discrimination (post):							0.028		0.15		0.29		0.16	
	Correlation with total (post):							0.08		-0.01		0.16		-0.04	
	Overall Alpha Rank (post):							31st		34th		22nd		36th	
	General/Specific Factor Loading:														
		Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
								143	94	123	108	105	212	73	60
Distribution of answers (%):	**a)**							76	82	7	3	5	3	4	7
	b)							8	10	10	19	18	19	16	20
	c)							9	3	14	29	25	21	30	25
	d)							**7**	**5**	**69**	**49**	**52**	**57**	**49**	**47**

7. Three of the following are graphical presentations of the same set of data. Which of the graphs 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 (Correct)
b) Box Plot
c) Cumulative Frequency
d) Stem and Leaf

Table 3-8: Annotation for SCI question 7.

G1	**Topic:**	Methods of displaying data, interpreting graphs.													
	Factor Group:	Graphical													
	Notes:	This question was added summer 2004. This question tests the ability to read and interpret four commonly used graphical representations of data, and to identify key characteristics in order to make comparisons. Answers are typically spread pretty evenly among all 4 choices on the pretest with (d) being slightly more popular. The histogram was chosen to be the different graph since it was expected to be the most familiar to students and it is thought that many will be able to distinguish it quickly from the stem and leaf. However, nearly half are unable to do this.													
	Semester (Item # on exam)**:**	**Fall 2002**		**Summer 2003**		**Fall 2003**		**Spring 2004**		**Fall 2004** (#7)		**Spring 2005** (#7)		**Summer 2005** (#7)	
	Discrimination (post):									0.48		0.38		0.41	
	Correlation with total (post):									0.31		0.28		0.25	
	Overall Alpha Rank (post):									7th		9th		16th	
	General/Specific Factor Loading:														
		Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
	# of Responses:									123	108	105	212	73	60
	Distribution of answers (%): **a)**									**20**	**38**	**10**	**25**	**26**	**30**
	b)									21	16	30	19	19	17
	c)									22	17	28	22	22	30
	d)									34	30	33	33	33	23

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 (Correct)
   d) None of these are always true

Table 3-9: Annotation for SCI question 8.

D3	**Topic:**	Percentiles													
	Factor Group:	Descriptive													
	Notes:	Most of the answers are split between (c) and (d). This was expected to be a relatively easy question since it was thought that most students would have had several exposures to percentiles in their academic career. Students choosing d) are most likely keying in on the "none" and "always" instead of a particular statistics related misconception. This fact was mentioned in focus groups. This question will be eliminated and replaced with a question focusing on the non-linearity of percentiles for the Fall 2005 post test.													
	Semester (Item # on exam):	**Fall 2002** (#10)		**Summer 2003** (#9)		**Fall 2003** (#31)		**Spring 2004** (#8)		**Fall 2004** (#8)		**Spring 2005** (#8)		**Summer 2005** (#8)	
	Discrimination (post):	0.45		0.22		0.36		0.41		0.28		0.56		0.12	
	Correlation with total (post):	0.30		0.10		0.30		0.25		0.27		0.32		0.12	
	Overall Alpha Rank (post):	4th		20th		15th		17th		14th		6th		26th	
	General/Specific Factor Loading:					0.47/0.05									
		Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
	# of Responses:		138	195	76	355	280	143	94	123	108	105	212	73	60
	Distribution of answers (%): **a)**		4	3	0	5	2	1	1	2	2	2	1	3	0
	b)		17	10	7	10	9	7	5	7	7	14	8	16	5
	c)		**70**	**72**	**82**	**70**	**74**	**64**	**70**	**71**	**81**	**54**	**67**	**53**	**78**
	d)		9	16	12	14	16	27	23	20	10	30	24	26	17

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. (Correct)
   d) It cannot increase without changing the median.

Table 3-10: Annotation for SCI question 9.

D4	**Topic:**	Median													
	Factor Group:	Descriptive													
	Notes:	This question has not undergone any changes since the first administration. We have found that many students write out the list on numbers in order on their test booklet to determine the median in answering this question. This would indicate that even though a large percentage of students answer this question correctly, many do not have an intuitive feel for the median. Students choosing response (d) may be confusing the median with the mean, consistent with research findings (Zawojewski and Shaughnessy 2000, Garfield 2002), or they may have an incorrect understanding of the median.													
	Semester (Item # on exam):	**Fall 2002** (#1)		**Summer 2003** (#1)		**Fall 2003** (#16)		**Spring 2004** (#9)		**Fall 2004** (#9)		**Spring 2005** (#9)		**Summer 2005** (#9)	
	Discrimination (post):	0.53		0.38		0.59		0.30		0.62		0.25		0.56	
	Correlation with total (post):	0.29		0.26		0.43		0.28		0.40		0.18		0.39	
	Overall Alpha Rank (post):			12th		3rd		12th		2nd		21st		7th	
	General/Specific Factor Loading:					0.64/0.34									
		Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
	# of Responses:		137	95	76	355	280	143	94	123	108	105	212	73	60
	Distribution of answers (%): **a)**		1	2	3	1	2	2	5	2	0	2	2	4	3
	b)		15	13	7	7	5	9	5	5	7	13	7	12	3
	c)		**53**	**61**	**78**	**71**	**71**	**63**	**75**	**65**	**74**	**48**	**78**	**63**	**77**
	d)		31	24	13	20	21	25	15	27	19	37	13	21	17

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)

Table 3-11: Annotation for SCI question 10.

I2	**Topic:**	Hypothesis testing, defining the alternative hypothesis													
	Factor Group:	Inferential statistics													
	Notes:	Initially, this question was posed using symbols in the answer choices. It was changed to words for the Spring 2004 version to eliminate any possible confusion caused by the symbolic language and to focus more on concept. Students choosing a) may be confusing the alternate hypothesis with the null hypothesis, response b) would indicate that students were failing to distinguish between a one and two tailed situation. Response c) has the inequality reversed.													
	Semester (Item # on exam)**:**	**Fall 2002**		**Summer 2003** (#23)		**Fall 2003** (#22)		**Spring 2004** (#10)		**Fall 2004** (#10)		**Spring 2005** (#10)		**Summer 2005** (#10)	
	Discrimination (post):			0.17		0.18		0.22		0.38		0.26		0.34	
	Correlation with total (post):			0.09		0.09		0.15		0.24		0.14		0.13	
	Overall Alpha Rank (post):			27th		29th		28th		17th		26th		25th	
	General/Specific Factor Loading:					0.12/0.27									
		Pre	Post	Pre*	Post*	Pre*	Post*	Pre	Post	Pre	Post	Pre	Post	Pre	Post
	# of Responses:			95	76	355	280	143	94	123	108	105	212	73	60
	Distribution of answers (%): a)			19	18	15	10	27	33	29	38	41	34	30	37
	b)			8	1	7	4	12	19	14	10	12	13	16	18
	c)			25	21	25	15	14	9	24	17	12	12	11	7
	d)			**47**	**59**	**50**	**72**	**46**	**40**	**32**	**35**	**34**	**40**	**42**	**38**

*This version contained "symbolic" responses, e.g. H_1 : mean = 20.

11. Which of the following statistics is least impacted by extreme outliers?

a) range
b) 3rd quartile (Correct)
c) mean
d) variance

Table 3-12: Annotation for SCI question 11.

D5 **Topic:**	Descriptive statistics, impact of outliers													
Factor Group:	Descriptive													
Notes:	This question was added Spring 04.													
Semester (Item # on exam)**:**	**Fall 2002**		**Summer 2003**		**Fall 2003**		**Spring 2004** (#11)		**Fall 2004** (#11)		**Spring 2005** (#11)		**Summer 2005** (#11)	
Discrimination (post):							0.34		0.38		0.30		0.49	
Correlation with total (post):							0.29		0.32		0.21		0.28	
Overall Alpha Rank (post):							13th		6th		17th		13th	
General/Specific Factor Loading:														
	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
# of Responses:							143	94	123	108	105	212	73	60
Distribution of answers (%): **a)**							6	6	6	6	12	6	10	7
b)							**62**	**66**	**65**	**66**	**47**	**62**	**53**	**62**
c)							17	9	11	13	24	14	14	13
d)							14	19	16	16	17	17	23	18

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) $\frac{3.24(2)+3.80(4)}{6}$ (Correct)

d) It is not possible to calculate the students overall GPA without knowing his GPA for each individual semester.

Table 3-13: Annotation for SCI question 12.

D6 **Topic:**	Weighted mean													
Factor Group:	Descriptive													
Notes:	Consistent with the findings of Pollatsek, Lima, and Well (1981). Response a) and b) would indicate that students possess only an algorithmic understanding of the mean. Response d) would also indicate that students are unable to move between individual data points and sums of the data. This would likely indicate difficulty constructing means from frequency data as well, though this has not been assessed specifically here.													
Semester (Item # on exam):	**Fall 2002** (#24)		**Summer 2003** (#28)		**Fall 2003** (#2)		**Spring 2004** (#12)		**Fall 2004** (#12)		**Spring 2005** (#12)		**Summer 2005** (#12)	
Discrimination (post):	0.45		0.49		0.44		0.38		0.37		0.53		0.68	
Correlation with total (post):	0.20		0.25		0.36		0.28		0.31		0.37		0.47	
Overall Alpha Rank (post):	23rd		18th		9th		14th		10th		2nd		4th	
General/Specific Factor Loading:					0.53/0.06									
	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
# of Responses:		137	95	76	355	280	143	94	123	108	105	212	73	60
Distribution of answers (%): **a)**		27	13	15	14	10	12	10	14	7	23	12	21	10
b)		15	11	9	4	7	8	3	10	6	11	5	5	12
c)		**45**	**68**	**74**	**72**	**79**	**65**	**76**	**62**	**73**	**50**	**71**	**59**	**63**
d)		8*	8	1	10	5	15	12	14	14	16	13	15	15

*Additional response: (3.24 + 3.80)/6, (6%).

13. 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 you have to dial?

    a) 10
    b) 15
    c) 25 (Correct)
    d) There is no way to estimate

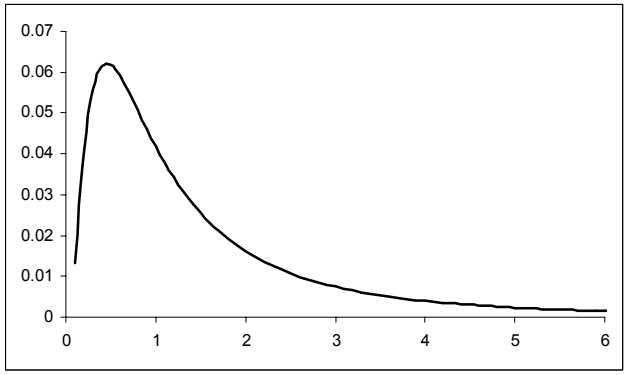
    Beginning Spring 2005: You have called your cell phone provider to discuss a discrepancy on your billing statement. Your call was received and placed on hold to "await the next available service representative". You are told that the average waiting time is 6 minutes. You have been waiting on hold for 4 minutes. How many more minutes do you anticipate you will have to wait before you speak to a service representative?

    a) 2
    b) 4
    c) 6 (Correct)
    d) There is no way to estimate

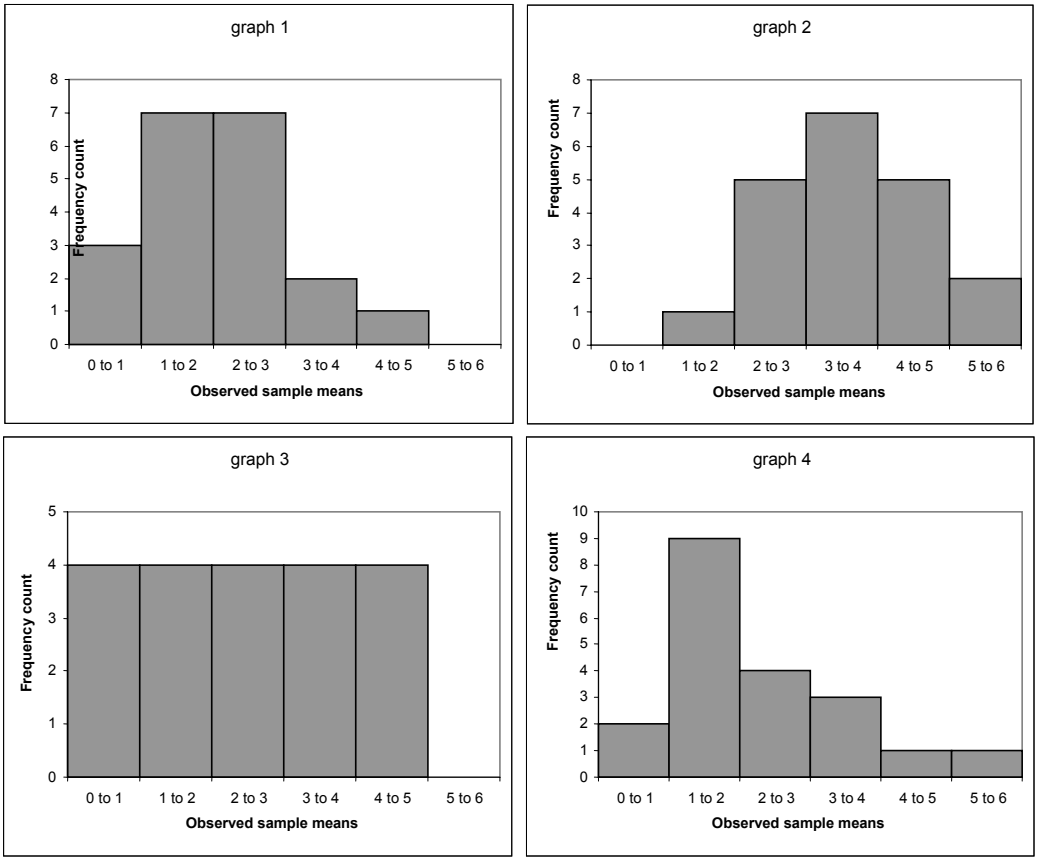
Table 3-14: Annotation for SCI question 13.

P4/P4a	Topic:	Memoryless property, Geometric distribution, Exponential distribution													
	Factor Group:	Probability													
	Notes:	Almost no one gets this question right. The discrimination for this question has not been consistent and is low. However, this is partially due to the overall low percentage of people who answer correctly. In Spring 2005, for example, no one at or below the 25th percentile answered this question correctly, and only 17% of those at or above the 75th percentile answered correctly. However, 4 of the top 5 scoring students answered correctly. Only 10 other students answered correctly, with 5 of them above the 75th percentile. The question was replaced with the new version because it was felt that the context of the original question was becoming obsolete													
	Semester (Item # on exam):	**Fall 2002** (#4)		**Summer 2003** (#4)		**Fall 2003** (#29)		**Spring 2004** (#13)		**Fall 2004** (#13)		**Spring 2005** (#13)		**Summer 2005** (#13)	
	Discrimination (post):	0.08		0.23		0.26		0.04		-0.06		0.17		-0.06	
	Correlation with total (post):	0.00		0.25		0.28		0.04		-0.16		0.26		-0.18	
	Overall Alpha Rank (post):	23rd		19th		18th		30th		33rd		18th		35th	
	General/Specific Factor Loading:					0.41/-0.44									
		Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post*	Pre	Post
	# of Responses:		138	95	76	355	280	143	94	123	108	105	163/106	73	60
	Distribution of answers (%): **a)**		51	67	84	54	51	62	70	68	56	67	66/63	64	48
	b)		4	3	1	6	5	4	1	2	3	7	7/5	1	3
	c)		**3**	**4**	**9**	**4**	**12**	**4**	**3**	**2**	**5**	**1**	**4/9**	**4**	**7**
	d)		42	25	5	35	31	30	26	27	37	26	23/23	30	42

*The second form of this question as shown above was introduced during the Spring 2005 post test. Some students had only one version of the question, some both. The answer distributions are shown for both versions as modem pool/cell phone.



14. From the above probability density function, 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 (Correct)
b) graph 2
c) graph 3
d) graph 4

Table 3-15: Annotation for SCI question 14.

G2 **Topic:**	Central limit theorem (graphical)													
Factor Group:	Graphical													
Notes:	Added summer 04. Students choosing d) may either be confusing a sampling distribution with a frequency distribution, or may believe that a sampling distribution should look like the parent distribution (Chance, delMas and Garfield 2004). Those choosing b) may be partially applying the central limit theorem, recognizing that the distribution should be normal but failing to center it appropriately.													
Semester (Item # on exam):	**Fall 2002**		**Summer 2003**		**Fall 2003**		**Spring 2004**		**Fall 2004** (#14)		**Spring 2005** (#14)		**Summer 2005** (#14)	
Discrimination (post):									0.05		0.10		0.21	
Correlation with total (post):									-0.07		0.05		0.13	
Overall Alpha Rank (post):									35th		33rd		24th	
General/Specific Factor Loading:														
	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
# of Responses:									123	108	105	212	73	60
Distribution of answers (%): **a)**									**11**	**17**	**15**	**16**	**10**	**15**
b)									8	7	14	14	7	8
c)									14	9	11	8	11	10
d)									65	67	59	62	73	67

15. For the following set of numbers, which measure will most accurately describe the central tendency?
3, 4, 5, 6, 6, 8, 10, 12, 19, 36, 83

a) Mean
b) Median (Correct)
c) Mode
d) Standard deviation

Table 3-16: Annotation for SCI question 15.

D7 **Topic:**	Measures of central tendency													
Factor Group:	Descriptive													
Notes:	Students choosing a) may believe that the mean is always the best measure of central tendency, even when there are outliers. Response a) or c) may indicate that average is not viewed as a representative number, or it may indicate confusion between the three measures (Watson and Moritz 1999, Garfield, Hogg, Schau and Whittinghill 2002).													
Semester (Item # on exam):	**Fall 2002**		**Summer 2003** (#10)		**Fall 2003** (#3)		**Spring 2004** (#15)		**Fall 2004** (#15)		**Spring 2005** (#15)		**Summer 2005** (#15)	
Discrimination (post):			0.51		0.28		0.197		0.40		0.33		0.36	
Correlation with total (post):			0.34		0.17		0.06		0.19		0.14		0.12	
Overall Alpha Rank (post):			16th		25th		32nd		19th		24th		27th	
General/Specific Factor Loading:					0.25/0.17									
	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
# of Responses:			95	76	355	280	143	94	123	108	105	212	73	60
Distribution of answers (%): **a)**			23	21	24	24	17	19	22	21	23	20	23	18
b)			**58**	**71**	**62**	**67**	**58**	**59**	**56**	**60**	**53**	**60**	**53**	**62**
c)			11	3	4	4	6	8	11	9	13	10	12	10
d)			8	5	9	5	17	13	10	9	10	10	11	10

16. A standard deck of 52 cards consists of 13 cards in each of 4 suits: hearts (♥), diamonds (♦), clubs (♣), and spades (♠). 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) ♥♥♥♥♥
    b) ♣♦♥♠♣
    c) ♠♥♠♥♠
    d) All three are equally likely. (Correct)

Table 3-17: Annotation for SCI question 16.

P5 **Topic:**	Probabilities of sequences, independence													
Factor Group:	Probability													
Notes:	This question was originally posed with a sequence of coin tosses. It was changed during Summer 2004 to cards so it would have less resemblance to the other coin toss question. The fall 03 through Spring 04 data below refer to the coin version. The fall 02 and summer 03 versions also had reasons attached to each answer such as a) HHHTH because the numbers of heads and tails should be more equal. The reasons were removed to avoid leading students' reasoning. Usually around 45% answered the coin toss version correctly. Incorrect responses to this question likely indicates students are reasoning using the representative heuristic or that they fail to acknowledge the independence of the card draws (Kahneman and Tversky 1982).													
Semester (Item # on exam):	**Fall 2002**		**Summer 2003**		**Fall 2003**		**Spring 2004**		**Fall 2004** (#16)		**Spring 2005** (#16)		**Summer 2005** (#16)	
Discrimination (post):					0.38		0.26		0.15		0.21		0.24	
Correlation with total (post):									0.03		0.07		0.04	
Overall Alpha Rank (post):					21st		29th		32nd		34th		30th	
General/Specific Factor Loading:					0.37/-0.45									
	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
# of Responses:									123	108	105	212	73	60
Distribution of answers (%): a)									36	39	36	35	45	22
b)									12	7	9	8	10	3
c)									2	1	8	3	1	0
d)									**50**	**53**	**52**	**54**	**44**	**75**

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 (Correct)
d) 5% of the measurements should be considered outliers

Table 3-18: Annotation for SCI question 17.

I3	**Topic:**	Confidence intervals													
	Factor Group:	Inference													
	Notes:	Answer choices have been edited over time to eliminate unpopular distracters, to restructure the answers so they look more similar in type and length, and to make them more precise. Response b) is a very strong distracter.													
	Semester (Item # on exam)**:**	**Fall 2002** (#8)		**Summer 2003** (#8)		**Fall 2003** (#19)		**Spring 2004** (#17)		**Fall 2004** (#17)		**Spring 2005** (#17)		**Summer 2005** (#17)	
	Discrimination (post):	0.11		0.48		0.55		0.51		0.49		0.22		0.36	
	Correlation with total (post):	0.02		0.29		0.37		0.38		0.30		0.12		0.24	
	Overall Alpha Rank (post):	29th		6th		5th		3rd		9th		29th		17th	
	General/Specific Factor Loading:					0.49/0.32									
		Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
	# of Responses:			95	76	355	280	143	94	123	108	105	212	73	60
Distribution of answers (%):	**a)**			15	4	12	8	10	9	6	11	15	14	8	7
	b)			32	32	43	50	41	36	46	38	31	41	49	38
	c)			**44**	**61**	**26**	**38**	**30**	**50**	**33**	**49**	**38**	**40**	**27**	**53**
	d)			*	*	17	4	17	4	13	2	15	7	15	2

*Alternate responses (pre%, post%): It is probable that 95% of the confidence intervals will be identical (2%, 0%), None of the above (6%, 3%)

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 0.10. Which of the following must be correct?

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 H_0 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
e) The probability that the null hypothesis is actually true 0.9

Table 3-19: Annotation for SCI question 18.

I4	**Topic:**	Hypothesis testing, p-value													
	Factor Group:	Inference													
	Notes:	This question has remained practically unchanged since the beginning. The percentage of students answering correctly has varied widely among individual classes, anywhere from 20% to 50%. Option e) was removed for the Fall 2004 version but replaced for Spring 2005 as IRT analysis showed the item was more discriminating with the option in place. The response distribution indicates widespread confusion about p-value and hypothesis testing. Response a) becomes more attractive while the number of correct responses almost always falls after instruction.													
	Semester (Item # on exam)**:**	**Fall 2002** (#26)		**Summer 2003** (#18)		**Fall 2003** (#25)		**Spring 2004** (#18)		**Fall 2004** (#18)		**Spring 2005** (#18)		**Summer 2005** (#18)	
	Discrimination (post):	0.17		0.63		0.21		0.43		0.34		0.17		0.47	
	Correlation with total (post):	0.13		0.50		0.17		0.28		0.16		0.03		0.42	
	Overall Alpha Rank (post):	24th		2nd		23rd		9th		22nd		36th		6th	
	General/Specific Factor Loading:					0.23/0.28									
		Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
	# of Responses:		131	95	76	355	280	143	94	123	108	105	212	73	60
	Distribution of answers (%): **a)**		25	17	21	21	30	17	27	15	37	17	33	19	47
	b)		18	11	16	8	16	11	11	15	22	6	10	14	15
	c)		**22**	**31**	**36**	**38**	**20**	**33**	**39**	**44**	**30**	**48**	**33**	**30**	**20**
	d)		22	27	13	18	20	24	16	20	10	19	20	23	13
	e)		12	11	14	9	13	6	7	*	*	11	3	10	5

* Option e) was removed for the Fall 2004 administration.

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 the same basic shape as a normal distribution but has less area in the tails
d) a & b are both true (Correct)
e) a, b & c are all true

Table 3-20: Annotation for SCI question 19.

I5	**Topic:**	t-distribution													
	Factor Group:	Inferential													
	Notes:	This question was rewritten for the Fall 04 version. "It describes a population" option was replaced with current (a) and option (e) was changed from b and c are true. This question assesses understanding of basic properties of the t-distribution. Most students recognize that the t-distribution is used for small samples. The most common misunderstanding seen here is the confusion of the amount of area in the tails relative to the normal distribution.													
	Semester (Item # on exam):	**Fall 2002** (#15)		**Summer 2003** (#16)		**Fall 2003** (#1)		**Spring 2004** (#19)		**Fall 2004** (#19)		**Spring 2005** (#19)		**Summer 2005** (#19)	
	Discrimination (post):	0.15		0.12		0.13		0.22		-0.07		0.12		0.28	
	Correlation with total (post):	0.01		0.03		0.09		0.12		-0.09		0.05		0.30	
	Overall Alpha Rank (post):	31st		31st		28th		27th		37th		35th		14th	
	General/Specific Factor Loading:														
		Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
	# of Responses:									123	108	105	212	73	60
	Distribution of answers (%): **a)**									8	21	4	16	5	22
	b)									6	5	13	8	14	7
	c)									11	6	5	3	4	7
	d)									**31**	**29**	**40**	**33**	**36**	**18**
	e)									40	39	38	38	36	47

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)

Table 3-21: Annotation for SCI question 20.

I6	**Topic:**	Parameter estimation, mean, standard deviation, sample													
	Factor Group:	Inferential													
	Notes:	This question has historically been one of the best in terms of alpha if deleted for each administration, with consistently good discrimination. Approximately half chose that they were more surprised by a single observation than by the shifted mean. This would indicate that they are not recognizing the sampling process, the parameter estimation process, or that the mean is a function of the entire sample and much less likely to be very different from the population mean than a single observation.													
	Semester (Item # on exam):	**Fall 2002** (#5)		**Summer 2003** (#5)		**Fall 2003** (#4)		**Spring 2004** (#20)		**Fall 2004** (#20)		**Spring 2005** (#20)		**Summer 2005** (#20)	
	Discrimination (post):	0.14		0.69		0.45		0.51		0.56		0.51		0.34	
	Correlation with total (post):	0.18		0.42		0.30		0.29		0.37		0.29		0.15	
	Overall Alpha Rank (post):	6th		1st		13th		7th		3rd		8th		23rd	
	General/Specific Factor Loading:														
		Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
	# of Responses:		137	95	76	355	280	143	94	123	108	105	212	73	60
	Distribution of answers (%): a)		5	5	4	29	31	19	29	25	21	23	41	29	32
	b)		44*	42*	43*	20	19	20	15	20	17	35	18	18	17
	c)		6	4	7	10	11	15	10	9	9	22	13	14	17
	d)		**9****	**48**	**46**	**40**	**38**	**43**	**47**	**45**	**52**	**20**	**29**	**40**	**35**

*On this version, response (b) was "One woman with height 77 inches".
**Additional response e) I am not surprised by anything (37%).

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)

Table 3-22: Annotation for SCI question 21.

P6	**Topic:**	Probabilities, independence, joint probability.													
	Factor Group:	Probability													
	Notes:	This question was introduced Spring 04. Students answering incorrectly may not be able to apply the probability laws for independent events. Students may be making use of the *adjustment and anchoring* heuristic which leads to overestimating the probability of conjunctive events (London and Chicago) and underestimating that of disjunctive events (London or Chicago) (Kahneman, et al. 1982).													
	Semester (Item # on exam):	**Fall 2002**		**Summer 2003**		**Fall 2003**		**Spring 2004** (#21)		**Fall 2004** (#21)		**Spring 2005** (#21)		**Summer 2005** (#21)	
	Discrimination (post):									0.37		0.27		0.05	
	Correlation with total (post):							0.19		0.17		0.14		-0.07	
	Overall Alpha Rank (post):									21st		25th		36th	
	General/Specific Factor Loading:														
		Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
	# of Responses:							143	94	123	108	105	212	73	60
	Distribution of answers (%): **a)**							1	6	5	3	5	4	3	2
	b)							34	32	37	35	26	27	25	22
	c)							19	9	8	4	19	8	4	3
	d)							**45**	**53**	**48**	**58**	**50**	**61**	**68**	**73**

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 (Correct)
    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

Table 3-23: Annotation for SCI question 22.

I7	**Topic:**	P-value, sample size effect													
	Factor Group:	Inferential													
	Notes:	This question assesses understanding of the relationship between p-value and sample size. It has had consistently high discrimination values and high alpha-if-item-deleted rankings. This question, as with the other p-value questions, has been consistently difficult for students, indicating that p-value is a difficult concept.													
	Semester (Item # on exam):	**Fall 2002** (#25)		**Summer 2003** (#29)		**Fall 2003** (#34)		**Spring 2004** (#22)		**Fall 2004** (#22)		**Spring 2005** (#22)		**Summer 2005** (#22)	
	Discrimination (post):	0.34		0.50		0.57		0.47		0.52		0.36		0.54	
	Correlation with total (post):	0.20		0.34		0.36		0.38		0.29		0.24		0.31	
	Overall Alpha Rank (post):	20th		7th		6th		4th		11th		12th		10th	
	General/Specific Factor Loading:					0.50/0.18									
		Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
	# of Responses:		137	95	76	355	280	143	94	123	108	105	212	73	60
	Distribution of answers (%): **a)**		**33**	**38**	**51**	**30**	**42**	**24**	**40**	**28**	**42**	**29**	**45**	**27**	**35**
	b)		35	24	21	32	22	29	17	24	23	30	25	30	18
	c)		12	13	9	9	11	14	10	15	3	13	10	14	10
	d)		*	23	18	25	25	24	33	29	31	29	19	25	37

*Alternate response: Insufficient information (20)%.

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

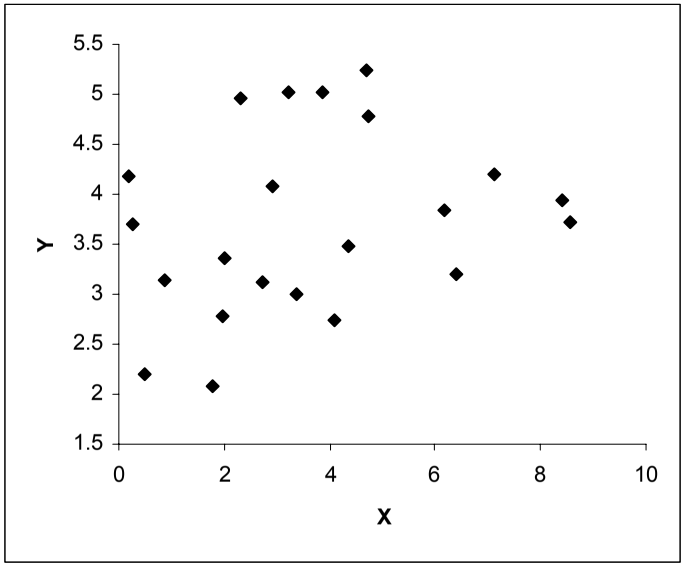
a) I & II (Correct)
b) II & III
c) I & III
d) All 3

Table 3-24: Annotation for SCI question 23.

D8 **Topic:**	Normal distribution													
Factor Group:	Descriptive													
Notes:	(a) and (d) are the most popular answers. Response set changed for Fall 2004 version, eliminated I only and II only which were very rarely chosen and added II & III for consistency within the response set. In previous version, answers were divided almost evenly between I&II and All 3. Fall 2002 and Summer 2003 versions asked about "Age of pennies in circulation" instead of "Shoe size of men".													
Semester (Item # on exam):	**Fall 2002**		**Summer 2003**		**Fall 2003** (#10)		**Spring 2004** (#23)		**Fall 2004** (#23)		**Spring 2005** (#23)		**Summer 2005** (#23)	
Discrimination (post):	0.13		0.45		0.63		0.37		0.44		0.35		0.48	
Correlation with total (post):					0.37		0.17		0.31		0.23		0.30	
Overall Alpha Rank (post):	14th		11th		6th		19th		8th		14th		11th	
General/Specific Factor Loading:					0.52/0.05 *									
	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
# of Responses:					355	280	143	94	123	108	105	212	73	60
Distribution of answers (%): **a)**					**44**	**47**	**42**	**45**	**45**	**63**	**30**	**41**	**38**	**38**
b)					**	**	***	***	2	6	10	4	7	8
c)					10	6	14	4	11	6	17	9	14	7
d)					39	43	38	44	41	25	42	45	41	47

*In probability group, have since moved it to descriptive. **Alternate responses (pre%, post%): I only (4%, 3%), II only (3%,1%).
***Alternate responses (pre%, post%): I only (3%, 6%), II only (0%, 1%).

24. Estimate the correlation coefficient for the two variables X and Y from the scatter plot below.

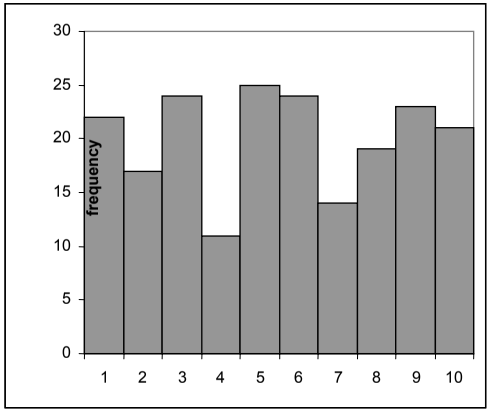


a) -0.3
b) 0
c) 0.3 (Correct)
d) 0.9
e) 1.6

Table 3-25: Annotation for SCI question 24.

G3	**Topic:**	Correlation coefficient, graphical													
	Factor Group:	Graphical													
	Notes:	Added Summer 2004. This question requires students to estimate the correlation coefficient from looking at a scatter plot of data. Students have to distinguish between negative and positive correlation, and the degree to which the data is correlated. Response (e) was added after focus group discussions. People indicated that they had tried to draw a line through the data and find y-intercept.													
	Semester (Item # on exam)**:**	**Fall 2002**		**Summer 2003**		**Fall 2003**		**Spring 2004**		**Fall 2004** (#24)		**Spring 2005** (#24)		**Summer 2005** (#24)	
	Discrimination (post):									0.12		0.30		0.30	
	Correlation with total (post):									-0.02		0.13		0.11	
	Overall Alpha Rank (post):									36th		27th		29th	
	General/Specific Factor Loading:														
		Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
	# of Responses:									123	108	105	212	73	60
	Distribution of answers (%): **a)**									2	0	7	4	0	2
	b)									16	19	19	25	23	27
	c)									**42**	**57**	**39**	**48**	**49**	**53**
	d)									21	15	16	10	12	5
	e)									15	9	19	13	12	12

25. Consider the sample distribution below. This sample was *most likely* taken from what kind of population distribution?



a) Normal
b) Uniform (Correct)
c) Skewed
d) Bimodal

Table 3-26: Annotation for SCI question 25.

G4	**Topic:**	Sample/parent distributions, uniform distributions													
	Factor Group:	Graphical													
	Notes:	This question asks students to identify the likely parent distribution from which this sample was taken. Students must have some familiarity with the shape of the distributions presented in the response set. The most common incorrect answer is d) bimodal. Focus groups have indicated that response a) is often chosen simply because they are most familiar with the normal distribution. This question has had consistently good discrimination and alpha-if-item-deleted values.													
	Semester (Item # on exam):	**Fall 2002** (#27)		**Summer 2003** (#21)		**Fall 2003** (#28)		**Spring 2004** (#25)		**Fall 2004** (#25)		**Spring 2005** (#25)		**Summer 2005** (#25)	
	Discrimination (post):	0.45		0.55		0.45		0.47		0.53		0.39		0.48	
	Correlation with total (post):	0.31		0.40		0.29		0.31		0.16		0.21		0.36	
	Overall Alpha Rank (post):	1st		4th		14th		11th		23rd		19th		8th	
	General/Specific Factor Loading:					0.40/0.46									
		Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
	# of Responses:		130	95	76	355	280	143	94	123	108	105	212	73	60
	Distribution of answers (%): **a)**		32	26	29	35	23	25	10	25	6	34	14	36	10
	b)		**31**	**32**	**26**	**31**	**36**	**26**	**37**	**33**	**46**	**25**	**34**	**29**	**43**
	c)		*	23	18	18	27	29	22	24	22	30	17	15	13
	d)		20	16	20**	14	14	15	29	15	25	11	34	21	33

* Alternate responses exponential (12%), lognormal (7%).
** Additional response (pre%, post%): exponential (2%, 7%).

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 (Correct)
d) The numbers are evenly spaced on both sides of the mean

Table 3-27: Annotation for SCI question 26.

D9	**Topic:**	Standard deviation													
	Factor Group:	Descriptive													
	Notes:	Added summer 03. Response (d) was changed during the Summer 2004 because variations of this were a common answer when this question was asked in a free response format. This change noticeably changed the answer distribution. This would suggest that a number of students feel that the standard deviation can give information about the symmetry of a distribution.													
	Semester (Item # on exam):	**Fall 2002**		**Summer 2003** (#34)		**Fall 2003** (#9)		**Spring 2004** (#26)		**Fall 2004** (#26)		**Spring 2005** (#26)		**Summer 2005** (#26)	
	Discrimination (post):			0.27		0.46		0.34		0.53		0.51		0.10	
	Correlation with total (post):			0.19		0.41		0.26		0.18		0.35		0.04	
	Overall Alpha Rank (post):			22nd		8th		15th		20th		5th		32nd	
	General/Specific Factor Loading:					0.63/0.48									
		Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
	# of Responses:			95	76	355	280	143	93	123	108	105	212	73	60
	Distribution of answers (%): **a)**			16	13	11	9	13	20	2	3	13	5	11	3
	b)			8	4	5	4	12	7	5	5	3	5	5	3
	c)			**73**	**78**	**75**	**80**	**63**	**69**	**59**	**63**	**36**	**71**	**45**	**63**
	d)			*	*	**	**	***	***	33	30	48	19	38	30

*Alternate responses (pre%, post%): A computational error was made (2%, 3%); The mean, median, and mode of these numbers are different (1%, 1%).
**Alternate responses (pre%, post%): A computational error was made (1%, 3%); The mean, median, and mode of these numbers are different (7%, 4%).
***Alternate response (pre%, post%): The mean, median, and mode of these numbers are different (8%, 2%).

27. In order to determine the mean height of American college students, which sampling method would *not* introduce bias?

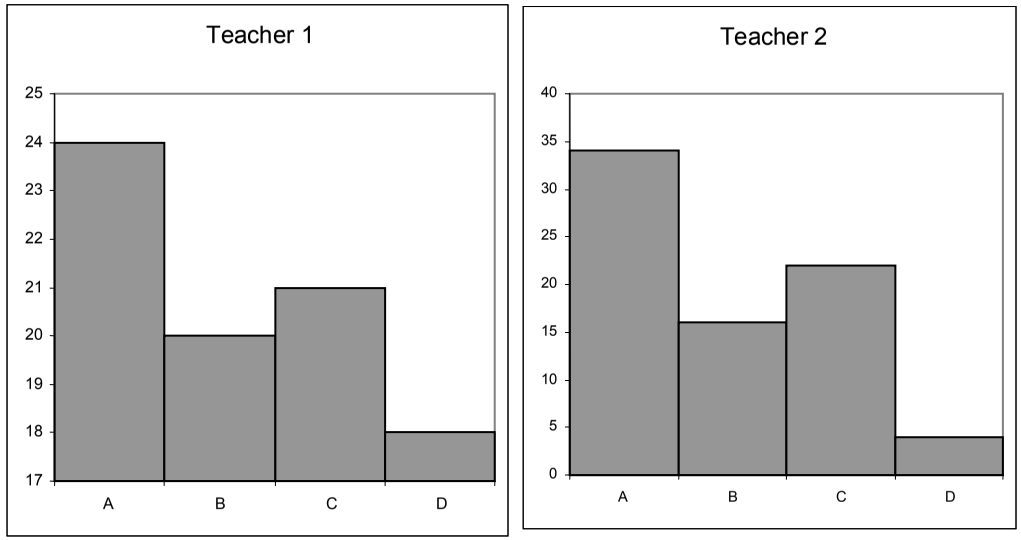
a) You randomly select from the university basketball team
b) You use a random number table to select students based on their student ID (Correct)
c) You roll a pair of dice to select from among your friends
d) None of the methods will have bias

Table 3-28: Annotation for SCI question 27.

I8	**Topic:**	Sampling, bias													
	Factor Group:	Inferential													
	Notes:	This question has undergone revision many times. We added a random element to each answer to make them more plausible. Most students answer this question correctly. This contributes to the low discrimination values.													
	Semester (Item # on exam):	**Fall 2002** (#6)		**Summer 2003** (#6)		**Fall 2003** (#11)		**Spring 2004** (#27)		**Fall 2004** (#27)		**Spring 2005** (#27)		**Summer 2005** (#27)	
	Discrimination (post):	0.24		0.22		0.19		0.38		0.29		0.24		-0.01	
	Correlation with total (post):	0.14		0.20		0.27		0.45		0.34		0.32		-0.02	
	Overall Alpha Rank (post):	22nd		23rd		20th		5th		12th		10th		31st	
	General/Specific Factor Loading:					0.50/0.10 (in Descriptive, have moved to Inferential)									
		Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
	# of Responses:					355	280	143	93	123	107	105	212	73	60
	Distribution of answers (%): **a)**					3	4	4	3	2	6	6	3	7	7
	b)					**90**	**92**	**77**	**81**	**89**	**90**	**86**	**89**	**79**	**85**
	c)					1	1	6	3	6	4	5	4	4	0
	d)					5*	3*	8	9**	2	1	4	3	10	7

*Additional response (pre%, post%): You flip a coin to select from a list of international students(1%, 1%).
**Additional response (pre%, post%): You flip a coin to select from a list of international students(2%, 3%).

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 (Correct)
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

Table 3-29: Annotation for SCI question 28.

G5	**Topic:**	Histograms													
	Factor Group:	Graphical													
	Notes:	Added Summer 04. Students must correctly read and interpret histograms. This question was added partially as a companion to question 30. It was not clear if reading histograms was the issue or understanding variability was the issue. The most common incorrect answer is d) that the distribution of grades is approximately the same. This could be because students are careless and do not look at the scales of the both graphs, or because they do believe the distributions are the same because they look the same regardless of the scale. Interview follow-ups would help to distinguish between the two. The discrimination for this question is low; this may be due to carelessness.													
	Semester (Item # on exam)**:**	**Fall 2002**		**Summer 2003**		**Fall 2003**		**Spring 2004**		**Fall 2004** (#28)		**Spring 2005** (#28)		**Summer 2005** (#28)	
	Discrimination (post):									0.21		0.22		-0.09	
	Correlation with total (post):									0.09		0.13		-0.18	
	Overall Alpha Rank (post):									28th		28th		38th	
	General/Specific Factor Loading:														
		Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
	# of Responses:									123	108	105	212	73	60
	Distribution of answers (%): **a)**									2	2	2	4	1	2
	b)									**68**	**70**	**67**	**67**	**60**	**62**
	c)									1	2	5	3	3	5
	d)									33	26	27	26	36	32

29. A scientist takes a set of 50 measurements. The standard deviation is reported as -2.30. Which of the following must be true?

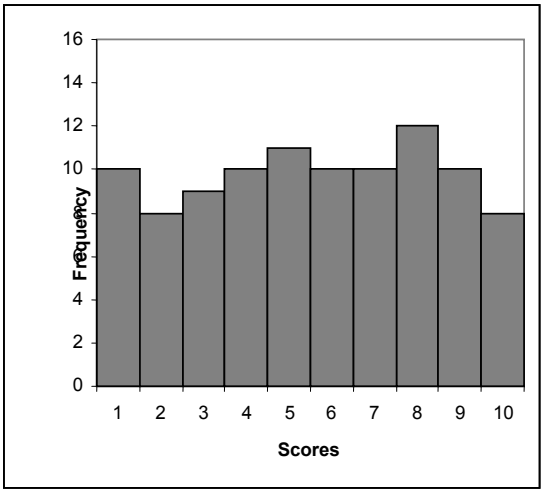
a) Most of the measurements were negative
b) All of the measurements were less than the mean
c) All of the measurements were negative
d) The standard deviation was calculated incorrectly (Correct)

Table 3-30: Annotation for SCI question 29.

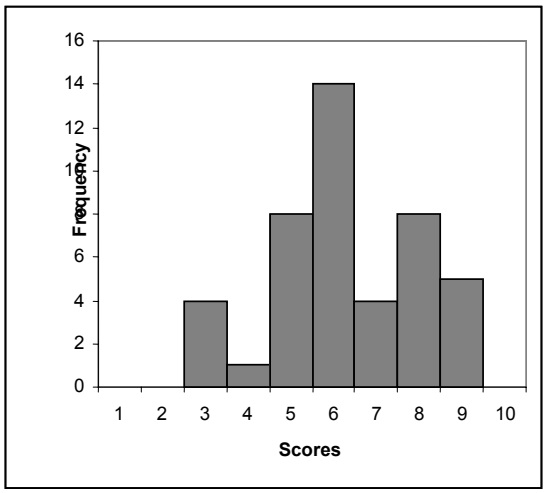
D10 **Topic:**	Standard Deviation													
Factor Group:	Descriptive													
Notes:	Students are asked to interpret a reported standard deviation. Between 30% and 40% of students do not recognize that the standard deviation cannot be a negative number. In addition, the popularity of response b) is somewhat alarming and displays a fundamental lack of understanding of the mean. This question is very discriminating and has had consistently high alpha-if-item-deleted rankings.													
Semester (Item # on exam):	**Fall 2002**		**Summer 2003** (#31)		**Fall 2003** (#18)		**Spring 2004** (#29)		**Fall 2004** (#29)		**Spring 2005** (#29)		**Summer 2005** (#29)	
Discrimination (post):			0.58		0.56		0.48		0.22		0.54		0.36	
Correlation with total (post):			0.41		0.44		0.48		0.26		0.34		0.18	
Overall Alpha Rank (post):			5th		4th		1st		16th		3rd		20th	
General/Specific Factor Loading:					0.63/0.62									
	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
# of Responses:			95	76	355	280	143	94	123	108	105	212	73	60
Distribution of answers (%): **a)**			18	16	13	13	13	13	7	8	26	7	14	22
b)			13	9	13	11	20	11	20	6	38	23	26	15
c)			11	3	5	4	8	3	2	3	4	8	5	3
d)			**56***	**66***	**68**	**72**	**55**	**72**	**69**	**83**	**33**	**62**	**55**	**60**

*Additional response (pre%, post%): Some of the measurements were zero (3%, 7%).

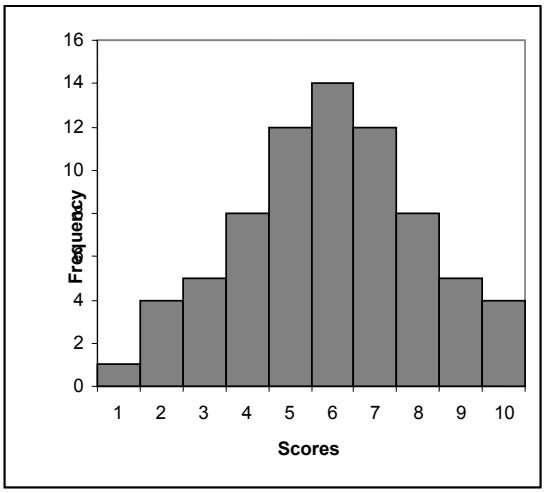
30. The following are histograms of quiz scores for four different classes. Which distribution shows the most variability?



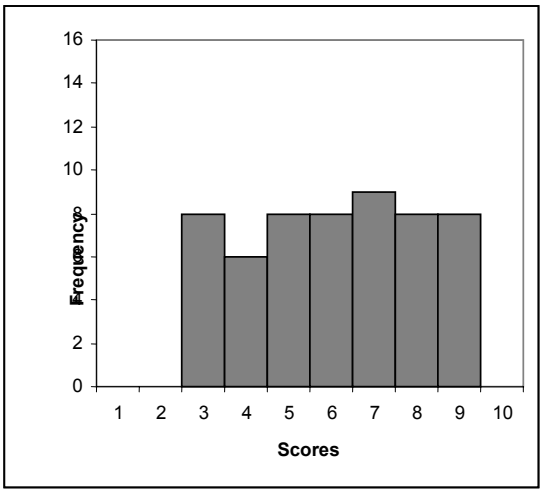
I



II



III



IV

a) I (Correct)
b) II
c) III
d) IV

Table 3-31: Annotation for SCI question 30.

G6	**Topic:**	Variability, histograms													
	Factor Group:	Graphical													
	Notes:	Response (b) is always the most common answer. Focus groups and constructed responses indicate that it is often chosen because it is the "bumpiest". Response (c) is also popular because of the "normal shape" and "they are familiar with it". We have found that many students do not interpret the histogram correctly, but read it instead as actual scores (as if it were a bar chart (non-frequency) or scatter plot) instead of frequency counts. Interestingly, in focus groups, many students who did well on the track question are unable to understand this question or draw parallels between the two. The Fall 2004 version was the first version to have the bars "touching". Earlier versions had a small space in between each bar.													
	Semester (Item # on exam):	**Fall 2002** (#30)		**Summer 2003** (#32)		**Fall 2003** (#21)		**Spring 2004** (#30)		**Fall 2004** (#30)		**Spring 2005** (#30)		**Summer 2005** (#30)	
	Discrimination (post):	0.14		0.40		0.30		0.22		0.31		0.33		0.27	
	Correlation with total (post):	0.06		0.19		0.27		0.15		0.14		0.23		0.32	
	Overall Alpha Rank (post):	31st		25th		16th		24th		25th		15th		12th	
	General/Specific Factor Loading:					0.42/0.19									
		Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
	# of Responses:		130	95	76	355	280	143	94	123	108	105	212	73	60
	Distribution of answers (%): a)		*	**	**	**15**	**19**	**13**	**18**	**26**	**31**	**10**	**23**	**18**	**17**
	b)		48	61	47	65	63	60	59	38	49	52	55	56	63
	c)		45	29	43	17	17	22	21	32	19	38	19	22	20
	d)		7	7	5	2***	1***	1***	1***	2	1	0	2	4	0

*The graph I was not included in this version. Graph IV was shifted to the left.

**The graph I was not included in this version. Graph IV was shifted to the left. Alternate responses (pre%, post%): The variability is equal for all three (7%, 1%); Insufficient information (1%, 1%).

***Graph IV was shifted to the left.

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 (Correct)
    c) Equal to 1 in 1000
    d) Insufficient information

Table 3-32: Annotation for SCI question 31.

P7 **Topic:**	Conditional probability, independence													
Factor Group:	Probability													
Notes:	The stem was changed during summer 2004. The phrase "errors often occur in groups, that is, they are not independent" was added to replace "errors often occur in bursts". This change seems to have made the question easier but also to have increased the discrimination of the question.													
Semester (Item # on exam):	**Fall 2002**		**Summer 2003** (#14)		**Fall 2003** (#13)		**Spring 2004** (#31)		**Fall 2004** (#31)		**Spring 2005** (#31)		**Summer 2005** (#31)	
Discrimination (post):			0.33		0.35		0.29		0.45		0.53		0.55	
Correlation with total (post):			0.40		0.22		0.20		0.26		0.30		0.47	
Overall Alpha Rank (post):			10th		19th		21st		13th		7th		3rd	
General/Specific Factor Loading:					0.31/0.41									
	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
# of Responses:			95	76	355	280	143	94	123	108	105	212	73	60
Distribution of answers (%): **a)**			13	9	8	10	13	13	8	10	11	13	10	10
b)			**36**	**33**	**36**	**30**	**38**	**21**	**46**	**57**	**47**	**54**	**49**	**43**
c)			35	42	34	39	30	43	24	19	26	20	22	33
d)			17	16	21	20	17	22	20	14	16	13	19	13

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)

Table 3-33: Annotation for SCI question 32.

I9	**Topic:**	Hypothesis testing, p-value													
	Factor Group:	Inferential													
	Notes:	Added spring 04, hoping to add a p-value question that was more "applied" and less definition oriented, and hopefully a little easier. In Spring 04, there were gains in 2 of three classes from pre to post, with 30%-50% answering correctly.													
	Semester (Item # on exam):	**Fall 2002**		**Summer 2003**		**Fall 2003**		**Spring 2004** (#32)		**Fall 2004** (#32)		**Spring 2005** (#32)		**Summer 2005** (#32)	
	Discrimination (post):							0.51		0.18		0.35		-0.04	
	Correlation with total (post):							0.28		0.05		0.23		0.01	
	Overall Alpha Rank (post):							8th		13th		13th		33rd	
	General/Specific Factor Loading:														
		Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
	# of Responses:							143	94	123	108	105	212	73	60
	Distribution of answers (%): **a)**							20	10	16	5	18	11	26	13
	b)							11	6	7	7	6	7	15	7
	c)							34	33	38	45	35	43	38	47
	d)							**20***	**42***	**23****	**42**	**41**	**38**	**18**	**33**

*Additional response (pre%, post%): Accept the alternate hypothesis (7%, 9%).
** Additional response: Accept the alternate hypothesis (11%).

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 (Correct)
b) 0.95
c) 0.997
d) 1.00

Table 3-34: Annotation for SCI question 33.

P8	**Topic:**	Properties of the normal distribution, standard deviation													
	Factor Group:	Probability													
	Notes:	This question requires students to recognize that the data is likely to be normally distributed and to know and apply the 68-95-99 rule for a normal distribution. Response (d) was added based on fill in responses for fall 02 administration. % correct has been very variable by class on the post test ranging from 20%-60%. The discrimination and alpha-if-item-deleted values are very high for this question.													
	Semester (Item # on exam):	**Fall 2002** (#18)		**Summer 2003** (#20)		**Fall 2003** (#23)		**Spring 2004** (#33)		**Fall 2004** (#33)		**Spring 2005** (#33)		**Summer 2005** (#33)	
	Discrimination (post):	0.25		0.50		0.64		0.58		0.69		0.61		0.54	
	Correlation with total (post):	0.16		0.43		0.46		0.37		0.45		0.44		0.44	
	Overall Alpha Rank (post):	8th		3rd		1st		2nd		1st		1st		5th	
	General/Specific Factor Loading:					0.65/-0.01									
		Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
	# of Responses:		133	95	76	355	280	143	94	123	108	105	212	73	60
	Distribution of answers (%): **a)**		**24**	**22**	**21**	**25**	**33**	**27**	**38**	**34**	**40**	**22**	**38**	**19**	**30**
	b)		29	33	29	33	31	27	29	26	30	35	26	32	25
	c)		29	18	17	18	14	20	10	15	23	23	14	12	20
	d)		*	20**	25**	22	21	21	21	21	17	19	22	36	25

* Additional responses: 0.50 (13%) , Other_____ (6%). **Additional response (pre%, post%): 0.50 (8%, 8%).

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 (Correct)
d) There is no way this can happen

Table 3-35: Annotation for SCI question 34.

P9	**Topic:**	Law of large numbers													
	Factor Group:	Probability													
	Notes:	This question demonstrates the misconception described as a belief in the law of small numbers where people expect that a process will be represented both globally and locally within a sample/event (Kahneman, et al. 1982). This question typically has good discrimination and alpha-if-item-deleted values.													
	Semester (Item # on exam)**:**	**Fall 2002** (#3)		**Summer 2003** (#3)		**Fall 2003** (#20)		**Spring 2004** (#34)		**Fall 2004** (#34)		**Spring 2005** (#34)		**Summer 2005** (#34)	
	Discrimination (post):	0.28		0.25		0.50		0.41		0.31		0.36		0.68	
	Correlation with total (post):	0.07		0.18		0.35		0.34		0.26		0.27		0.48	
	Overall Alpha Rank (post):	21st		30th		10th		6th		15th		11th		2nd	
	General/Specific Factor Loading:					0.55/0.29									
		Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
	# of Responses:		130	95	76	355	280	143	94	123	108	105	212	73	60
	Distribution of answers (%): **a)**		12	19	13	13	12	14	12	11	10	17	17	18	15
	b)		1	4	3	6	3	8	5	5	2	8	2	3	13
	c)		**45**	**64**	**68**	**65**	**68**	**64**	**62**	**63**	**74**	**58**	**71**	**64**	**60**
	d)		10*	12	16	15	15	10	20	16	14	17	10	15	12

*Additional response: This is possible for any of the trials (31%).

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 (Correct)
    b) Wider
    c) The same width
    d) It depends on the confidence level

Table 3-36: Annotation for SCI question 35.

I10	**Topic:**	Confidence intervals													
	Factor Group:	Inferential													
	Notes:	This question examines the effect of sample size on the width of confidence intervals. This problem was reworded for Fall 2004. The response set was changed from "Smaller", "Larger", "No change", and "It depends on the confidence level". The stem was also rewritten slightly to hopefully make the question clearer. This change greatly increased the discrimination of this question.													
	Semester (Item # on exam):	**Fall 2002** (#11)		**Summer 2003** (#11)		**Fall 2003** (#32)		**Spring 2004** (#35)		**Fall 2004** (#35)		**Spring 2005** (#35)		**Summer 2005** (#35)	
	Discrimination (post):	0.32		0.41		0.31		0.33		0.55		0.51		0.68	
	Correlation with total (post):	0.17		0.22		0.15		0.19		0.33		0.34		0.49	
	Overall Alpha Rank (post):	11th		17th		26th		22nd		5th		4th		1st	
	General/Specific Factor Loading:					0.25/0.15									
		Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
	# of Responses:		137	95	76	355	280	143	94	123	108	105	212	73	60
	Distribution of answers (%): **a)**		**42**	**38**	**59**	**30**	**42**	**28**	**35**	**24**	**44**	**30**	**37**	**21**	**53**
	b)		42	51	16	37	30	38	29	28	25	32	32	43	30
	c)		7	5	11	15	19	13	18	34	18	26	19	27	15
	d)		9	6	14	14	9	16	17	9	14	13	10	10	13

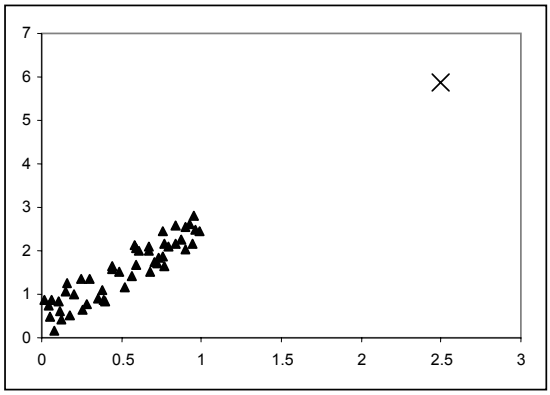
36. A chemical company has decided to begin producing a new product. They want to use existing equipment. An engineer is assigned to determine which of two reactor settings will yield the most pure product. He performs ten runs at each of the settings and measures the purity. Which test is most appropriate for this analysis?

    a) two-sample Z test
    b) paired comparison t test
    c) two-sample t test (Correct)
    d) one-sample t test

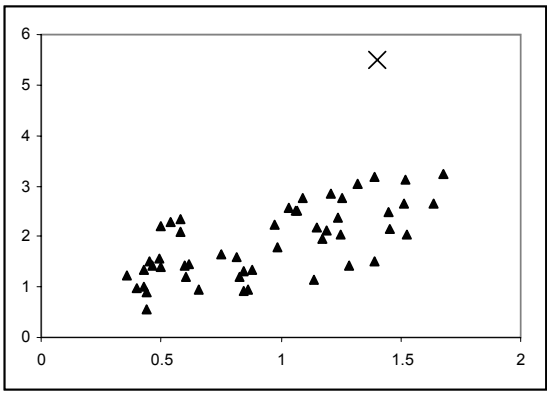
Table 3-37: Annotation for SCI question 36.

I11/I11a **Topic:**	Hypothesis testing, two sample t-test													
Factor Group:	Inference													
Notes:	Added during summer 04. This question requires students to choose the most appropriate statistical test. They must consider the sample size and whether the observations are independent. The discrimination for this question is very low. From the distribution of the answers, there appears to be considerable guessing. There is evidence that many students are able to rule out the z-test based on sample size, but are unable to distinguish which t-test would be appropriate. For spring and summer 2005, the stem was changed to read: "will yield the highest quality product".													
Semester (Item # on exam):	**Fall 2002**		**Summer 2003**		**Fall 2003**		**Spring 2004**		**Fall 2004** (#36)		**Spring 2005** (#36)		**Summer 2005** (#36)	
Discrimination (post):									0.12		0.11		0.09	
Correlation with total (post):									0.05		0.04		0.01	
Overall Alpha Rank (post):									30th		37th		34th	
General/Specific Factor Loading:														
	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
# of Responses:									123	108	105	212	73	60
Distribution of answers (%): a)									24	18	19	17	26	12
b)									28	28	34	35	29	27
c)									**34**	**45**	**38**	**40**	**33**	**40**
d)									9	9	9	7	7	22

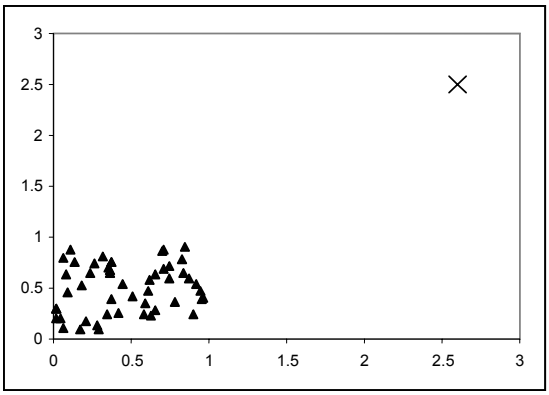
37. Consider the correlation coefficients of the scatter plots below. If the data point that is marked by an × is *removed*, which of the following statements would be true?



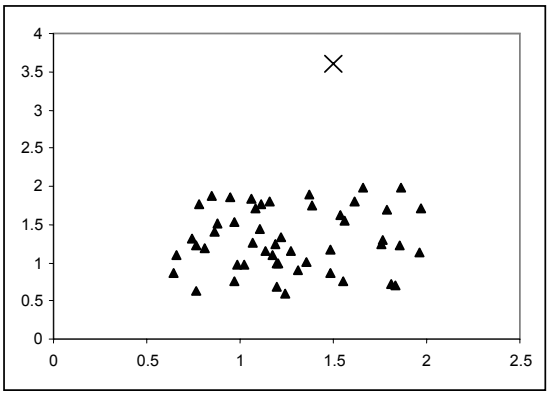
I



II



III



IV

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 (Correct)
d) correlation of (II) increases, correlation of (III) increases

Table 3-38: Annotation for SCI question 37.

G7 **Topic:**	Correlation, effect of outliers, scatter plots													
Factor Group:	Graphical													
Notes:	This question was added summer 04 and requires students to estimate how the correlation coefficient would change if a particular data point was removed. There is evidence of guessing for this question, and the discrimination values are not very high. These facts along with the way the responses are constructed make it difficult to pinpoint specific misconceptions that students may hold.													
Semester (Item # on exam):	**Fall 2002**		**Summer 2003**		**Fall 2003**		**Spring 2004**		**Fall 2004** (#37)		**Spring 2005** (#37)		**Summer 2005** (#37)	
Discrimination (post):									0.16		0.22		0.29	
Correlation with total (post):									0.07		0.11		0.17	
Overall Alpha Rank (post):									29th		31st		21st	
General/Specific Factor Loading:														
	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
# of Responses:									123	108	105	212	73	60
Distribution of answers (%): **a)**									15	7	15	10	19	10
b)									16	10	17	23	21	20
c)									**39**	**53**	**38**	**43**	**42**	**40**
d)									25	29	31	23	18	25

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 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 (Correct)
c) Engine size is a poor predictor of engine revolutions per mile
d) Engine size is independent of engine revolutions per mile

Table 3-39: Annotation for SCI question 38.

I12	**Topic:**	Correlation													
	Factor Group:	Inferential													
	Notes:	Added Spring 2005 post to have a correlation question that was in a non-graphical format.													
	Semester (Item # on exam):	**Fall 2002**		**Summer 2003**		**Fall 2003**		**Spring 2004**		**Fall 2004**		**Spring 2005** (#38)		**Summer 2005** (#38)	
	Discrimination (post):											0.17		0.61	
	Correlation with total (post):													0.33	
	Overall Alpha Rank (post):													9th	
	General/Specific Factor Loading:														
		Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
	# of Responses:												106	73	60
	Distribution of answers (%): **a)**												11	15	12
	b)												**63**	**42**	**47**
	c)												23	27	23
	d)												3	15	13

Chapter 4 An Item Response Theory Perspective

As discussed in Chapter 1, item response theory provides an additional set of tools that we can utilize in constructing and analyzing the SCI. Several advantages over classical test theory methods can make it a valuable addition. It provides sample independent estimates about the individual items. This enables us to be able to make decisions that are not overly influenced by an individual semester's idiosyncrasies. It also gives us much more insight into the behavior of the question over the range of the ability distribution, including the measurement error.

The ability or theta distribution is assumed to have a mean of zero and a standard deviation of 1. The pattern of item responses is modeled by a mathematical function that relates the latent trait/ability to the probability of answering a question correctly. The models used here are the two parameter logistic model and the nominal response model.

4.1 The Data Set

In order to achieve a large enough sample size to carry out an IRT analysis, the questions on the fall 2004 version were divided into groups by topic area and assigned a master number so they could be tracked backward through the previous versions of the instrument. The questions on each previous version of the SCI were compared to the fall 2004 version. Then a new data set was created for each semester that included the item responses for those questions that were the same as the fall 2004 version. The questions

that were different or that were no longer on the fall 2004 version were marked as not presented. Finally these data sets were combined into a single master data set. The same method was followed for subsequent semesters so that a master data set has been created with all data from fall 2002 to summer 2005 with each question having a unique identifier.

A few questions had undergone minor revisions for fall 2004 and had been unchanged for several semesters prior to fall 2004. These questions were included in the data set but were divided into their two versions, for example P2 (earlier version) and P2a (newer version). The data included in the master set are shown in Table 4-1. By including both versions of these questions, we can evaluate the changes that were made and decide whether the changes were an improvement or not. This method is essentially a horizontal equating scheme with common items and non-equivalent groups (Kolan and Brennan 1995). The common items serve as "anchor items" and item parameters are estimated simultaneously. The two forms of the questions can then be compared.

Table 4-1: Historical matrix for data selection. Data from the shaded areas were included in the analysis.

			Fall 2002	*Summer 2003*	*Fall 2003*	*Spring 2004*	*Summer 2004*	*Fall 2004*	*Spring 2005*	*Summer 2005*
	Item	*Total # of Responses*	171	103	281	94	16	211	213	60
Topic Area	P1/P1a	374/109						P1		P1a
	P2/P2a	773/203			P2			P2a	P2	
Probability	P3	483							P3	
	P4/P4a	1037/166				P4				P4a
	P5	483							P5	
	P6	499							P6	
	P7/P7a	493/483			P7				P7a	
	P8	976			P8					
	P9	976			P9					
Descriptive	D1	272							D1	
	D2	483							D2	
	D3	1146					D3			
	D4	1146					D4			
	D5	593						D5		
	D6	976					D6			
	D7	976					D7			
	D8/D8a	390/483				D8			D8a	
	D9	483							D9	
	D10	873					D10			
Inferential	I1	483							I1	
	I2	593						I2		
	I3	873					I3			
	I4/I4a	878/268			I4			I4a	I4	I4a
	I5	499							I5	
	I6	873					I6			
	I7	976					I7			
	I8	483							I8	
	I9	483							I9	
	I10/I10a	663/483			I10				I10a	
	I11/I11a	374/109						I11		I11a
	I12	166								I12
Graphical	G1	499							G1	
	G2	499							G2	
	G3	499							G3	
	G4	873					G4			
	G5	499							G5	
	G6	873					G6			
	G7	499							G7	

4.2 The Two Parameter Logistic Model (2PL)

Once the data set had been created, the IRT analysis was carried out using the analysis software BILOG-MG (Zimowsky, Muraki, Mislevy and Bock 2003). The data were modeled with a 2-parameter logistic model. In this model, two parameters for each item are estimated that define the item characteristic curve (ICC) for that item; a slope or discrimination parameter, a, and a threshold parameter, β. The threshold parameter is the value of theta (the ability level) for which the probability of answering the question correctly is 0.5. The discrimination parameter is the slope of the ICC at the point $\Theta=\beta$. For the estimation routine, Bayesian priors were used for both the slope and the threshold parameters. The following analysis is in the logit metric, for which the model is

$$P(X_i = 1 \mid \Theta) = \frac{\exp(a_i(\Theta - \beta))}{1 + \exp(a_i(\Theta - \beta))}.$$

Table 4-2 contains the item statistics and item parameter estimates. Recall that higher values of the discrimination parameter a are desirable, the normal range of values is from 0 to 2. The threshold parameter β is a measure of the item difficulty and it the point along the ability distribution where the probability of answering correctly is 0.5. For example, consistent with previous findings, the parameter estimates for question P4 indicate a very difficult question (β=5.752) with low discrimination (α=0.329). Similar results are found for question G2 (β=5.251, α=0.307).

Table 4-2: Item Statistics and Parameter Estimates

	Item Statistics				*Item Parameters*		
Item	**N**	**% Correct**	**Pearson Correlation**	**Biserial Correlation**	**Slope Parameter *a***	**Threshold Parameter β**	**Factor Loading**
XD1	272	31.2	0.121	0.158	0.454	1.754	0.413
XD2	483	56.1	0.039	0.049	0.303	-0.871	0.29
XD3	1146	72.3	0.208	0.279	0.755	-1.43	0.603
XD4	1146	69.8	0.276	0.363	1.073	-0.969	0.731
XD5	593	63.6	0.26	0.333	0.733	-0.884	0.591
XD6	976	71.4	0.299	0.397	1.044	-1.059	0.722
XD7	976	61.5	0.167	0.213	0.533	-0.921	0.47
XD8	390	46.2	0.314	0.394	0.872	0.249	0.657
XD8A	483	41.4	0.254	0.321	0.778	0.436	0.614
XD9	483	64.8	0.229	0.294	0.732	-0.986	0.591
XD10	873	67.6	0.341	0.444	1.152	-0.824	0.755
XG1	499	26.5	0.305	0.411	0.943	1.206	0.686
XG2	499	16	0.018	0.028	0.307	5.251	0.294
XG3	499	54.5	0.062	0.078	0.307	-0.643	0.293
XG4	873	38	0.243	0.31	0.654	0.801	0.547
XG5	499	67.3	0.064	0.083	0.353	-2.122	0.333
XG6	873	20.6	0.216	0.307	0.711	2.072	0.58
XG7	499	43.7	0.103	0.13	0.396	0.602	0.368
XI1	483	47.8	0.146	0.183	0.42	0.152	0.387
XI2	593	41	0.069	0.087	0.324	1.098	0.308
XI3	873	43.5	0.255	0.321	0.642	0.427	0.54
XI4	878	26.7	0.151	0.204	0.553	1.98	0.484
XI4A	268	32.1	-0.004	-0.005	0.332	2.133	0.315
XI05	499	30.3	0.029	0.039	0.306	2.671	0.292
XI06	873	36.2	0.291	0.373	0.888	0.727	0.664
XI07	976	41.6	0.315	0.398	0.921	0.443	0.678
XI08	483	88.8	0.21	0.348	0.884	-2.681	0.662
XI09	483	43.3	0.11	0.139	0.346	0.731	0.327
XI10	663	42.1	0.171	0.215	0.484	0.738	0.435
XI10A	483	43.1	0.307	0.387	0.834	0.316	0.64
XI11	374	39.6	-0.031	-0.039	0.256	1.54	0.248
XI11A	109	49.5	0.167	0.209	0.554	0.04	0.485
XI12	166	57.8	0.2	0.253	0.591	-0.599	0.509
XP1	374	58.6	0.231	0.292	0.704	-0.622	0.576
XP1A	109	45	0.147	0.184	0.579	0.371	0.501
XP2	773	34.3	0.346	0.447	1.018	0.821	0.714
XP2A	203	41.4	0.297	0.375	0.806	0.357	0.628
XP3	483	14.7	0.086	0.133	0.49	3.629	0.44
XP4	1037	13.2	0.046	0.072	0.329	5.752	0.312
XP4A	166	7.8	0.172	0.316	0.724	3.543	0.586
XP5	483	56.1	0.063	0.079	0.322	-0.825	0.307
XP6	499	56.9	0.134	0.169	0.466	-0.675	0.423
XP7	493	29.6	0.257	0.34	0.71	1.44	0.579
XP7A	483	48.9	0.291	0.364	0.857	-0.005	0.651
XP8	976	34.3	0.373	0.482	1.196	0.704	0.767
XP9	976	67	0.315	0.409	1.035	-0.826	0.719

4.2.1 Question Comparisons

The questions included with multiple versions are discussed next. Question P1 (#1) has undergone many revisions since it was originally piloted in fall 2002. The contingency table was added in its current format in fall 2004, but option (e) was not included until spring 2005:

P1. 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 result 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 (Correct)
d) 0.95*0.98
e) 0.02+0.05

The addition of response (e) makes the question more difficult, the percentage of people answering correctly drops from more than 60% to below 50%, with response (e) being a strong distractor. However, while the threshold parameter increases, the discrimination/slope parameter decreases. The item characteristic curves for the two versions of the question are shown in Figure 4-1. The probability of answering this question correctly at the upper end of the ability distribution is not as high for version P1a. Thus, it is not clear how to treat this question. Eliminating choice (e) makes for a stronger question psychometrically, but it eliminates a powerful distractor. From an

instructor's point of view, being able to identify the presence of this error/misconception may be of more value.

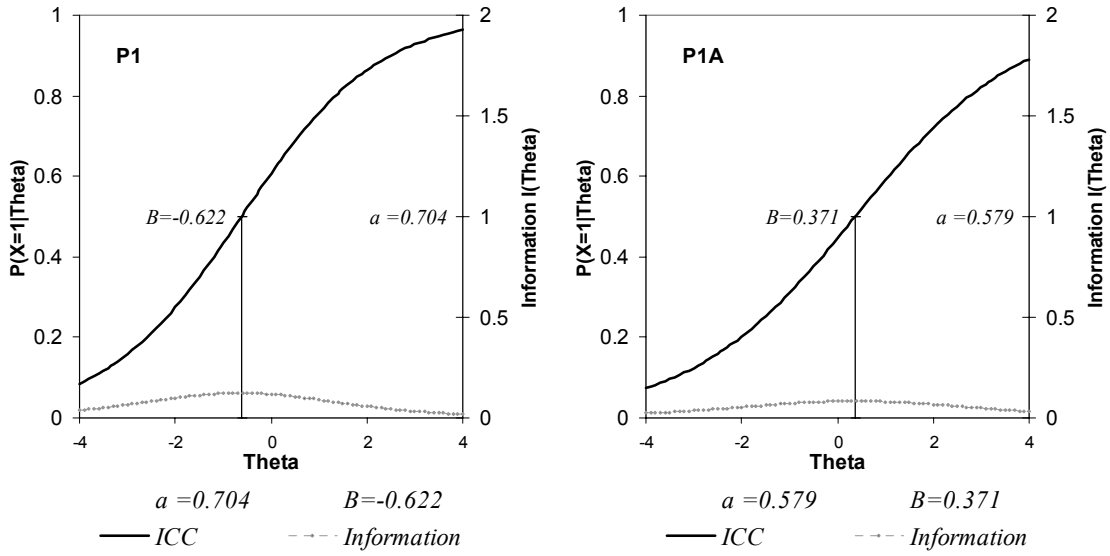


Figure 4-1: Item characteristic and information curves for items P1 and P1a.

Question P2 (#4) is shown below. This is a question about the law of large numbers. It has consistently been a good question in terms of discrimination (0.4-0.6) and ranking in the top few for alpha-if-item-deleted. For the fall 2004 version, P2a, response (d) was deleted. Generally about half of all people choose (c). Answer (d) generally attracted less than 10% of people and we felt that (d) wasn't really an appropriate way to answer the question posed, so it was deleted. The item characteristic curves for the two questions are shown in Figure 4-2.

P2. 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 (Correct)
b. Urban
c. Equally likely
d. Both are extremely unlikely

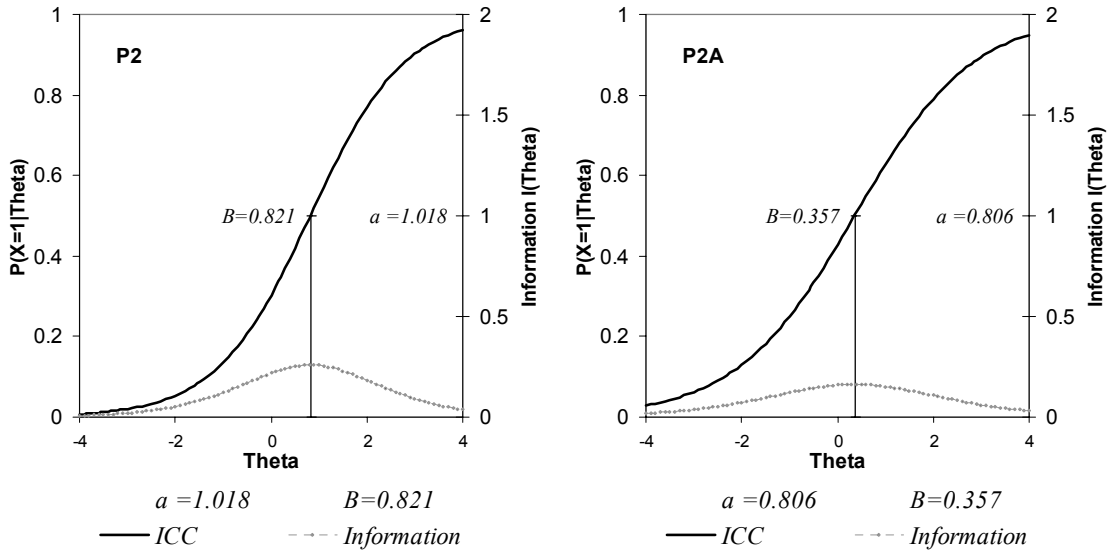


Figure 4-2: Item Characteristic Curves for items P2 and P2a.

Interestingly, this perceived minor revision changed the item characteristic and item information curves. The distribution of answers to (b) and (c) remained very similar but a much greater percentage of people chose the correct answer (a) in the P2a version. Note that the original version has a higher threshold value and is more discriminating. This larger slope parameter increases the information function, which subsequently decreases the standard error. Based on this, we decided to go back to the original version of the question. It would also be interesting to conduct student interviews to find out what types of student thinking lead to the different responses.

Question P7 (#31) is a conditional probability question. The original version included the engineering term "bursts"; to make the question more appropriate for a more general audience this question. There has been concern that the rewrite made the question too easy by giving away too much information with the independence statement. The

answer distribution changed greatly between the two versions. The typical distribution of answers is shown below to the right of the question.

P7. 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?

P7a. 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?

	P7	P7a
a) Less than 1 in 1000	10%	10%
b) Greater than 1 in 1000 (Correct)	30%	50%
c) Equal to 1 in 1000	40%	25%
d) Insufficient information	20%	15%

The test characteristic curves in Figure 4-3 show that while the newer version is easier, it is more discriminating. It contributed more information at the center of the Θ distribution. Furthermore, as the independence is explicitly stated in the question, the new wording removes any doubt about how the situation should be interpreted. Thus we decided to retain the newer version of this question.

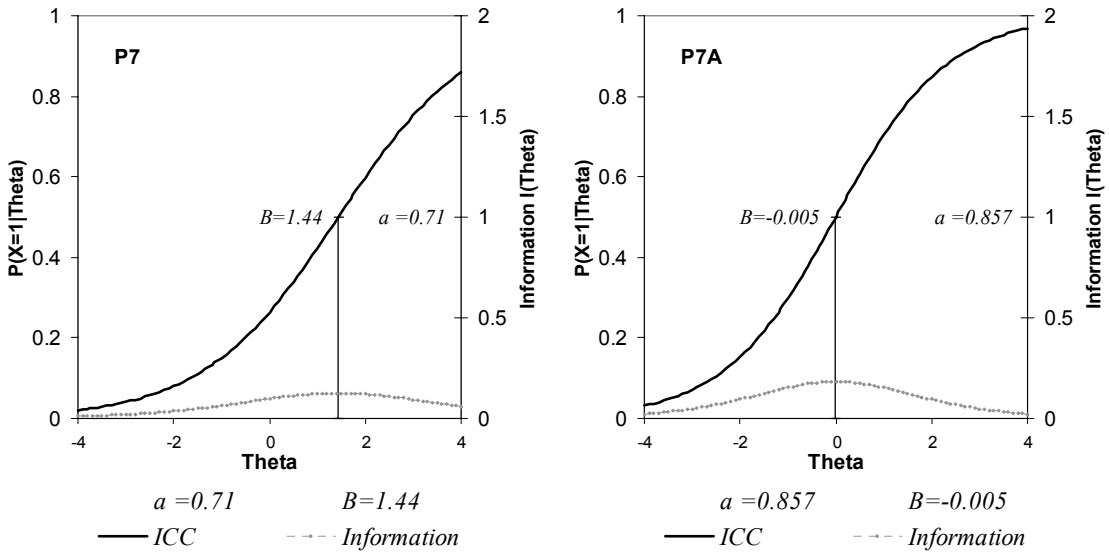


Figure 4-3: Item characteristic and information curves for questions P7 and P7a.

Question I4 (#18), shown below, is about hypothesis testing and p-values. For the fall 2004 version, response (e) was deleted. Each answer in the I4 version received roughly 15%-20% of the responses. For version I4a, answers (a) and (c) received approximately 30%, b) 20%, and d) 10%:

I4. 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 $\alpha = 0.05$ significance level
b) The power of the test statistic used was 90%
c) Assuming H_0 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
e) The probability that the null hypothesis is actually true is 0.9

The new question version had about the same level of difficulty but the discrimination was lower, though neither version had a particularly high discrimination parameter. The item characteristic and information curves are shown in Figure 4-4. The distribution of answers may indicate that the influence of guessing may be too great, contributing to the low discrimination. We retained the previous version of this question. Student interviews could be helpful for improving this question.

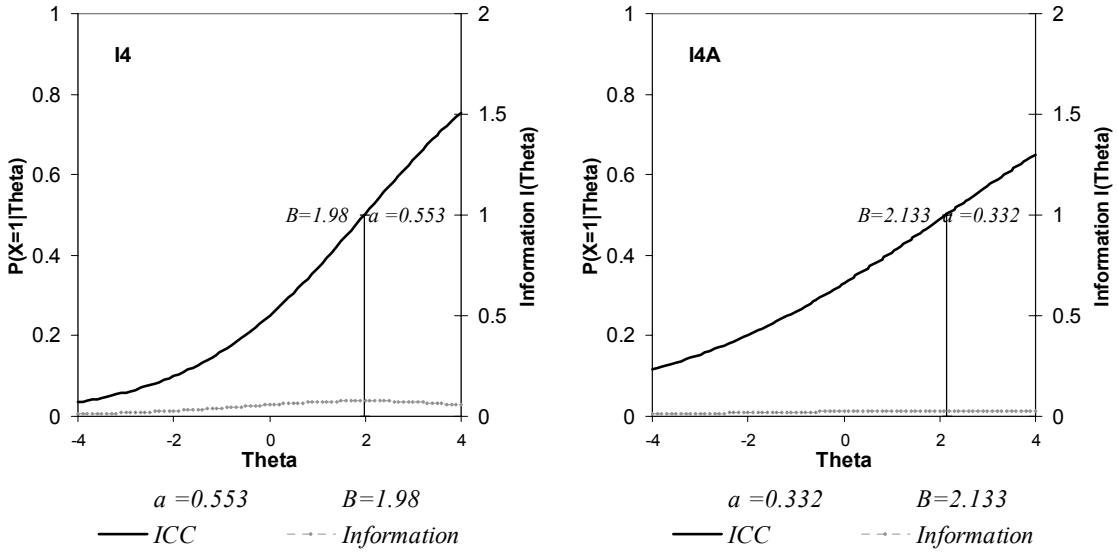


Figure 4-4: Item characteristic and information curves for items I4 and I4a.

Question I10 (#35), which is about confidence intervals, was reworded in both the stem and the response set for fall 04. The response distribution appears to be similar for both forms of the question:

I10a. When calculating a confidence interval on a given population with a fixed significance level, using a larger sample size will make the confidence interval:

a) Smaller (Correct)
b) Larger
c) No Change
d) It depends on the significance level

I10. 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 (Correct)
b) Wider
c) The same width
d) It depends on the confidence level

The discrimination index for this question had generally been around 0.3 and it has usually had a midrange alpha-if-item-deleted ranking. The new version had a considerably higher discrimination index, over 0.5 and one of the highest alpha-if-item deleted rankings. The item characteristic curves are shown in Figure 4-5. The item difficulty for the two questions was about the same, but the newer version of the question was more discriminating. The questions are virtually the same at face value, but their behavior is different. The newer version I10a, which is more precise and seems to work better, was retained.

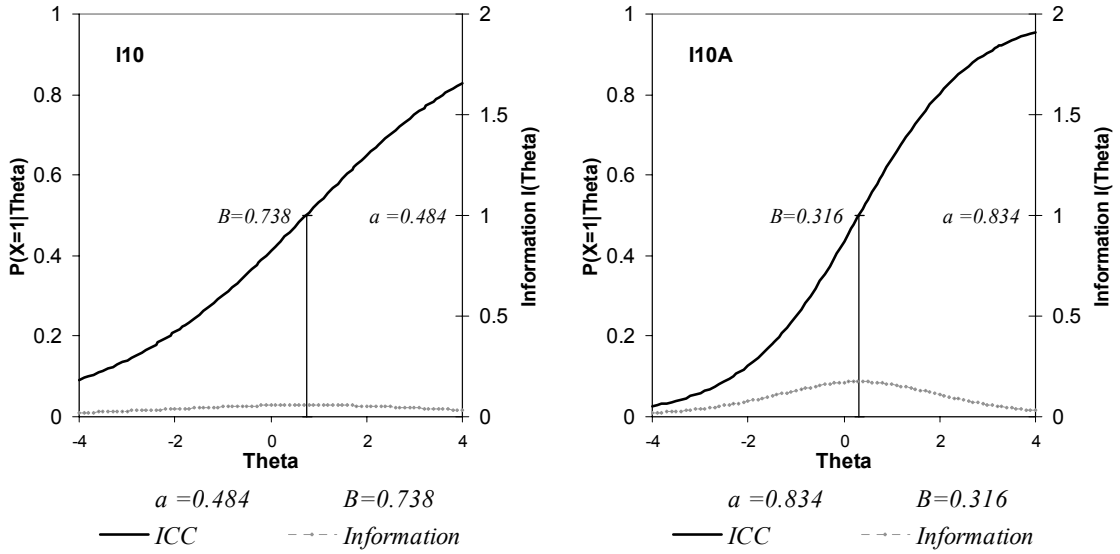


Figure 4-5: Item characteristic and information curves for items I10 and I10a .

The stem of question I11 (#36) was reworded slightly during spring 2005. The words "most pure" were replaced with "more quality" and the quality is measured in the new version. This change was made to incorporate a quality engineering viewpoint:

I11. 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 (Correct)
d) one-sample t test

The discrimination index was very low for this question in either form and the alpha-if-item-deleted rankings were equally poor. The IRT analysis provides further evidence of this. The item characteristic curves are shown in Figure 4-6. It would seem that this change should have little impact on the way this question is interpreted, though

the newer version does appear to have somewhat different properties. It should be remembered, however, that since this question was relatively new at the time of this analysis, the number of data points for the new version was small in IRT terms (only 109 responses). The newer version was retained for now, but this question is a good candidate for further revisions.

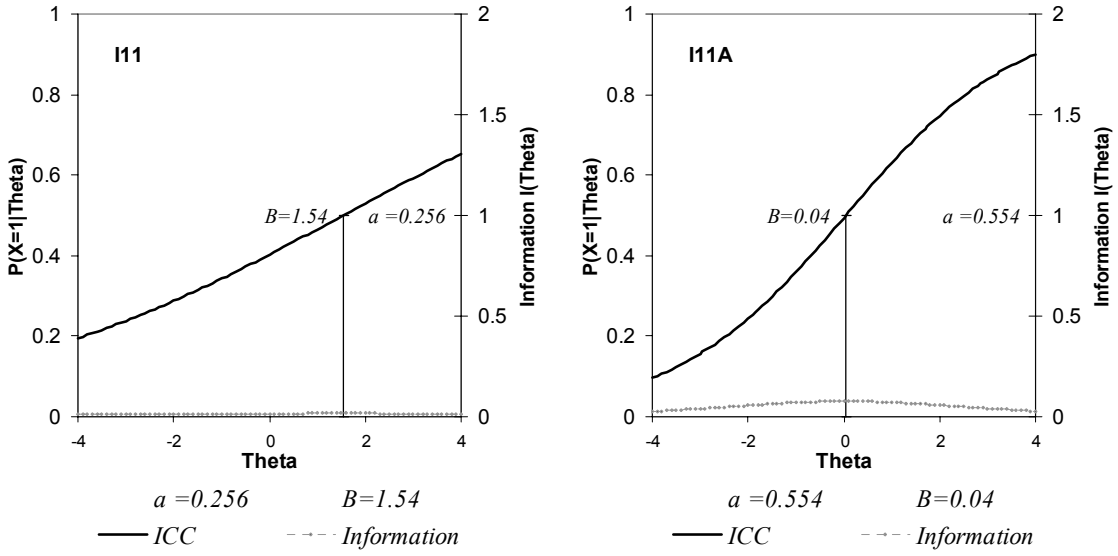


Figure 4-6: Item characteristic and information cureves for items I11 and I11a.

Question D8 (#23) is the final question that has been analyzed in this manner. The response set was changed for fall 2004 to eliminate unpopular distractors:

D8. 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 only
b) II only
c) I & II
d) I & III
e) All 3

D8a. 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 (Correct)
b) II & III
c) I & III
d) All 3

As can be seen from the ICC and information curves in Figure 4-7, this change did not have dramatic impact on this question, though the discrimination parameter was actually higher in the original version.

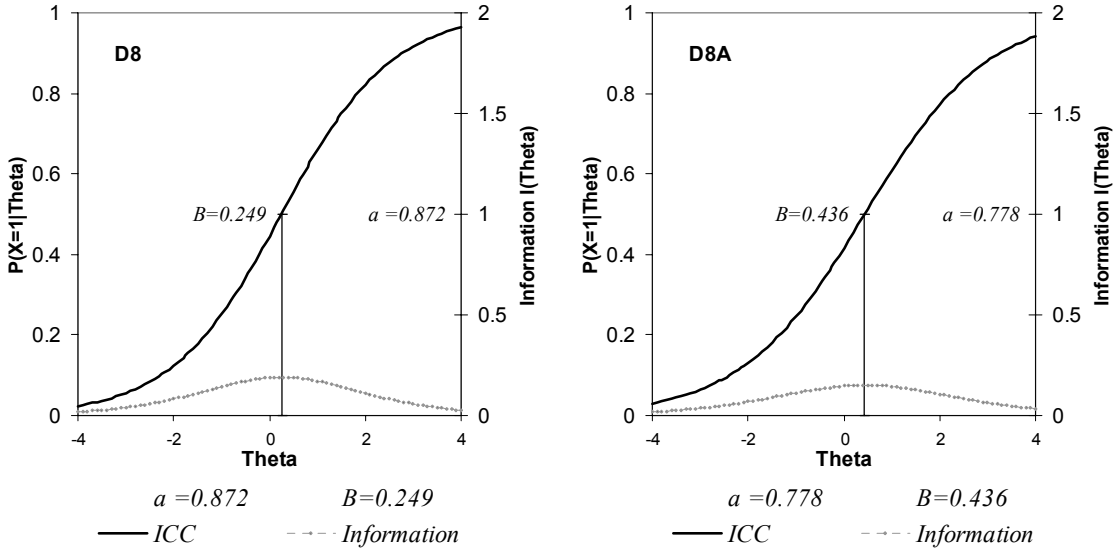


Figure 4-7: Item characteristic and information curves for items D8 and D8a.

Item analysis of this type can help to guide further refinements of the SCI. Being able to make comparisons based on information that is derived from all the data at once instead of from a single semester can lend increased confidence to the subsequent decisions. The item characteristic curves also provide a sense of the question behavior over the entire ability distribution. The response curves for all the questions are included

in Appendix A. The remaining analysis is based only the data for the versions of the 38 questions that were in use at the time of this writing.

4.2.2 The Test as a Whole

In addition to detailed item behavior, item response theory can also provide information about the test as a whole. We can get a sense of how the test covers the ability distribution by looking at the distribution of the item threshold parameters (see Figure 4-8). From this, we can see that the SCI is a somewhat difficult test. The mean threshold value is 0.346 and the median is 0.223. The majority of the questions are concentrated in the middle of the ability distribution.

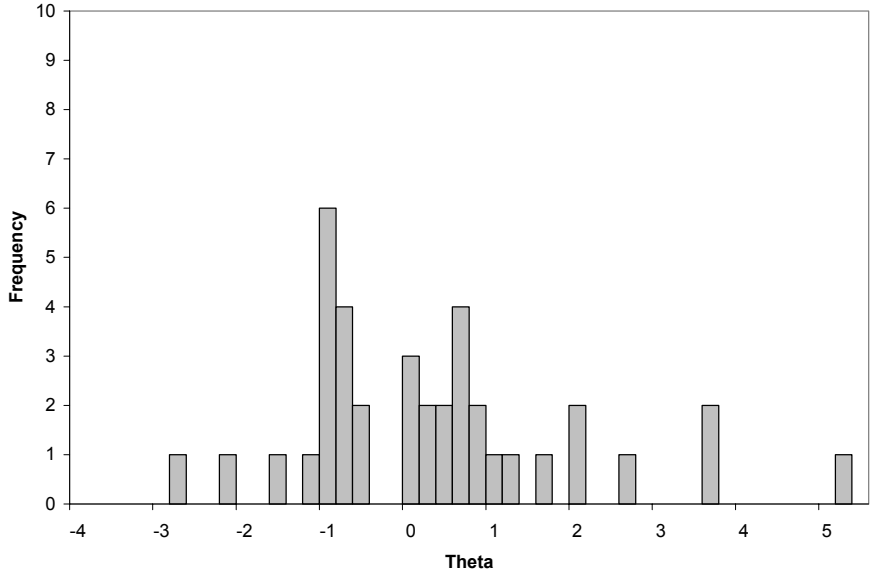


Figure 4-8: Distribution of threshold (β), parameters for SCI items.

The IRT analysis provides an estimate of the measurement error in estimating the ability level Θ. The item information function for dichotomously scored items is defined

for each item as

$$I_i(\Theta) = \frac{[P_i'(\Theta)]^2}{P_i(\Theta)(1 - P_i(\Theta))}.$$

The denominator of the information function is the item variance, so as the variance decreases we obtain better estimates of the latent trait, Θ and thus more information. The numerator is the slope of the ICC at Θ, so steeper slope values also increase the information. The total test information is the sum of all the item information functions. The information function is used to estimate the standard error of measurement for the estimation of theta:

$$SE = \frac{1}{\sqrt{I(\Theta)}}.$$

Since the maximum item information occurs at the threshold value of theta, the standard error is also lowest in this area of the ability distribution. The total test information and standard error curves over the ability distribution are shown in Figure 4-9. The standard error can be used to obtain a reliability estimate for the SCI. Since the standard error is not constant over the theta distribution, the reliability estimate is an average over the theta distribution. The reliability estimate obtained for the SCI is 0.787.

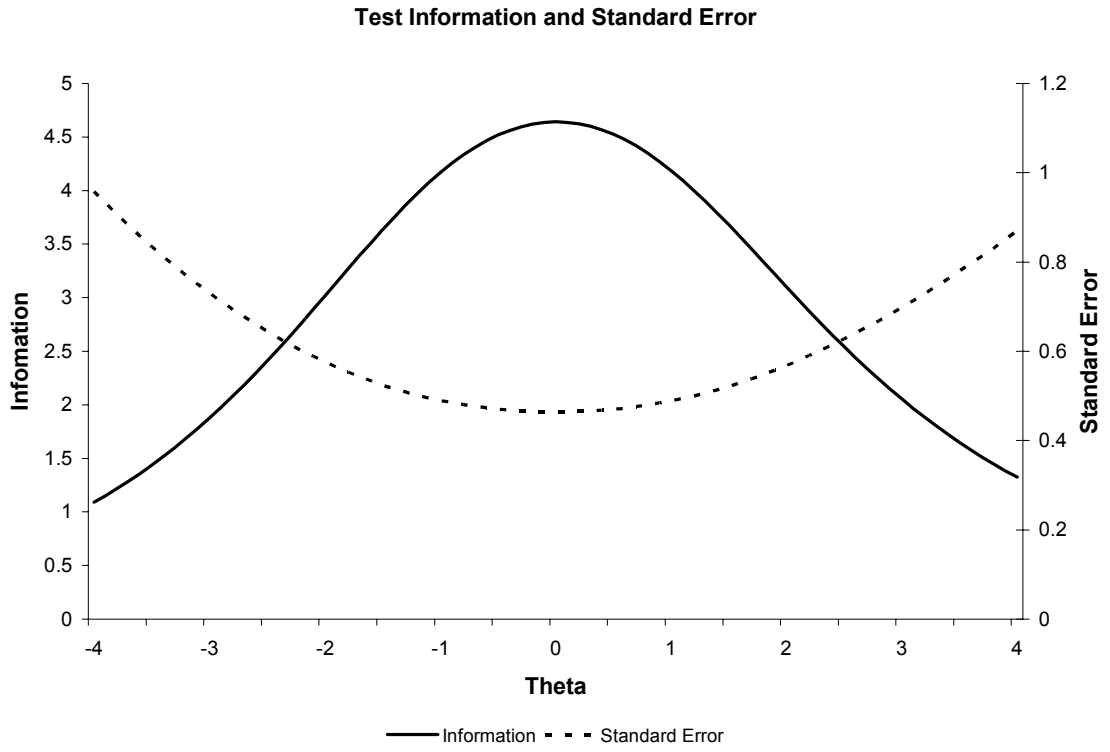


Figure 4-9: Total test information and standard error curves in the logit metric.

4.3 The Nominal Response Model

Another item response theory model which has potential to be helpful in the future development of the SCI and other concept inventories is Bock's nominal response model (Bock 1972). When multiple choice items are dichotomously scored, information contained in the incorrect responses is essentially lost because all the incorrect answers are collapsed into one category. One of the main ideas underlying the concept inventory movement is that important information about student understanding is contained in the incorrect responses as well as the correct responses.

For each item, the nominal response model provides a response curve for every response alternative, not simply the correct one. In this way all the information in the response pattern is used and this can help increase the accuracy of the theta estimates. In addition, it provides a more accurate picture of the item behavior across the theta distribution, including which distractors are more likely to be chosen at each point along the distribution.

Under the nominal response model, for an item X_i with *m* possible responses, the probability that an examinee will choose a particular response option *k* is represented by:

$$P_{ik} = \frac{\exp[a_{ik}(\Theta - c_{ik})]}{\sum_{j=1}^{m} \exp[a_{ij}(\Theta - c_{ij})]}.$$

For each value of Θ, the sum $\sum_{k=1}^{m} P_{ik} = 1$. There is no assumption that the response alternatives are ordered. The item parameters to be estimated are the parameters a_{ik}, c_{ik}.

The 38 items used on the SCI during the spring and summer 2005 semesters were analyzed using the nominal response model. The same data set was used for this analysis as for the 2PL model discussed before. The parameter estimation was conducted using MULTILOG (Thissen 2003). The parameter estimates obtained are included in Table 4-1.

Table 4-3: Nominal response model item parameter estimates.

Response:	a)	b)	c)	d)	e)	f)	a)	b)	c)	d)	e)	f)
Parameter	a_{i1}	a_{i2}	a_{i3}	a_{i4}	a_{i5}	a_{i6}	c_{i1}	c_{i2}	c_{i3}	c_{i4}	c_{i5}	c_{i6}
XD1	-1.51	-0.3	0.27	0.76	-0.04	0.83	-3.89	0.23	-0.04	2.62	-0.95	2.04
XD2	-1.37	0.69	0.17	0.5			-2.39	0.27	0.6	1.52		
XD3	-0.87	-0.36	0.87	0.36			-2.61	-0.21	2.17	0.65		
XD4	-0.84	-0.62	1.23	0.23			-2.25	-0.79	2.21	0.84		
XD5	-0.75	0.7	0.15	-0.1			-1.34	1.43	-0.23	0.13		
XD6	-0.29	-0.24	0.74	-0.22			-0.33	-0.85	1.64	-0.46		
XD7	-0.21	0.39	-0.17	0			0.29	1.29	-0.84	-0.74		
XD8	0.74	-0.46	1.01	-1.29			-0.21	-1.45	2.24	-0.59		
XD9	-0.77	-0.3	0.79	0.28			-1.36	-1.19	1.77	0.77		
XD10	-0.48	-0.27	-0.27	1.02			-0.42	-0.15	-1.03	1.61		
XG1	0.64	-0.36	-0.23	-0.05			0.09	-0.93	0.49	0.35		
XG2	0.14	-0.23	-0.06	0.15			-0.11	-0.36	-0.79	1.26		
XG3	-0.34	0.4	0.2	-0.12	-0.14		-1.74	0.46	1.49	-0.11	-0.1	
XG4	-0.59	0.53	0.02	0.04			-0.53	0.47	-0.07	0.14		
XG5	-0.37	0.69	-1.06	0.75			-1.63	2.11	-1.63	1.14		
XG6	0.81	0.15	-0.09	-0.87			0.45	1.67	0.52	-2.65		
XG7	-0.54	0.11	0.26	0.16			-0.91	-0.06	0.74	0.22		
XI1	-0.24	0.18	0.06				-0.53	0.53	0			
XI2	0.27	-0.33	-0.15	0.21			0.44	-0.5	-0.6	0.66		
XI3	-0.65	0.21	0.63	-0.18			-0.77	1.11	1.08	-1.41		
XI4	-0.08	0.17	0.39	-0.44	-0.04		0.46	-0.22	0.34	-0.02	-0.55	
XI5	0.41	-0.53	-0.25	0.2	0.17		-0.04	-0.36	-1.14	0.67	0.87	
XI6	0.36	-0.59	-0.62	0.86			0.46	-0.17	-0.76	0.47		
XI7	0.79	0.1	-0.74	-0.15			0.63	0.15	-1.04	0.26		
XI8	0.54	1.03	-0.87	-0.71			-0.16	2.97	-1.49	-1.31		
XI9	-0.44	-0.24	0.34	0.35			-0.7	-1.09	0.84	0.96		
XI10A	0.68	0.03	-0.17	-0.54			0.7	0.35	-0.32	-0.72		
XI11A	-0.45	0.09	0.37	-0.02			-0.88	0.21	0.86	-0.19		
XI12	-0.27	0.42	0.3	-0.46			-0.46	1.23	0.33	-1.11		
XP1	-0.38	-0.05	0.53	-0.1			0.04	-0.14	1.28	-1.17		
XP1A	-0.18	-0.19	0.64	-0.84	0.57		0.02	-0.06	1.4	-2.23	0.87	
XP2	0.87	-0.93	-0.04	0.09			0.56	-1.28	1.11	-0.39		
XP3	-0.69	0.69	-0.4	0.4			-1.24	0.42	-1.3	2.13		
XP4A	0.25	-0.79	0.47	0.06			1.41	-1.57	-0.57	0.74		
XP5	0.39	-0.28	-0.57	0.46			1.21	-0.64	-2.25	1.67		
XP6	-0.33	0.2	-0.36	0.49			-1.56	0.86	-0.72	1.42		
XP7A	-0.27	0.51	-0.37	0.13			-0.5	0.84	0.09	-0.43		
XP8	0.8	-0.19	-0.15	-0.46			0.16	0.26	-0.25	-0.18		
XP9	-0.14	-0.68	0.87	-0.04			-0.06	-1.66	1.65	0.07		

Consider question I3 (#17):

I3. 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 (Correct)
d) 5% of the measurements should be considered outliers

Based on the classical test theory analysis, this question was reasonably discriminating (usually the discrimination index was 0.4 or greater), its correlation with the total score was generally around 0.3. The distribution of responses was usually approximately 10%, 35%, 50%, 5%. This told us only the relative popularity of the distractors in the response set and that more people in the fourth quartile answered correctly than in the first. But, we have no sense of how the middle of the distribution answered this question or whether any of the alternative distractors remained popular at the upper ends of the distribution.

The item characteristic curve generated by the 2PL model affirms that the question discriminated reasonably well between high and low ends of the distribution, see Figure 4-10. The probability of answering correctly was less than one even at the upper end of the theta distribution indicating that this question was answered incorrectly by a good number of these students. This may indicate a persistent misconception, but does not identify which one.

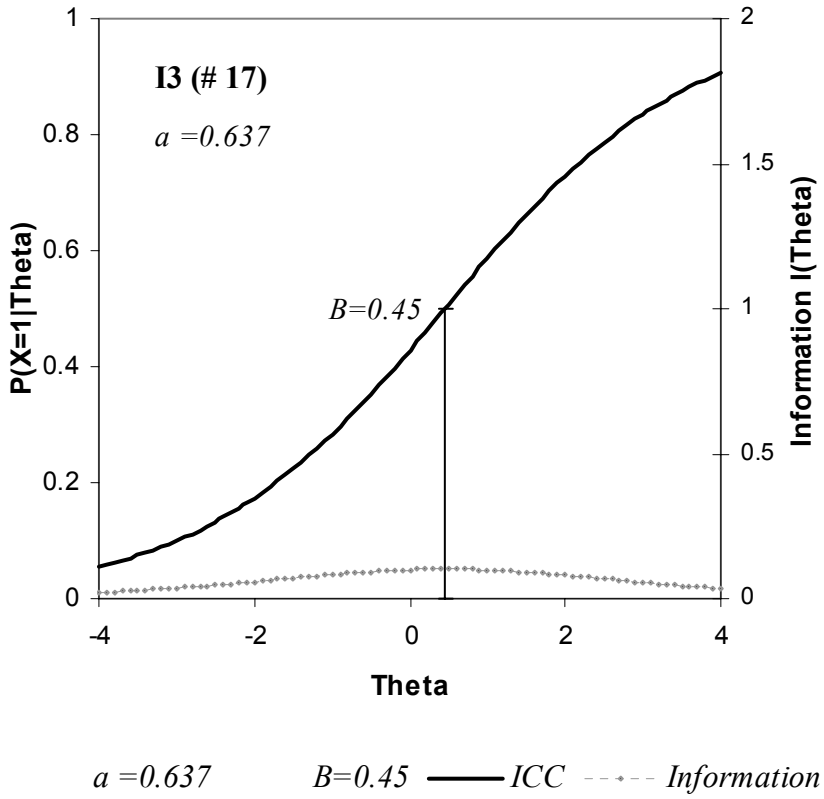


Figure 4-10: ICC for item I3 from the 2PL model.

By looking at the response curve generated by the nominal response model, a much clearer picture emerged about this question, see Figure 4-11. The probability of choosing response (a) was only greatest at the low end of the theta distribution and then tapered off to zero, indicating that this error was corrected as the understanding of statistics concepts increases. Response (b), however, was the most likely response in the interval between -2 and 0. The response curve did not quickly approach zero as theta increases, and this response accounted for a good proportion of students at the upper end of the distribution. This most likely indicates a misconception that persists along all points of the theta distribution.

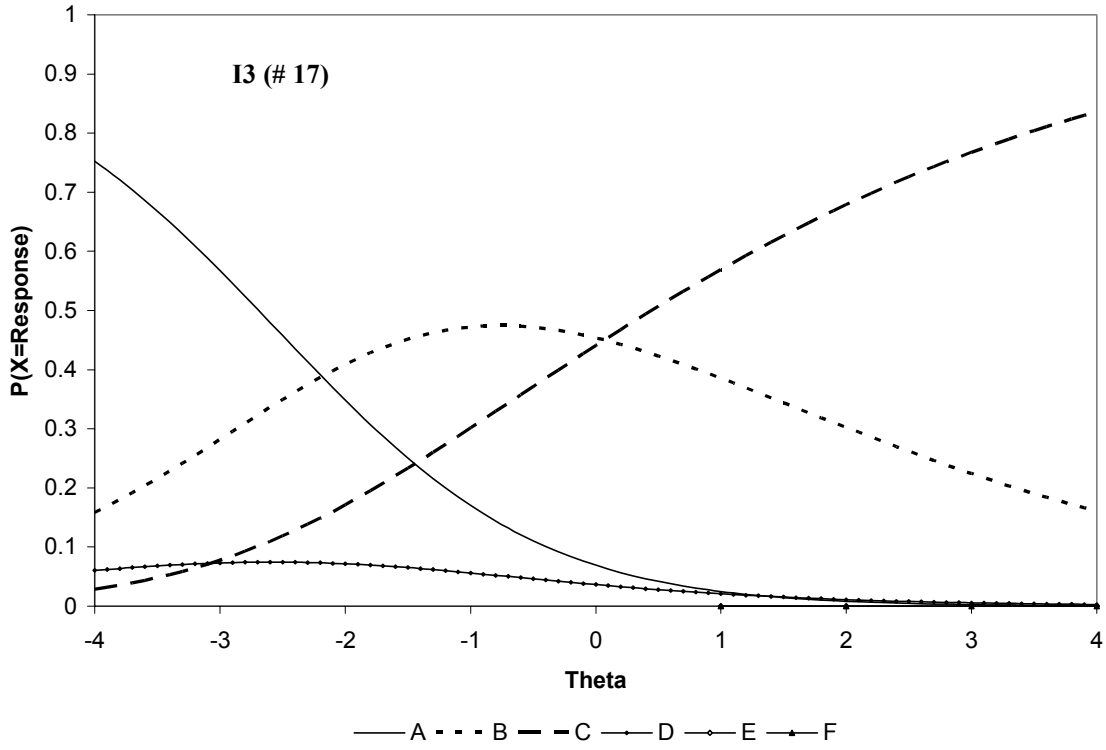


Figure 4-11: Response curves for item I3 from the nominal response model.

The response curves for question P1 and P1a (discussed above) are shown in Figure 4-12. The addition of response (e) lowered the discrimination parameter. From the nominal response model, it appears that response (e) was a strong distractor for the upper part of the distribution only. This may be a misconception that is generated after instruction and is a good candidate for targeted instruction. From a pedagogical standpoint, leaving this response alternative in the question may be more beneficial because it identifies a widely held misconception than any reliability gains that may be had by eliminating it.

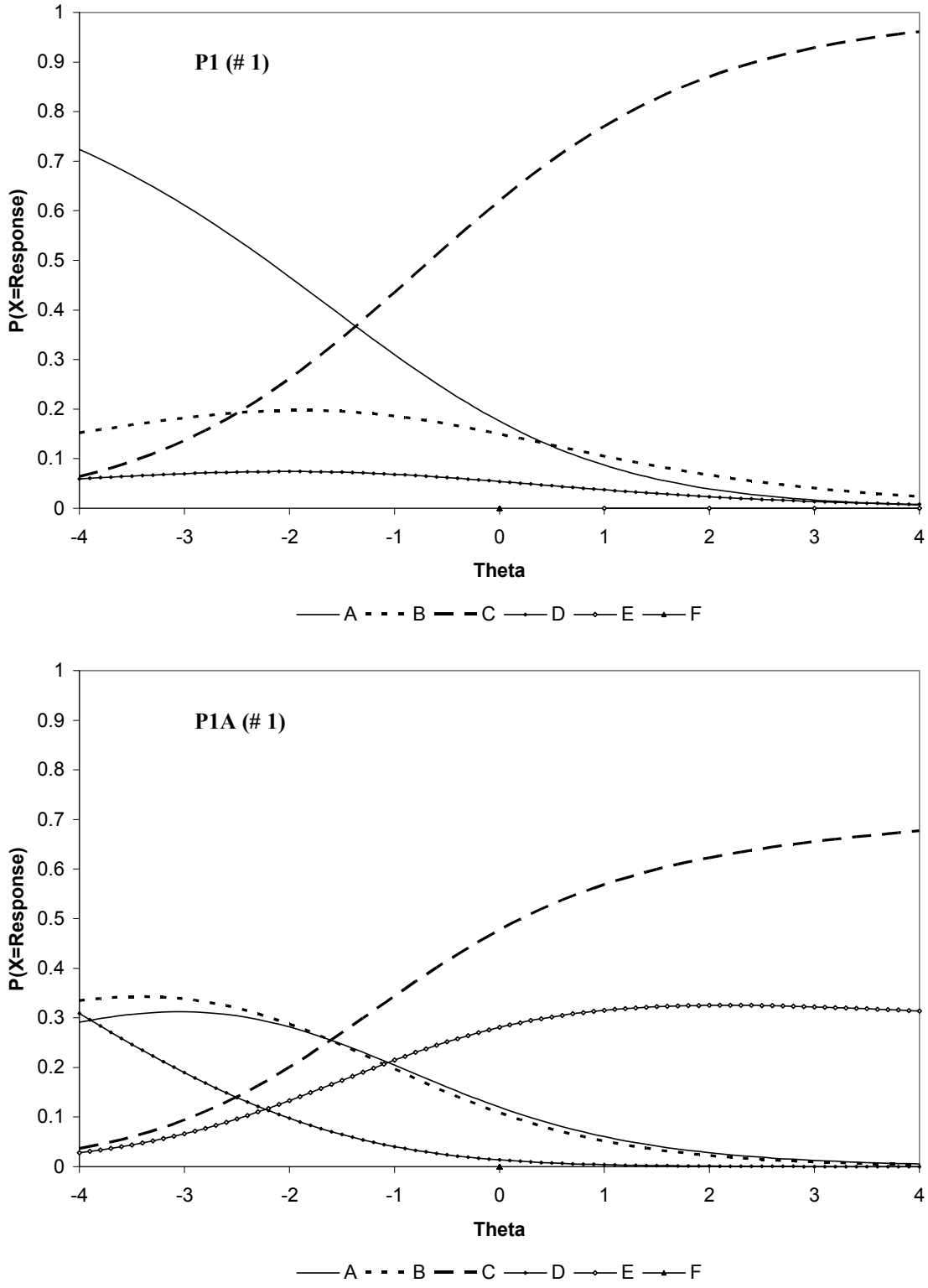
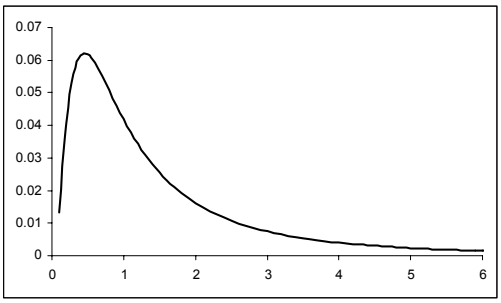
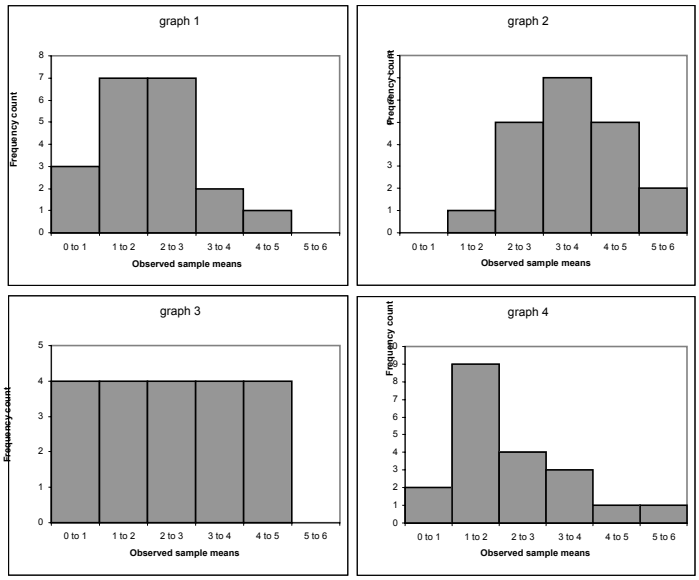


Figure 4-12: Response curves for item P1 and P1a for the nominal response model.

Question G2 (#14) asks students to identify the most likely distribution of sample means from a given probability density function:



G2. From the above probability density function, 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 (Correct)
b) graph 2
c) graph 3
d) graph 4

Typically between 60 and 70% of examinees chose response (d), 15% chose the correct response (a), with the remainder split between choices (b) and (c). The discrimination index for this question was low. The response curves shown in Figure 4-13 show that,

overwhelmingly, response (d) was favored by people at every level of theta. This question does not discriminate well between any examinees. However, it does indicate that confusion about sampling distributions and frequency distributions is pervasive.

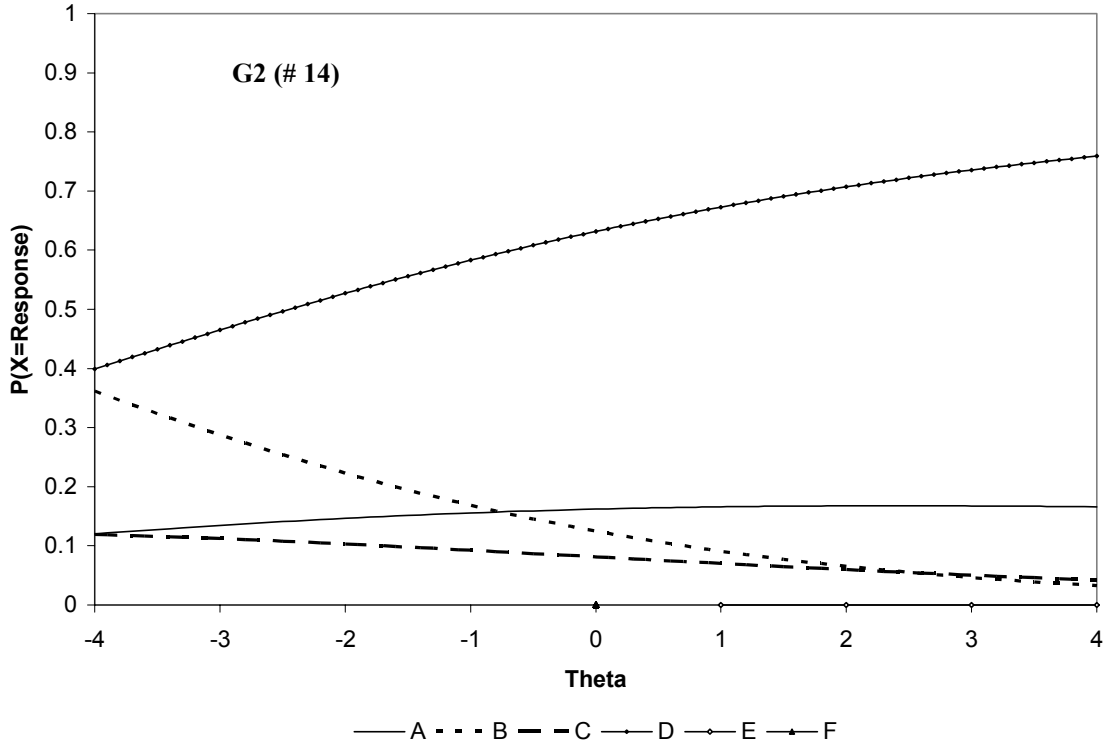


Figure 4-13: Response curves for item G2 for the nominal response model.

Question D9 (#26) typically had a very high discrimination index each semester. The majority of the responses were split between (c) and (d). Response (d) was added based on student responses to an open ended version of this question. Many variations of this same theme were submitted.

D9. 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 (Correct)
d) The numbers are evenly spaced on both sides of the mean

The response curves are shown in Figure 4-14. Response (d) was a strong distractor in the lower half of the theta distribution, but was also present in the upper half indicating that this idea --that the data are somehow mirrored on each side of the mean and this symmetry is captured by the standard deviation-- is a persistent misconception.

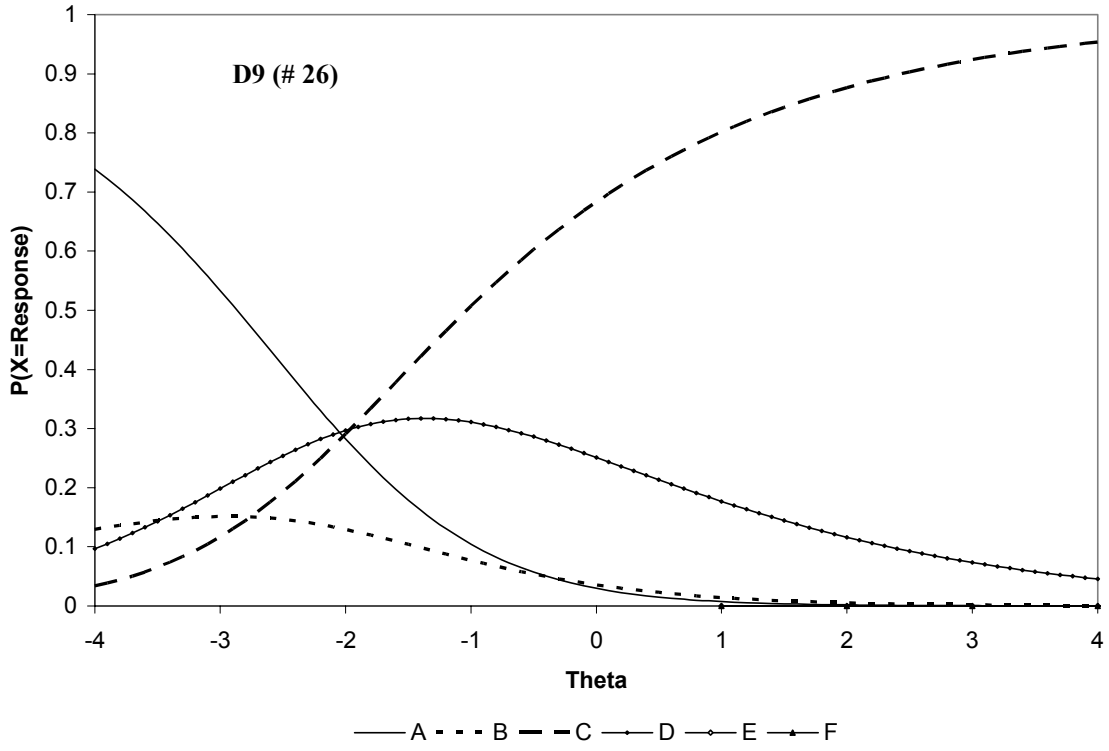
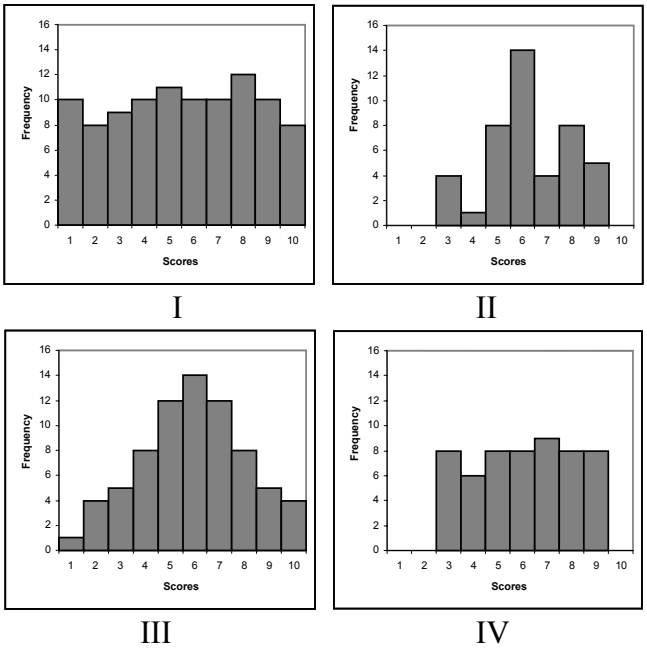


Figure 4-14: Response curves for item D9 for the nominal response model.

In question G6 (#30), students were asked to identify which distribution would have the greatest variance. Routinely response (b) was chosen by almost 60% of examinees. Focus group interviews indicated students focus on the bumpiness or

raggedness of the shape of the distribution with no thought given to any notion of spread or wideness or relation to center. This type of reasoning would indicate a fundamental lack of understanding about variance or at the very least a lack of visual representation for the concept.

G6. The following are histograms of quiz scores for four different classes. Which distribution shows the most variability?



a) I (Correct)
b) II
c) III
d) IV

The response curves, shown in Figure 4-15 indicate that this belief is widespread throughout the theta distribution. Since variation is one of the key ideas of statistics, this is an important misconception that could be addressed in a statistics course.

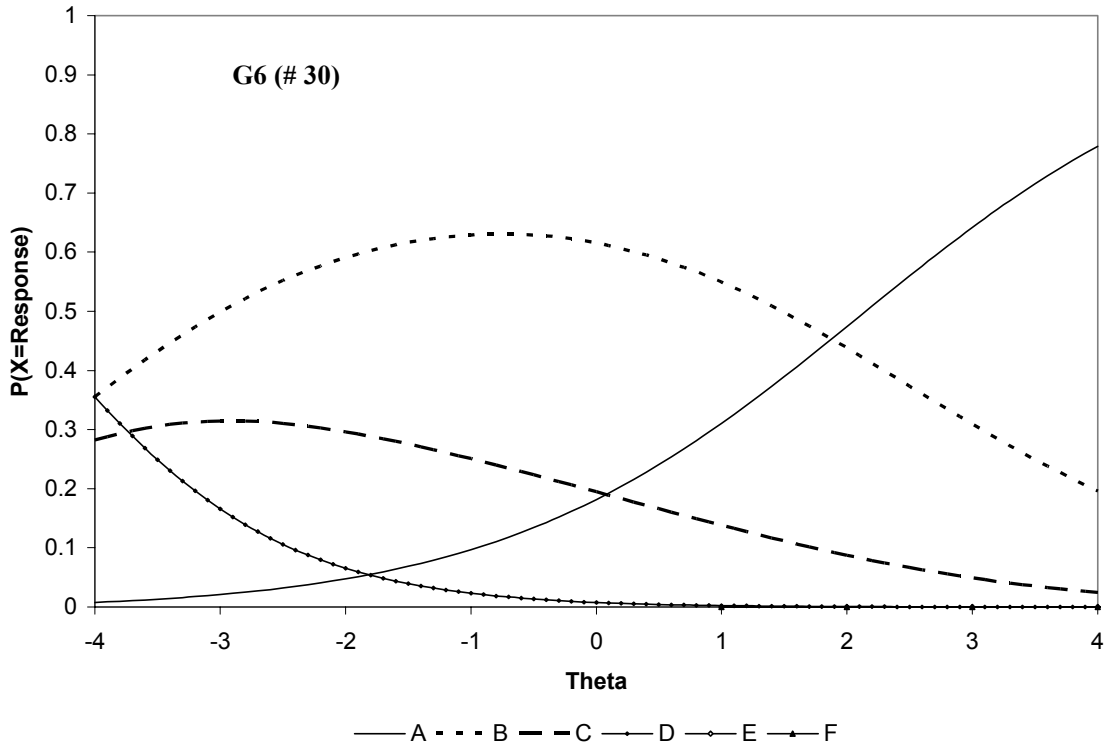


Figure 4-15: Response curves for item G6 for the nominal response model.

The nominal response model curves for all of the SCI items are included in Appendix B. This type of analysis will be used when making further revisions to the SCI. In addition, it may help pinpoint specific errors or misconceptions that may be useful in developing instructional strategies. Due to the large sample size required to employ these techniques, they typically could not be used in the beginning phases of the development of a concept inventory. However, once a suitably large data set has been amassed, the techniques are promising.

Chapter 5 Discussion and Directions for Future Research

The Statistics Concept Inventory (SCI) is a multiple choice instrument which seeks to assess conceptual understanding of material typically presented in an introductory statistics course. The SCI is modeled after the very successful Force Concept Inventory (FCI) developed by Halloun and Hestenes (Halloun and Hestenes 1985, Hestenes, Wells and Swackhamer 1992). Widespread use and research with the FCI over the last thirty years have led to a better understanding of physics instructional needs and many instructional innovations.

The SCI is still new to the concept inventory movement. Outcomes have been consistent with those found in other disciplines with new concept inventories. Total scores are low and gains are minimal with traditional instruction. It is hoped that with further use and research, the SCI will inform efforts to develop instructional strategies for statistics content. This chapter outlines future research goals for the SCI and presents some preliminary findings from research in these directions.

5.1 Scoring

Ultimately, the goal of an assessment instrument is to determine a meaningful score or set of scores for each examinee. To date, scoring results from the SCI have been based on the classical test theory model. In this framework, score results are reported as the number correct or percent correct obtained on a set of items. This score provides an

estimate of the true score for the examinee. Certainly, this method is widely used, but it has limitations. In particular, the score is dependent on the particular set of items from which it is obtained. This can make it difficult to compare different tests measuring a single domain or different forms of a test, since differences in length and difficulty will result in different scores for the same person.

An alternative scoring model is provided by item response theory. The underlying assumption of the item response theory model is that each examinee has an unobserved *ability, Θ,* which determines the probability of answering an item correctly. Instead of simply summing the number of correct answers, the pattern of correct and incorrect answers is used to determine an ability score or Θ estimate for each examinee. This score places the examinee along the latent trait distribution. Once the item parameters have been estimated, these parameters and the individual response patterns are used to determine the Θ estimate. The advantage of the IRT ability score is that it is independent of the particular set of items used. Items can be added or subtracted without affecting the score. The ability scale provides an "absolute scale" which could make comparisons between different forms and between different examinees more meaningful (Hambleton, Swaminathan and Rogers 1991).

5.1.1 Obtaining Ability Estimates

Once the item parameters have been estimated, they are used along with the individual's response pattern to estimate the individual's ability estimate. Procedures for estimating theta can be found in Hambleton and Swaminathan (1985) or Hambleton et al. (1991). For an instrument with *n* dichotomously scored items, each individual's response pattern can be written as a vector $(u_1, u_2, ... u_n)$ where each u_i is either 0 or 1. Because of

the assumption of local independence, the joint probability of the specific response pattern is the product of the individual item probabilities and can be written:

$$P(u_1, u_2, ...u_n \mid \Theta) = \prod_{i=1}^{n} P(u_i \mid \Theta)^{u_i} (1 - P(u_i \mid \Theta))^{1-u_i} .$$

When applied to an observed response pattern, this is called the likelihood function, denoted

$$L(u_1, u_2, ...u_n \mid \Theta) = \prod_{i=1}^{n} P(u_i \mid \Theta)^{u_i} (1 - P(u_i \mid \Theta))^{1-u_i} .$$

The likelihood function is a function of theta and the item parameters as defined by the item response model. In the case of the 2PL model, it is a function of Θ, the discrimination parameter α, and the threshold parameter β. Since the item parameters have been determined, the likelihood function can be evaluated for each value of Θ. The theta estimate for a given response pattern is the value of theta that maximizes the likelihood function.

Use of IRT for analysis of the SCI is in an early phase. Ability scores for the Summer 2005 SCI data were obtained using the item parameters described in chapter 4 for the 2PL model. The estimation procedure was carried out using BILOG analysis software and Bayesian estimation procedures (EAP) with a normal prior distribution. The percent correct scores on the SCI and the corresponding theta estimates are shown along with the test characteristic curve (TCC) in Figure 5-1. The TCC is defined by

$$TCC = \frac{1}{n} \sum_{i=1}^{n} P_i(\Theta)$$

where $P_i(\Theta)$ are the item characteristic curves. The TCC is an average of the item characteristic curves for the item included on the test. The two measures are highly

correlated for the Summer 2005 SCI data ($\rho = 0.964, p < 0.0001$), as expected. In general, when the model and data fit is acceptable, the data points are expected to be scattered along the TCC. This type of analysis is one suggested method for assessing model-data fit (Hambleton, et al. 1991). Scatter along the TCC is expected due to measurement error. The SCI data do appear to fall closely along the TCC. It should be noted that the sample size here is small and similar analyses should be conducted with larger sample sizes.

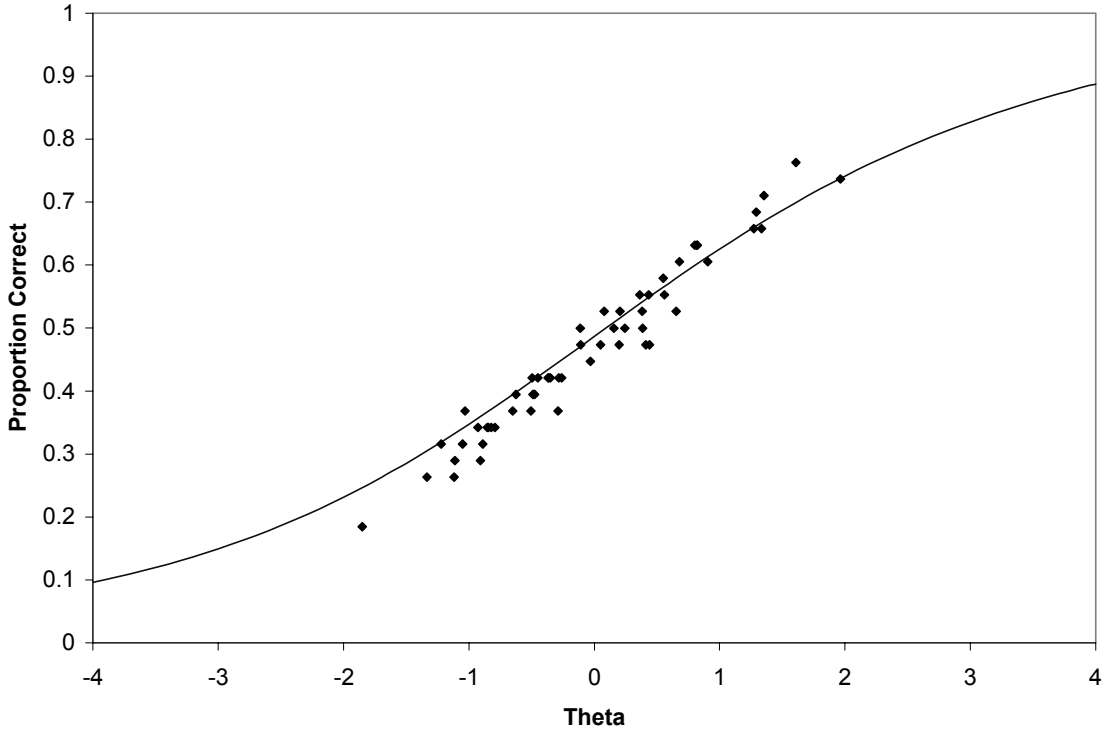


Figure 5-1: Proportion correct vs. theta for the SCI post test data from the summer 2005 administration. The data is superimposed over the test characteristic function which shows the expected proportion correct for each point along the theta distribution.

Once theta estimates have been obtained, decisions about how the scores should be reported must be made. Throughout this analysis, the theta distribution has been set to have a mean of 0 and a standard deviation of 1. However, when reporting test scores, this

scale is not the most easily interpreted. The ability scores can be rescaled using a suitable transformation to make score interpretation easier. The true-score scale is often used for this purpose. This transformation yields an estimate of the true score or the domain score, as from classical test theory. Under the classical test theory model, the true score is estimated by the total correct or the proportion of correct responses. The domain score estimate is obtained from the theta estimate $\hat{\Theta}$ by the transformation

$$f(\hat{\Theta}) = \frac{1}{n}\sum_{i=1}^{n} P_i(\hat{\Theta}),$$

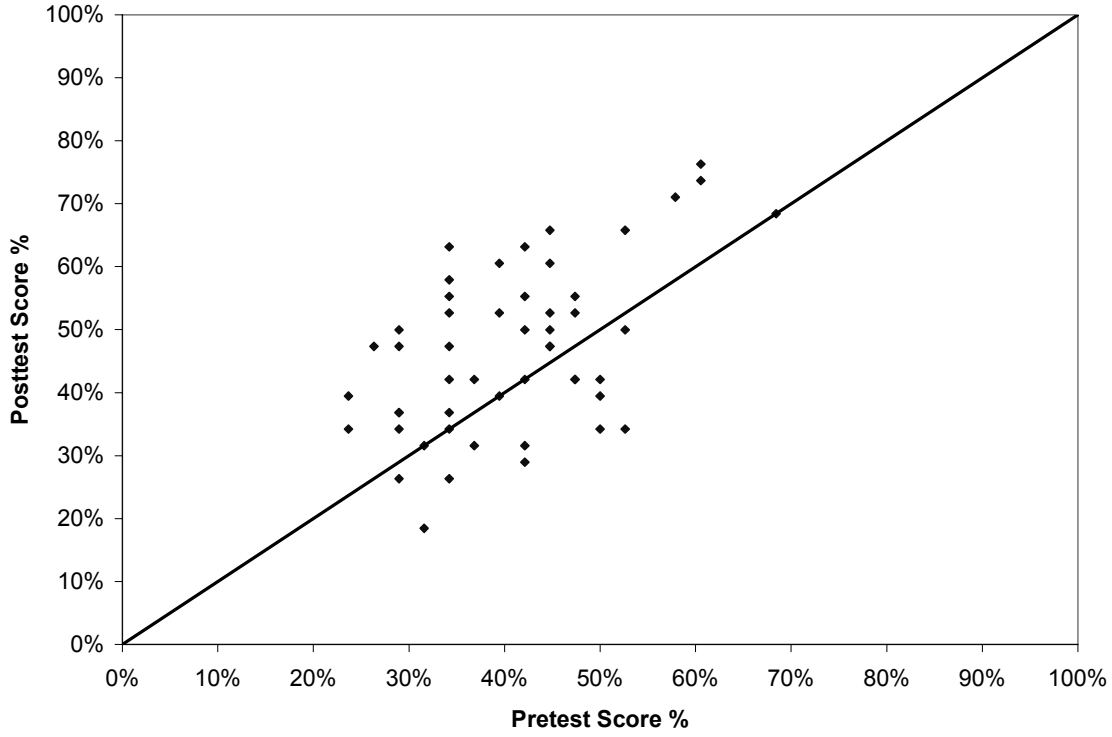
and is simply the value of the test characteristic function for the estimated value of theta. The scores can range from 0 to 1 (or from 0% to 100%).

The advantage of the estimated domain score over the observed number correct score is that the estimated score is independent of the particular items used. This makes it possible to compare multiple forms, perform test equating, and even to make score predictions for items which the examinee has not taken, but for which item parameters are known. These characteristics can be utilized as the SCI is revised further.

5.1.2 Gains

The theta scale also provides another mechanism for looking at gains from pre to posttest. Gains on the SCI have typically been reported as raw gain, which is the change in percent from pre to posttest, or as normalized gain, which is the ratio of the change from pre to posttest to the total possible gain. Because these measures are dependent on the items and the version of the test used, semester to semester comparisons can be difficult to interpret. Also, in the past, we have refrained from making revisions to the SCI between pre and posttest administrations in order to make pre and posttest

comparisons meaningful. Using the IRT scoring methods will eliminate these problems and allow gains to be considered along the "absolute scale" that the theta scale provides. Figure 5-2 shows the pre and posttest scores for the summer 2005 administration as both the observed percent correct and the estimated theta values. Note that the two metrics provide very similar information as expected.



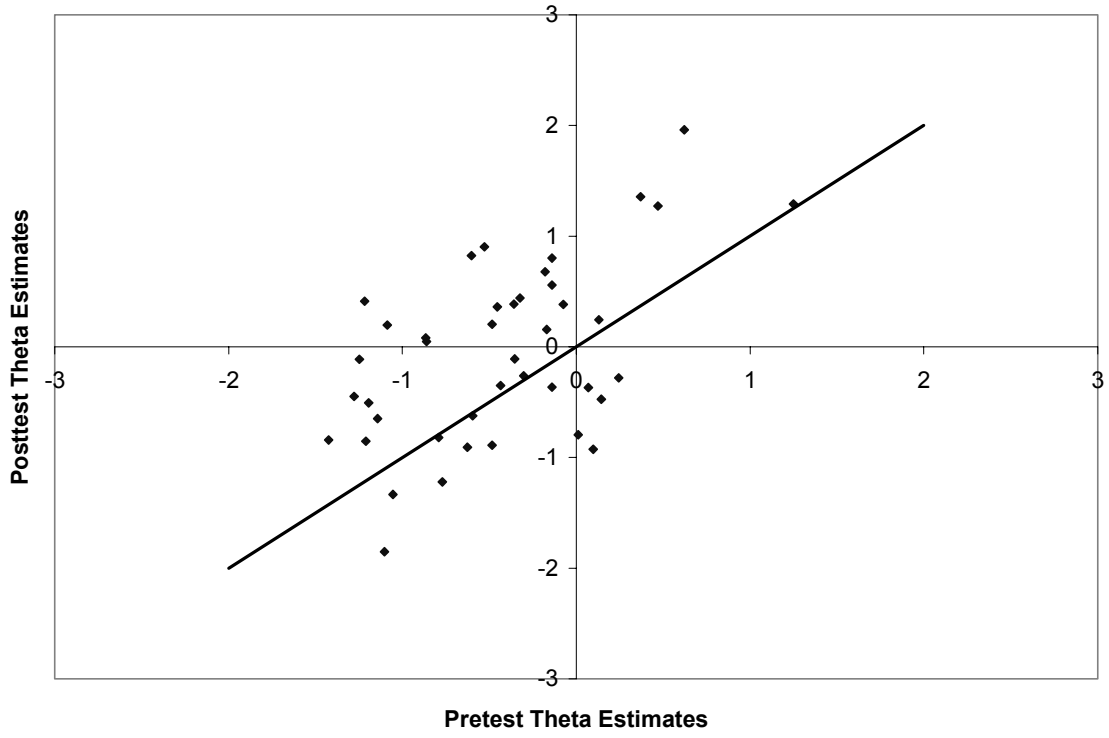


Figure 5-2: Pre and posttest scores for the summer 2005 administration. The top plot shows the observed scores as correct percentages. The bottom plot shows the scores as theta estimates. The line represents no change in the scores. Points above the line demonstrated a positive gain from pre to posttest. Those below represent negative gains.

5.2 Model-Data Fit

In order to make reliable inferences using IRT analysis methods, it is very important to determine whether the model fits the data. Hambleton et al. (1991) recommend utilizing a variety of analysis methods rather than relying on statistical goodness of fit tests, since these tests are sensitive to sample size. They advocate assessing three areas to determine model-data fit: the validity of the assumptions of the model, the extent to which the expected properties of the model hold, and the accuracy of the model predictions. Some preliminary analysis of the SCI is presented here, but more in depth analysis should be conducted.

The assumptions of the 2PL IRT model are that the test is unidimensional, involves minimal guessing, and is non-speeded (i.e. time is not a factor in test performance). While it is not expected that any real data set will completely meet all of the model assumptions, it is desirable to determine whether the assumptions are reasonably met.

5.2.1 Unidimensionality Assumption

One method for assessing the unidimensionality of the model is to plot the eigenvalues of the inter-item correlation matrix. A high ratio of the first to the second eigenvalue provides evidence of a dominant first factor. Tetrachoric correlations from the SCI data used in the IRT analysis were computed using the analysis software TESTFACT (Wood 2003). Figure 5-3 shows the plot of the first eighteen eigenvalues of the inter-item correlation matrix. The ratio of the first to second eigenvalue is 2.54. Ideally, this ratio would be larger; however, since the second eigenvalue is very similar to the remaining smaller eigenvalues, some evidence of a dominant first factor is obtained.

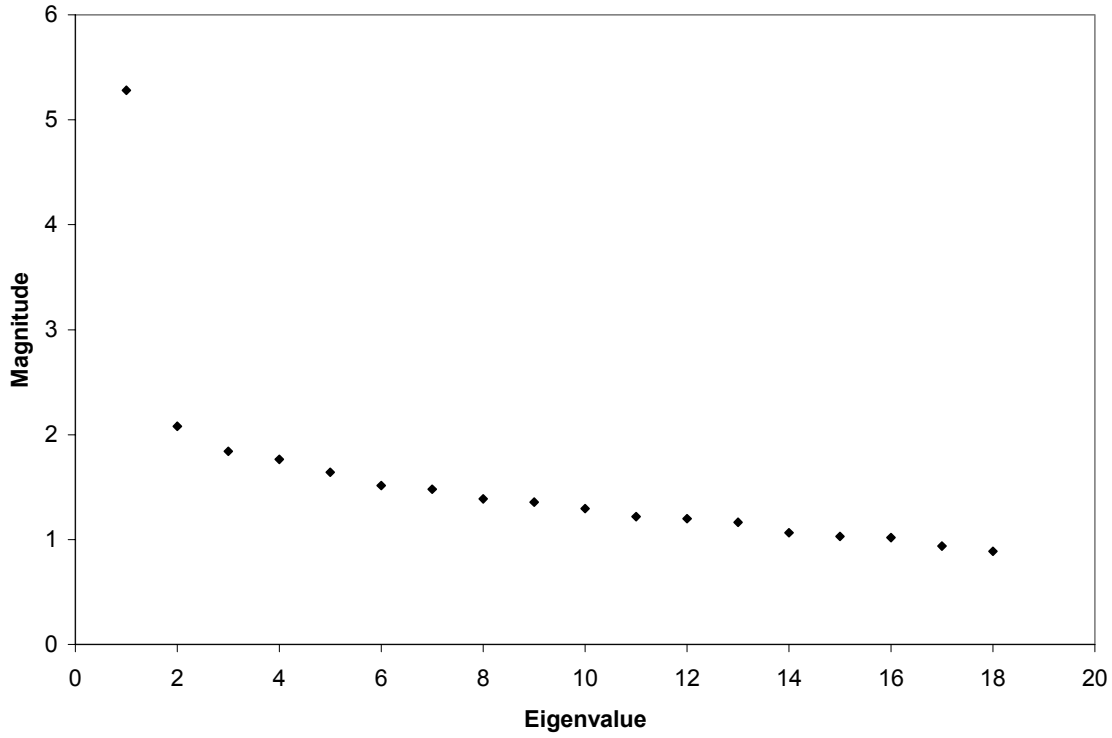


Figure 5-3: Plot of the 18 largest eigenvalues of the inter-item tetrachoric correlation matrix.

Another suggested method for assessing unidimensionality is to select a subset of items that appear to measure a different ability than the whole test. Item parameter estimates for these items are obtained as part of the whole test and as an isolated subtest. The two sets of estimates are then compared. If the estimates are equal within error, this gives evidence of unidimensionality. If the estimates are not equal, then test performance would be dependent on the selection of items, which is a violation of the unidimensionality assumption.

To examine this for the SCI data, the test items were divided into the four topic areas: probability, descriptive statistics, inferential statistics, and graphical. Parameter estimates for the 2PL model were obtained for the four subtest areas using the same data

set used in the whole test analysis. Figure 5-4 shows the threshold estimates for each of the four subgroups compared to the estimates obtained from the whole test. Nearly all of the questions lie closely along the line of equality.

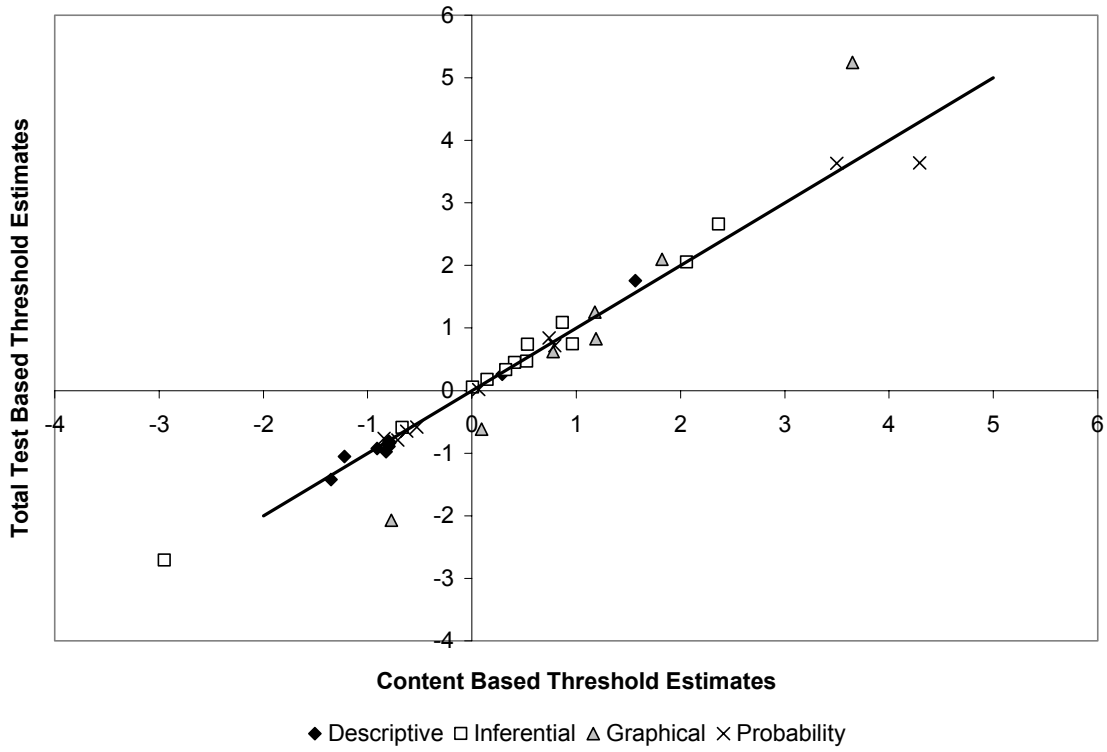


Figure 5-4: A comparison of threshold (β) estimates obtained from the whole test and from subgroups of items divided by content area.

A similar plot is included in Figure 5-5 for the slope parameter estimates. The scatter around the line of equality is much greater for the slope estimates. However, with the exception of the graphical subgroup, almost all of the questions are within error of the line of equality. The questions within this grouping cover topics which would fall within the other three subgroups, but share a common graphical format. For instance, one question asks about the relative variation among four distributions presented as

histograms. So, this question requires knowledge of variation (a descriptive statistics concept) and knowledge of reading and interpreting histograms. Even though graphical representation and interpretation is a key component of the statistics curriculum, this separate presentation format may be a second smaller dimension.

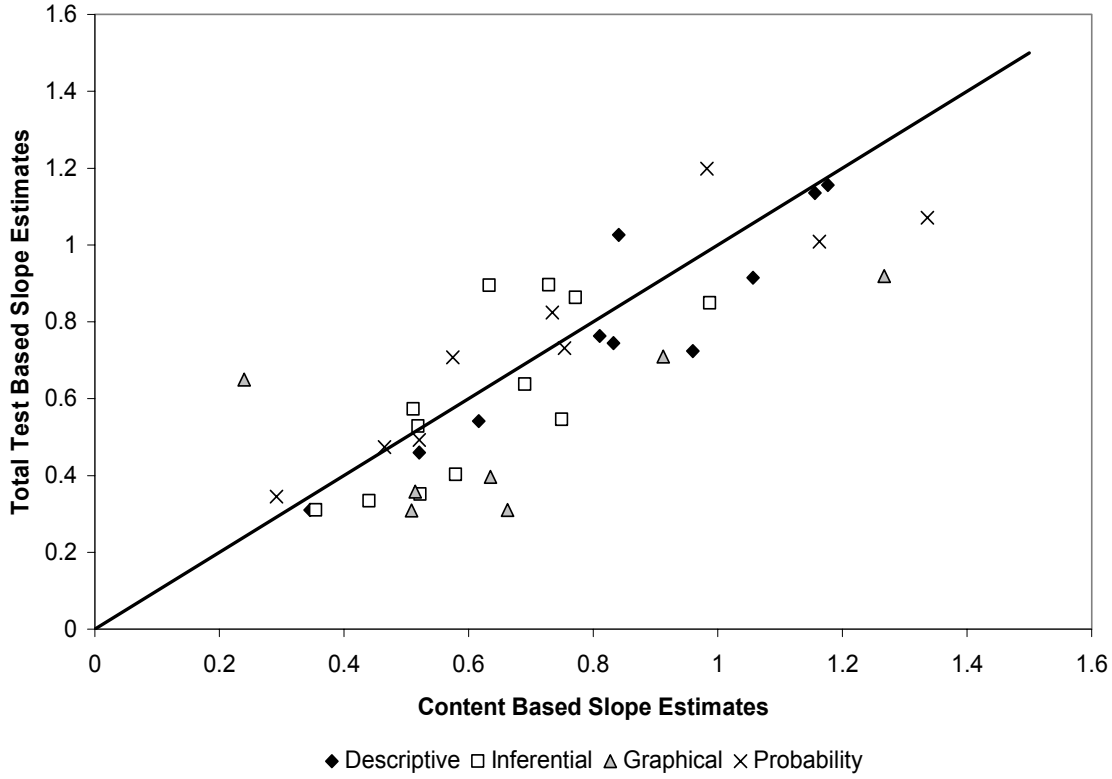


Figure 5-5: A comparison of slope (α) estimates obtained from the whole test and from subgroups of items divided by content area.

The methods presented here provide only some heuristic evidence in support of using the IRT model. Other possible methods of unidimensionality assessment that could be used include non-linear factor analysis, residual analysis of the one factor model and Stout's technique for assessing essential dimensionality. However, no single method for

assessing dimensionality is clearly recommended at this time and this is an area of active research (Embretson and Reise 2000). There is evidence that the IRT models are somewhat robust to the presence of multidimensionality, though to what extent is not clear (Reckase 1979, Kirisci, Hsu and Yu 2001).

5.2.2 Minimal Guessing Assumption

To check the assumption of minimal guessing, one suggested method is to check the performance of low-ability students on the most difficult items. If low ability students score near zero on these questions, this provides evidence that the assumption of minimal guessing may be reasonable. To examine this with the SCI data, the performance of the students at or below the 10th percentile on the total score for the spring 2005 posttest administration was examined on the ten most difficult questions. The most difficult questions were chosen based on their threshold parameter estimates. Figure 5-6 shows the percentage of these students who answered correctly or incorrectly on these 10 questions. While the percent correct for many of the questions is at or near zero, for almost half of the questions, the percent correct is close to 20%, near what would be expected from random guessing. So, as expected for most multiple choice tests, guessing does seem to be a factor in some questions on the SCI. The 3PL model may achieve a better model-data fit. However, at this time the SCI data set is not large enough to reach a convergent solution for the 3PL model. Once more data have been collected this model should be evaluated.

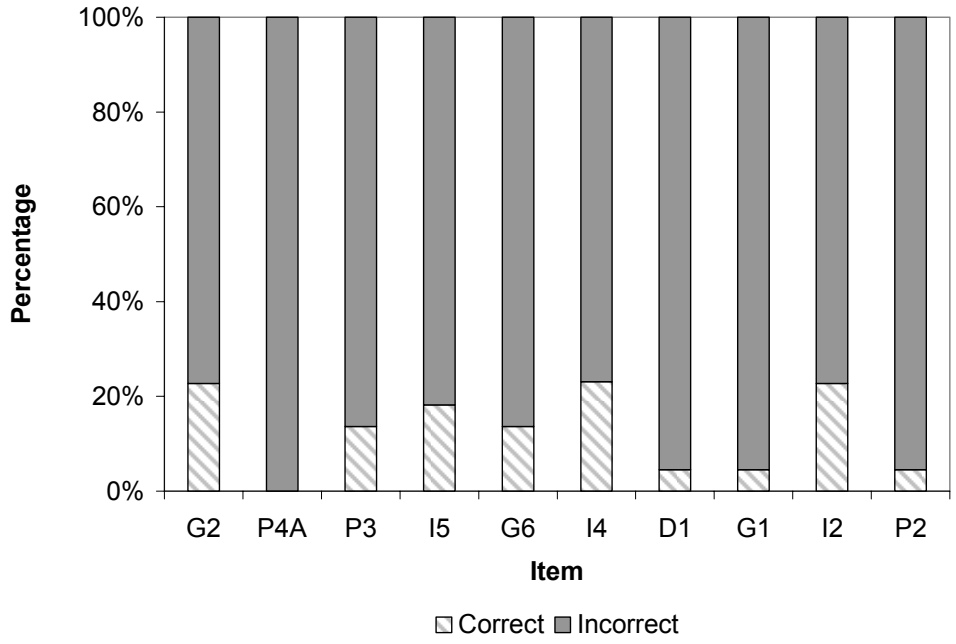


Figure 5-6: The percent of students whose total scores were at or below the 10th percentile who answered correctly or incorrectly on the 10 most difficult SCI questions. The data is from the spring 2005 post test. The number of students at or below the 10th percentile was 22.

5.2.3 Non-speeded Assumption

The third assumption of the 2PL IRT model is that the test is non-speeded. Since almost all of the examinees complete all of the questions on the SCI, it can be assumed that speed is not a critical factor in test performance and the non-speeded assumption is satisfied by the SCI. While the assumptions of the IRT model are not all perfectly satisfied, there is some evidence that the data partially satisfy the assumptions. The 2PL model may yield sufficiently accurate and stable parameter estimates to be a viable and useful research tool. Examining the model features and behavior can help to determine this.

5.2.4 Model Features and Behavior

The key feature of the IRT models is the invariance of the parameter estimates. The item parameter estimates should be independent of the specific group of examinees used to determine them. The ability (theta) estimates should be independent of the specific items used to estimate them. The parameter invariance feature is what makes it possible to make meaningful comparisons between examinees, even if they have taken different forms of the exam.

To check the invariance of the item parameter estimates, the parameters (α and β) can be estimated from different subgroups of the population. The parameters should be invariant for any subgroup: based on gender, race, performance, etc. The relationship between the estimates for two groups should be linear except for scatter due to error. Randomly equivalent groups can be used to obtain baseline data for comparison to subgroups selected for a specific criterion.

A preliminary check of item parameter invariance for the SCI was conducted by generating two random subgroups from the data used in the initial analysis. A random number between 0 and 1 was generated for each examinee and those greater than 0.5 were assigned to group 1 and the others to group 2. Parameter estimates were then obtained for each group separately using the BILOG analysis software. The scatter plots for the slope and threshold parameter estimates are shown in Figure 5-7 and Figure 5-8. The threshold parameters are quite stable between the two groups, the regression equation obtained is $y = 1.0089x - 0.0573$ and the correlation 0.96. The parameter stability is less for the slope parameter, the correlation is only 0.70. It should be noted that with the data set divided, only a few of the items within each group had a sample size

of at least 500. Most items had less than 300 responses and a few had less than 100 responses. Large sample sizes are required to achieve good parameter estimates as the small number of responses to each item would increase the error in the parameter estimation.

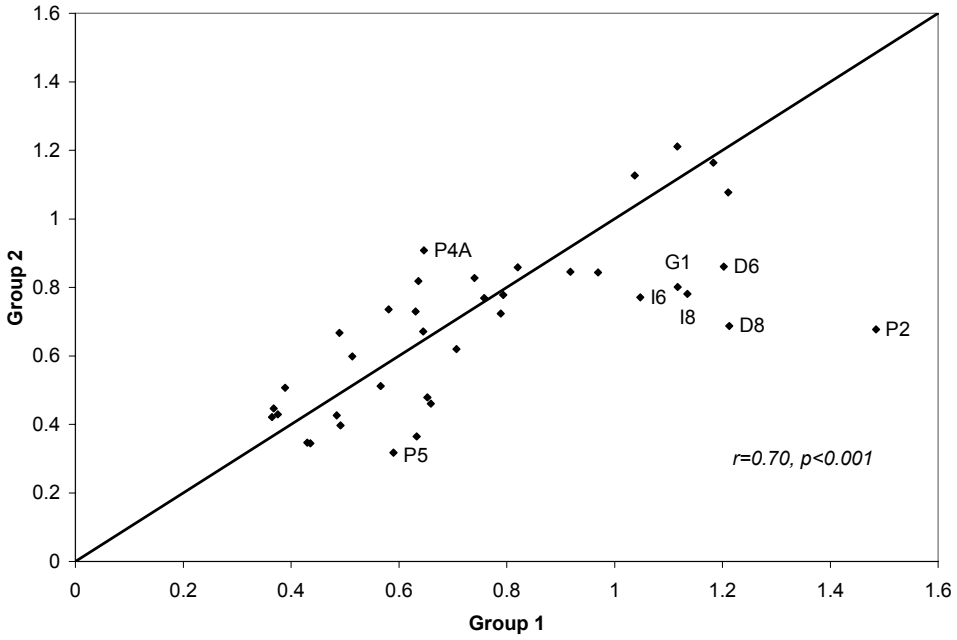


Figure 5-7: A comparison of slope (α) parameter estimates obtained for two randomly selected subgroups of examinees. Items lying farthest from the line of equality are labeled with their master number. The correlation coefficient is also shown.

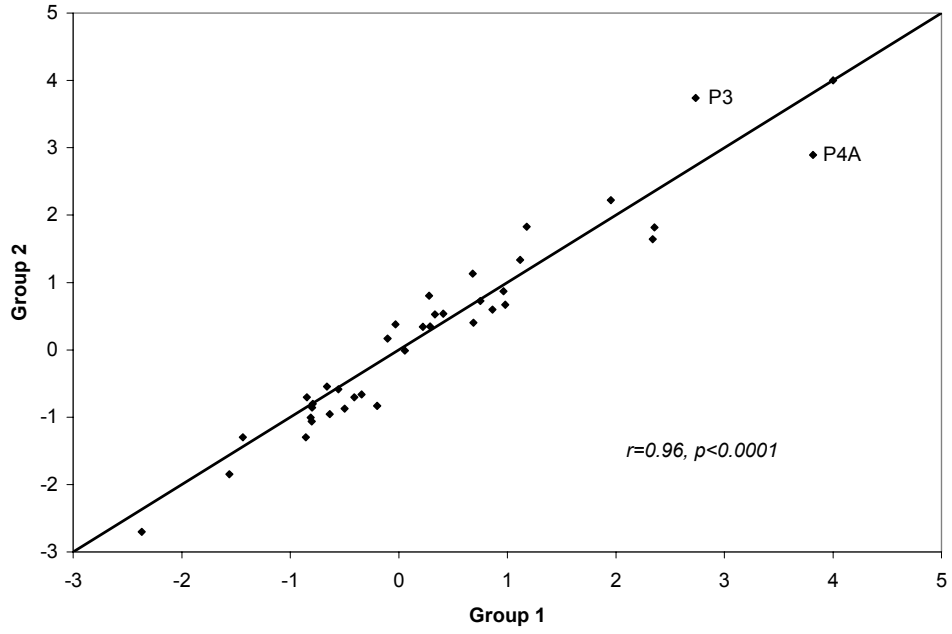


Figure 5-8: A comparison of threshold (β) parameter estimates for two randomly selected subgroups of examinees. Items lying farthest from the line of equality are labeled with their master number. The correlation coefficient is also shown.

Ability parameter estimate invariance can be checked by comparing ability estimates obtained from different subsets of the items. Test items can be divided based on a variety of criteria, including item difficulty or content. To evaluate the stability of ability estimates, two analyses were conducted with the SCI data. The item parameter estimates used in each analysis were those obtained from calibrating the whole test. For the first analysis, the items were placed in order by their master number and split every other item into two groups: even and odd. The master numbers organize the questions topically, so this split placed half of the descriptive question into each group, half of the probability, etc. Ability estimates were then obtained for each examinee based on each "subtest" using BILOG. The comparisons based on all post test data are shown in Figure

5-9. The correlation between the two scores is 0.55. Most of the estimates agree within error, but about 15% differ in excess of the estimated error.

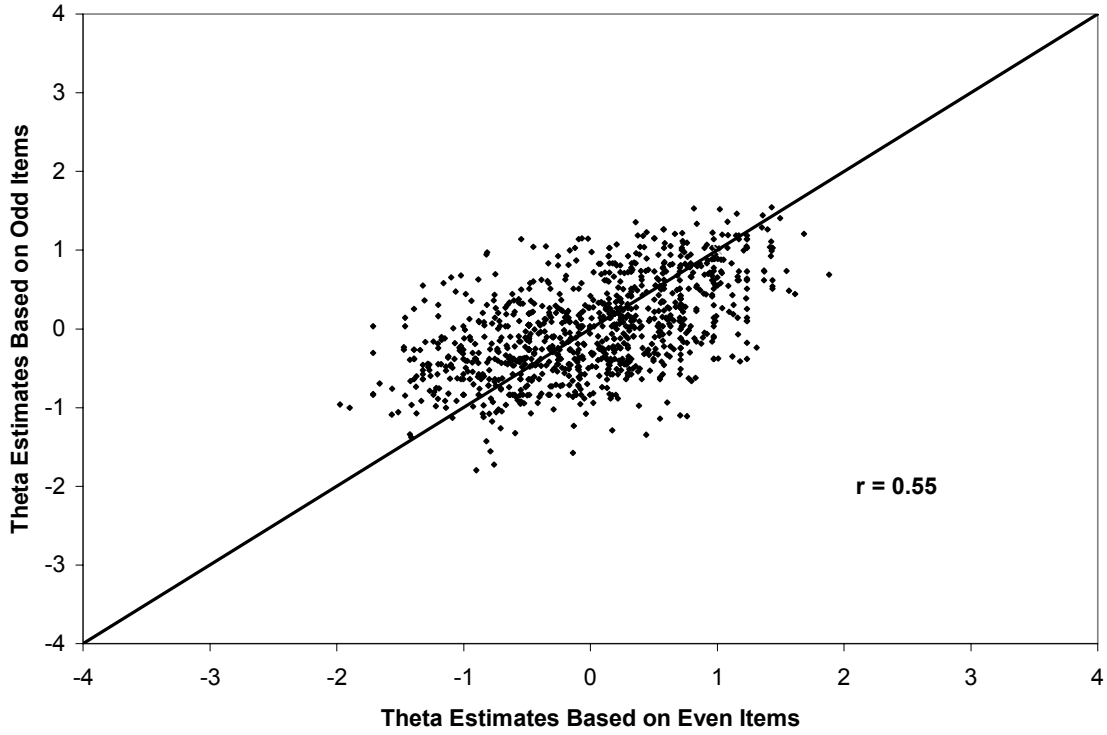


Figure 5-9: Comparison of ability estimates based on even numbered items and odd numbered items only.

For the second analysis, the SCI items were divided into two groups based on their item difficulty. The items were ranked based on their threshold parameter estimate from easiest to hardest and easiest 19 questions form one subtest while the hardest 19 formed the other subtest. Ability estimates were obtained for all examinees for each subtest, again using item parameter estimates based on the whole test and BILOG. The correlation between the scores was 0.47. The results are shown in Figure 5-10. About 23% of the scores differ in excess of the error estimate. A higher correlation between the two scores would be more desirable. Since the plots are similar and most of the scores

agree within error, the low correlation may be more reflective of too large a standard error of measurement than lack of parameter invariance.

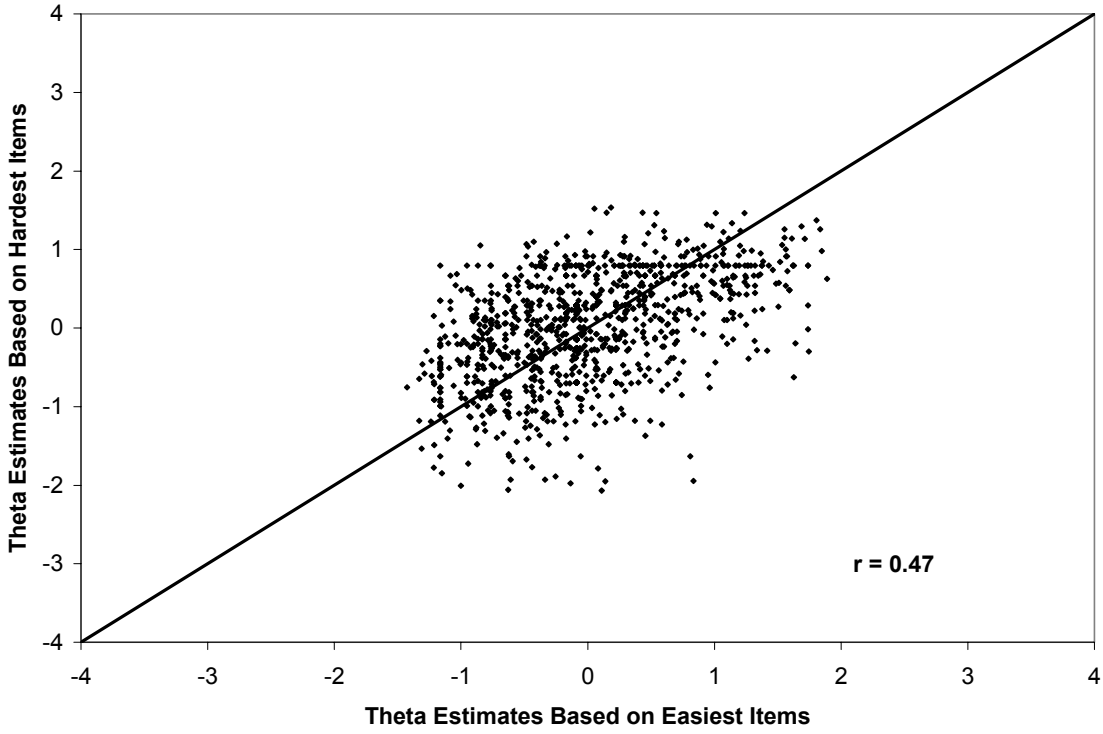


Figure 5-10: Comparison of ability parameter estimates based on subsets of the easiest and hardest items for data from summer 2005.

The third area to check for model data fit recommend by Hambleton et al (1991) is to assess how well the model prediction matches observed and simulated test results. There are several methods proposed to carry out this type of analysis. One method is to compare the estimated ability scores to the observed proportion correct score as shown in Figure 5-1. Another method is to compare the observed proportion correct scores for different intervals along the theta distribution to the expected proportion correct predicted by the model for individual items.

This type of analysis was carried out for a few of the items from SCI. Item D10 is shown in Figure 5-10. The theta scale between -3 and 3 was divided into 10 intervals of

width 0.6 and the proportion of examinees answering item D10 correctly was calculated for each interval. These observed proportions are plotted over the item characteristic curve. The corresponding residuals are also shown. The data are expected to be scattered along the item characteristic curve. For comparison, the same type of plot was prepared for the 1PL model. This is shown in Figure 5-12. From this comparison, we see that the 2PL model fits the observed data much better than the 1PL model.

The 2PL item characteristic curves and observed proportion correct are also shown for items D6 and P2 in Figure 5-13 and Figure 5-14 respectively. For these three items, we see that the 2PL model predictions are fairly accurate. This type of analysis can be carried out for each item on the SCI. Once the 3PL model can be evaluated, the item plots can be compared for the two models to ascertain whether significant improvements in model fit are obtained.

These preliminary results indicate that, while the model assumptions are not ideally satisfied (as with any real data set), the 2PL model is viable for research purposes in further developing the SCI. Future research should include additional model-data fit analysis of the type presented here, an assessment of the fit of the multiple choice model presented in chapter four, and fitting and assessing the 3PL model when sufficient data have been gathered.

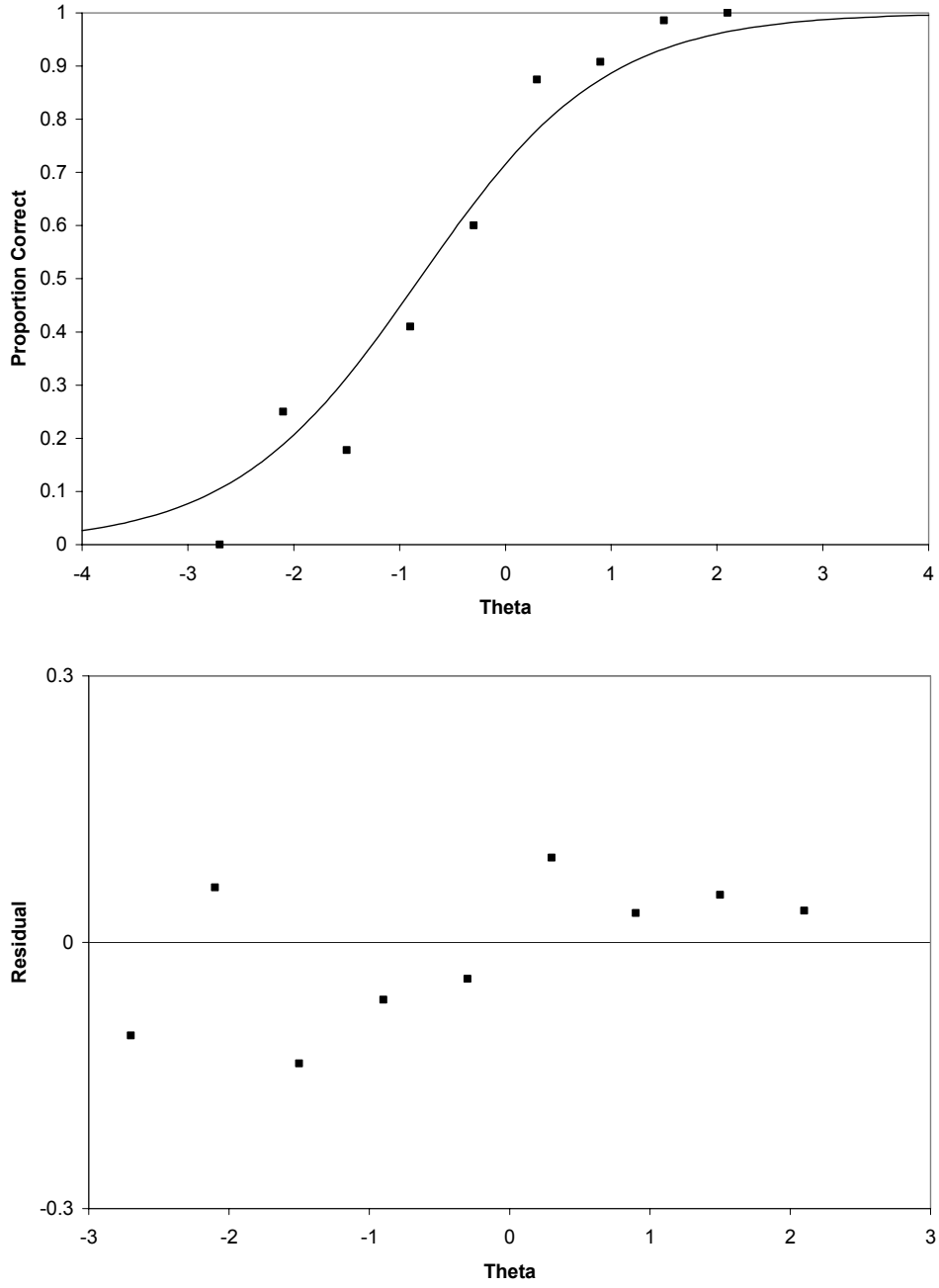


Figure 5-11 Observed and expected proportion correct for item D10 (#29) based on the 2PL model. The line represents the item characteristic curve (and the expected proportion correct) and the data points are the observed proportion correct within the interval. The interval widths are 0.6; the midpoint of the interval is used as the observed theta value. The corresponding residual plot is shown below.

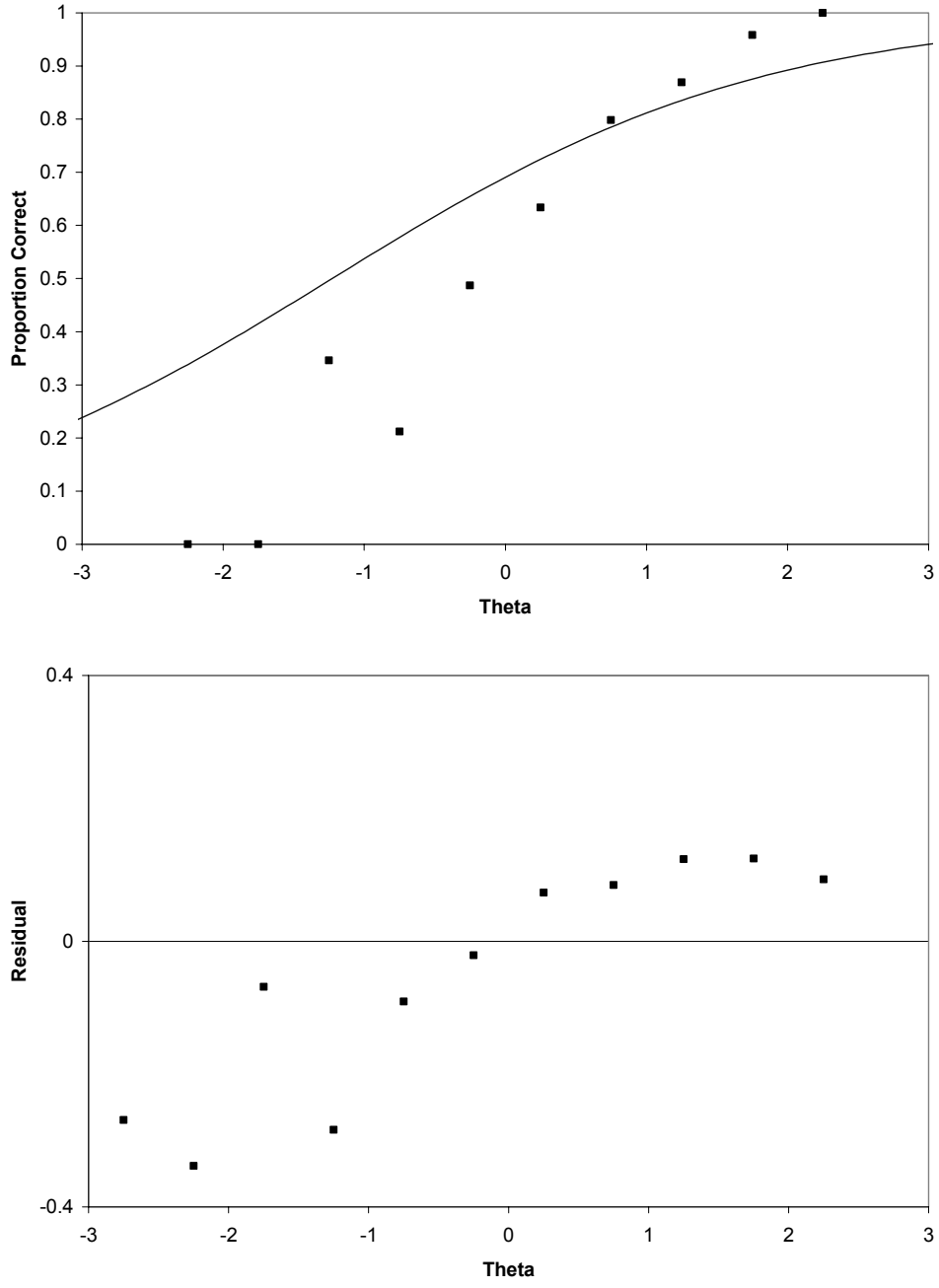


Figure 5-12: Item D10 observed proportion correct versus predicted proportion correct based on 1PL model. The corresponding residuals are also shown.

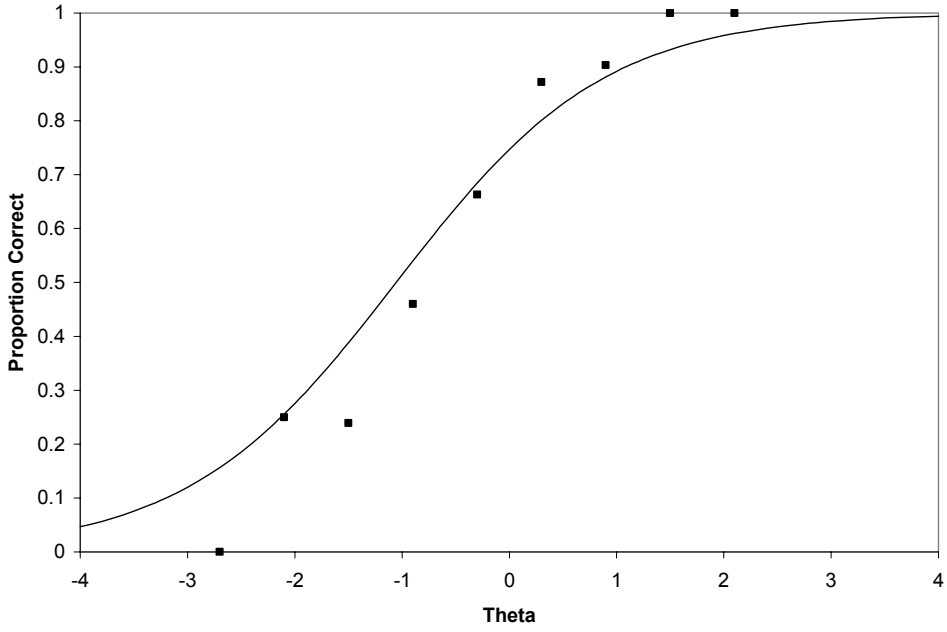


Figure 5-13: Observed and predicted proportion correct for item D6.

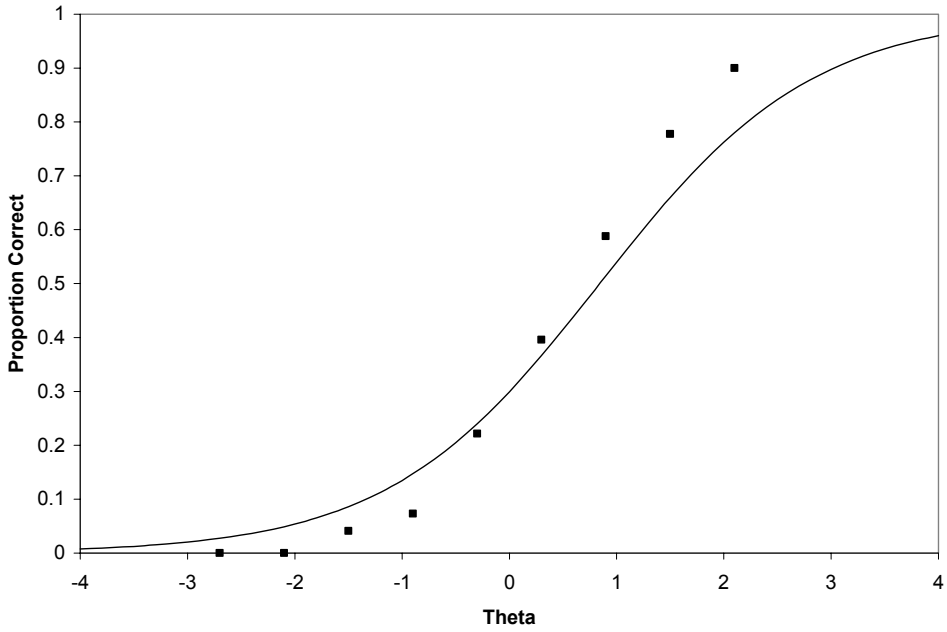


Figure 5-14: Observed and predicted proportion correct for item P2.

5.3 Further analysis

Further analysis of the SCI data is planned in three key areas including factor analysis, test bias, and confidence analysis. An investigation of the factor structure of the instrument will provide additional information on the dimensionality of the SCI and will aid in structuring the score reporting. This type of analysis will also provide evidence for the validity of the instrument. The second area includes methods for detecting possible test or item bias and is an important component of establishing test validity. The third area involves an analysis of examinee confidence in their answers to items. This type of analysis will yield deeper insight into item behavior and may help to distinguish between misconceptions and guessing.

5.3.1 Factor Analysis

Full information maximum likelihood non-linear factor analysis was carried out with the fall 2003 post test data using the TESTFACT analysis software (Wood 2003). Questions that had been eliminated based on content and item analysis considerations were omitted from the factor analysis. The sample size was 280. A single factor model accounted for 17.7% of the total variance. Only three of the items had a negative factor loading. This would indicate that nearly all questions exhibited a positive loading on the single factor.

In addition, a bifactor model was fit. The bifactor analysis assumes the presence of a general factor and additional group factors. All items are expected to load positively on the general factor. The group factors are comprised of subsets of items, each item may be assigned to only one group. For this model, the SCI questions were divided into groups based on topical content: probability, descriptive, inferential, and graphical. For

the SCI, the model includes a general statistics factor and the four more specialized group factors. This factor structure accounted for almost 30% of the total variance, the general factor accounted for 19.8%.

These findings indicate the presence of considerable unique variance among the items. This is not surprising however, since the items cover a broad range of topics within the statistics content domain, with few concepts repeated across multiple items. This also suggests that reporting subscores for the topic areas in addition to a total score may be a reasonable approach. This needs to be investigated further. The sample size for this original analysis is relatively small, and the instrument has undergone many changes since the fall of 2003. Determining whether this model still fits the current data or whether other models are more representative is an important area for further research. Further results can be found in Allen (2006).

Factor analysis can be an important tool for evaluating and interpreting any testing instrument. It can provide evidence for the validity of score interpretation. It provides anther method for assessing instrument reliability: coefficient omega can be determined once item factor loadings are known. In addition, this type of analysis can provide further evidence for assessing whether dimensionality assumptions are met for both classical test theory and item response theory.

5.3.2 Investigation of Test Bias

Another important issue that should be considered is that of test bias or fairness. Item response theory provides effective tools for understanding and evaluating test bias at the item level. It item is said to exhibit differential item functioning (DIF) “if individuals having the same ability, but from different groups, do not have the same probability of

getting the item right" (Hambleton, et al. 1991 p. 110). Since the IRT item characteristic curve shows the probability of success for a given ability level, it provides a mechanism for evaluating DIF. An item which exhibits no DIF would have identical ICCs over all subgroups. Estimating the item parameters separately for each subgroup of interest and comparing the resulting ICCs gives an effective method for detecting DIF, and thus bias, within the test instrument.

There are key advantages to using the IRT framework to investigate DIF. First, it differentiates between cases where differences in performance are due to DIF (differences in the probability of success for people of the same ability level) and cases where differences are due to actual between-group differences in ability (differences in the mean ability levels for each group, but the probability of success at a given ability level is the same for each group). Secondly, when an item exhibits DIF, comparing the ICCs for each group can reveal whether the DIF is uniform across all ability levels, that is, the probability of success is higher for one group across the entire ability range, or non-uniform, that is the probability of success is greater for one group at one end of the ability range and another at the other end of the ability range.

The major disadvantage of this method is once again the large sample size that is required to obtain good parameter estimates. A sufficiently large sample size must be available for *each* subgroup. To date the SCI data set is not large enough to perform DIF analysis. Soon, we expect to be able to conduct DIF analysis for gender. Another interesting analysis would be to consider the subgroup of the population comprised of science, engineering, and mathematics majors and the subgroup comprised of other majors. Performance differences have been observed between some courses taken by

predominately science, engineering and mathematics majors and those taken predominately by other majors (see section 2.3.1). The non-engineering courses are typically taken by younger students with less mathematics instruction. It is not clear however whether these differences are due to actual between-group differences in ability. Since the SCI was written with an engineering population in mind, this question remains relevant as the instrument is used within a broader population.

5.3.3 Confidence Analysis

Administration of the SCI is currently shifting to an optional online, web-based format. This allows for easier administration and data collection at local and distant sites, and permits additional flexibility in course schedules. With this implementation, it has been possible to add an additional component to each question to assess how confident examinees are in their answers. After answering each question, examinees are asked to rank their confidence on a scale from 1 ("Not confident at all") to 4 ("Very confident").

Some preliminary data from Fall 2005 indicates that this information may be useful in identifying true misconceptions from questions that students simply do not know, as well as for identifying questions for which many students are guessing. Figure 5-15 shows the distribution of confidence levels for each answer for question P1 (#1). The correct answer is (c), but choice (e) is clearly popular and chosen with a high level of confidence. This information along with the information provided by the nominal response model in chapter 4 indicates that this is a probable misconception held by many students. More confidence analysis will be presented in Allen (2006).

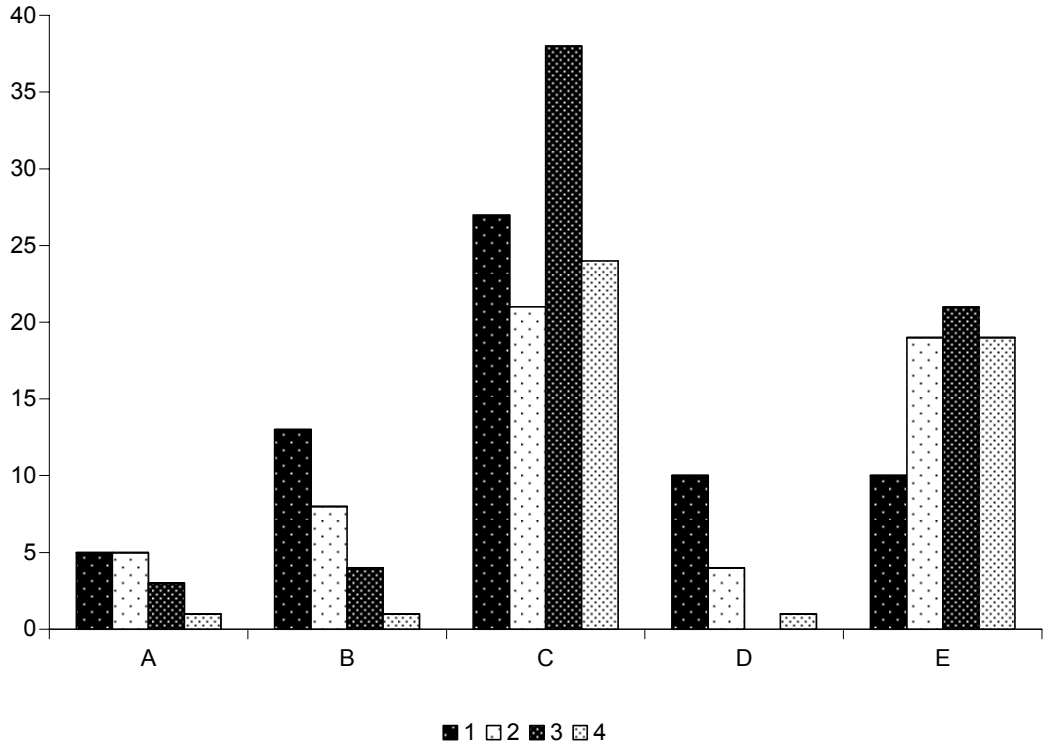


Figure 5-15: Frequency of confidence ranking for each response to question P1 (#1). Confidence was ranked on a scale from 1 ("Not confident at all") to 4 ("Very confident").

5.4 Future Revisions

The main goal for further development of the SCI should be to make revisions that reduce the measurement error. The classical test theory data assembled in chapter three can be combined with the IRT perspective introduced in chapter four. Items that have poor discrimination indices and correspondingly low slope parameter estimates are candidates for revision or elimination. Improving these questions will help to reduce the measurement error and thus increase the reliability of the SCI.

Items should also be reevaluated to ensure that they focus on concepts and do not simply require knowledge of a definition to answer. Some of the questions still seem to require more recall than conceptual understanding. Items I4 (#18) and I5 (#19) may

benefit from this type of evaluation. Very difficult questions can also be considered for revision. Two of the current items on the SCI, I4A (#13) and G2 (#15), have estimated threshold parameters greater than 3. This would indicate extremely difficult questions. Are these questions difficult because students are not covering the content during instruction or because they do not understand the instruction? Are the topics important enough to be retained? Could a different item assess the concept better? These items should be considered in this way. The topic list should be reevaluated to ensure that topic coverage is being maintained as revisions are made. For example, there are currently no questions on error or regression. This process is expected to be ongoing as the instrument continues to evolve.

One method that could be very useful for improving the SCI considerably is to introduce a relatively large number of new questions and pilot them to see how they compare to items currently on the SCI. In the past this has been very difficult to do. The length of the test cannot be increased significantly due to time constraints and the need to be able to administer the SCI within one class period. Generally new questions have been introduced only as others have been eliminated. Then, incorporating new questions made comparisons between versions of the SCI difficult.

These difficulties can be overcome by employing IRT methods. For example, it is now possible to select a core set of "best" questions from those in the current version of the SCI. Items can be selected that are more highly discriminating (higher slope parameter values) since these questions contribute to higher test information and thus lower error. Items should also be selected to maintain topic coverage and a variety of difficulty levels (threshold parameter values). This core set of questions can then be

administered with a new set of pilot questions. The known item parameter estimates from the core set of questions can be used calibrate the item parameter estimates for the new questions. The new questions can thus be equated with all of the items currently on the SCI. Theta estimates for examinees can be used to compare examinee performance and item parameter estimates can be used to compare questions. Superior items can then be retained.

5.5 Reliability and Validity

No instrument discussion is complete without an assessment of its reliability and validity. Reliability is most often measured using coefficient alpha, which gives a lower bound for reliability. Since coefficient alpha is dependent on the sample of examinees, reliability estimates have varied by semester and by class. In general, the coefficient alpha estimate for the SCI is around 0.7, higher reliability estimates are usually obtained in courses taken predominately by engineering majors (Allen, Stone, Rhoads and Murphy 2004, Allen 2006).

Other reliability measures can also be used. Coefficient omega is considered to be a better estimate of reliability, especially when a test is not strictly unidimensional (McDonald 1999). When test items do not have equal covariances and equal factor loadings, coefficient omega is strictly greater than coefficient alpha. Once a factor analysis has been completed and item factor have been established, coefficient omega can be obtained.

A third measure of reliability can be obtained through the item response theory framework. The test information curve can be used to determine the standard error of measurement for the test across the ability distribution. The standard error is not constant,

but is a function of theta. This is one of the advantages of IRT, but it does not give a single estimate of reliability. In order to obtain a single reliability estimate, the error must be averaged across the ability distribution. The reliability estimate obtained for the SCI from IRT methods is 0.787. This estimate is important because it is *not* sample dependent and is based on all of the data, not single administrations. For these reasons, it should be considered the best reliability estimate that is available for the SCI. While higher reliability is always desirable, for the current purposes of the SCI, this reliability is adequate (Nunnally 1967).

As discussed in Chapter 1 (section 1.5.5), addressing test validity is a very important part of the test construction process. The validity of a test must be established before meaningful claims can be made about test outcomes. Reliability is a necessary component of test validity, but it is not sufficient evidence of validity. Neither is there a single measure of validity. Instead, evidence of validity must be accumulated from multiple sources and the process of establishing validity is ongoing. Messick (1989) discusses three components of construct validity which should be addressed when assessing validity claims: a substantive component, a structural component, and an external component.

The structural component focuses on the test at the item level and addresses the inclusion of items based on topical content and psychometric analyses. In constructing the SCI, the domain was initially specified as statistics and probability topics. Important topics were chosen based on input from instructors of statistics in the engineering department at the University of Oklahoma. In addition, the Advanced Placement topics list was also consulted in order to obtain a broader perspective on the introductory

statistics curriculum (College Entrance Examination Board 2001). Items were written with a focus on the conceptual nature of the topics rather than on problem solving or computation. However, initially items were included that were very recall oriented or numerically focused. The test has been administered several times and student focus groups have been used to discuss the individual items. Based on this feedback, items that contained confusing wording or that examinees indicated they were answering based on reasoning other than the intended concept have been eliminated or rewritten.

We have worked to remove or rewrite items that appear to function on the level of definition recall. In addition, analyses of the data generated from these administrations have been carried out. Item analysis and response distribution analysis has been conducted for all items to determine item difficulty, discrimination, and the effect on test reliability. Based on this type of analysis, items that have had poor psychometric properties have been revised and deleted. Additionally, item response theory modeling has been conducted and this information has been paired with classical test statistics to identify questions which would benefit from further revision.

A factor analysis of the data from the Fall 2003 administration was conducted using a model that included a general factor and specific group factors. For this analysis the items were assigned to one of four groups based on their topical content. The four subgroups were probability, descriptive statistics, inferential statistics, and graphical methods. Further factor analysis will be carried out with a larger sample size for the current version of the instrument to ascertain whether the questions work together in groups as expected.

The structural component of construct validity encompasses the scoring procedures and reporting format. The scoring model for the SCI is a cumulatively scored criterion measure. Currently scores are reported as the total percent correct. Factor loadings that were determined by the 2PL item response theory analysis were included in Table 4-2. These factor loadings were all positive, indicating that it is reasonable to consider a total score.

The statistics content does not appear to be strictly unidimensional and we would like to explore reporting sub-scores in addition to total correct scores. Questions have been grouped into sub-groups of the content domain as it is generally encountered in instructional methods: probability, descriptive statistics, inferential statistics, and graphical methods. As more data are collected, we will refine these groupings and explore alternative groupings of items for sub-score reporting.

The external component involves how the test relates to other variables. A valid instrument should perform in predictable ways to other measures of test and non-test behaviors. There is not another measure that is comparable to the SCI. Other measures of statistics knowledge rely heavily on problem solving and computation skills. The SCI has been analyzed for correlation with final course scores. So far, the correlation has been low. This is as expected since we believe course grades are generally largely a measure of problem solving and may include components of other classroom behavior such as attendance. We believe that the problems on the SCI are somewhat novel in nature to the introductory statistics student as they are outside of the traditional textbook fare. There may be higher correlations with project evaluations or possibly final test scores since they

generally are a broader measure of course content. Assessment of validity will be an ongoing activity as more data are collected and further revisions are made.

5.6 Conclusions

The SCI is a unique instrument for evaluating statistics understanding. There is no other instrument currently available which focuses on conceptual understanding and which covers the scope of a typical introductory statistics course. It has been demonstrated to be a reasonably reliable instrument for research use. The SCI should be used in classroom settings as a posttest and optionally as a pretest for the purposes of evaluating instructional methods. Baseline data is available that can be used as a benchmark for comparison.

It is hoped that instructors find that the content on the SCI corresponds to what they expect their students to have mastered upon leaving the introductory statistics course. As such, the SCI can fulfill the role that other concept inventories have in initiating widespread interest in instructional research and innovations for statistics within the classroom setting.

References

Aliaga, M., Cobb, G. W., Cuff, C., Garfield, J., Gould, R., Lock, R., et al. (2005). Guidelines for assessment and instruction in statistics education (GAISE) project college report. Retrieved April 14, 2005, from http://it.stlawu.edu/~rlock/gaise/

Allen, K., Stone, A., Rhoads, T. R., & Murphy, T. J. (2004). *The statistics concept inventory: Developing a valid and reliable instrument.* Paper presented at the American Society for Engineering Education Annual Conference and Exposition, Salt Lake City, UT.

Allen, K. C. (2006), "*The Statistic Concept Inventory: Development and Analysis of a* Cognitive *Assessment Instrument in Statistics*," Unpublished Doctoral Dissertation, University of Oklahoma.

Bar-Hillel, M. (1982). Studies of representativeness. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 69-83). Cambridge: Cambridge University Press.

Beichner, R. J. (2004). Testing student interpretation of kinematics graphs. *American Journal of Physics, 62*(8), 750-762.

Ben-Zvi, D., & Garfield, J. (2004). Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In D. Ben_Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 3-15): Kluwer.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29-51.

Chance, B., delMas, R., & Garfield, J. (2004). Reasoning about sampling distributions. In D. Ben_Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 295-323). Dordrecht: Kluwer.

Cobb, G. W. (1993). Reconsidering statistics education: A national science foundation conference, *Journal of Statisitcs Education* (Vol. 1).

College Entrance Examination Board. (2001). Advanced placement program course description: Statistics. Retrieved September 10, 2002, 2002, from http://apcentral.collegeboard.com/repository/ap01.cd_sta_4328.pdf

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

Crouch, C., & Mazur, E. (2001). Peer instruction: Ten years of experience and results. *American Journal of Physics, 69*(9), 970-977.

Embretson, S. E., and Reise, S. P. (2000), *Item Response Theory for Psychologists*, Mahwah, NJ: Lawrence Erlbaum.

Engelhardt, P. V., & Beichner, R. J. (2004). Students' understanding of direct current resistive electrical circuits. *American Journal of Physics, 72*(1), 98-115.

Engineering Accreditation Commission. (2003). 2004-2005 criteria for accrediting engineering programs. Retrieved April 12, 2005, from http://www.abet.org/criteria_eac.html

Evans, D. L., Gray, G. L., Krause, S., Martin, J., Midkiff, C., Notaros, B. M., et al. (2003, November 5-8). *Progress on concept inventory assessment tools.* Paper presented at the 33rd ASEE/IEEE Frontiers in Education Conference, Boulder, CO.

Fischbein, E., & Schnarch, D. (1997). The evolution with age of probabilistic, intuitively based misconceptions. *Journal for Research in Mathematics Education, 28*(1), 96-105.

Foundation Coalition. (2001, 2/22/2005). Foundation coalition key components: Concept inventories. Retrieved February 22, 2005, from http://www.foundationcoalition.org/home/keycomponents/concept/index.html

Gal, I., & Garfield, J. (1997a). Curricular goals and assessment challenges in statistics education. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 1-16). Amsterdam: IOS Press.

Gal, I., & Garfield, J. (Eds.). (1997b). *The assessment challenge in statistics education*. Amsterdam: IOS Press.

Garfield, J. (2002). The challenge of developing statistical reasoning, *Journal of Statistics Education* (Vol. 10).

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

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

Garfield, J., Hogg, B., Schau, C., & Whittinghill, D. (2002). First courses in statistical science: The status of educational reform efforts, *Journal of Statistics Education* (Vol. 10).

Gibb, B. G. (1964). *Test-wiseness as secondary cue response.* Stanford University.

Gray, G. L., Costanzo, F., Evans, D., Cornwell, P., Self, B., & Lane, J. (2005). *The dynamics concept inventory assessment test: A progress report and some results.* Paper presented at the 2005 American Society for Engineering Education Annual Conference & Exposition.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*, 255-282.

Hake, R. R. (1987). Promoting student crossover to the Newtonian world. *American Journal of Physics, 55*(10), 878-883.

Hake, R. R. (1992). Socratic pedagogy in the introductory physics laboratory. *The Physics Teacher, 30*, 546-552.

Hake, R. R. (1998a). Interactive-engagement versus traditional methods: A six-thousand-student

survey of mechanics test data for introductory physics courses. *American Journal of Physics, 66*(1), 64-74.

Hake, R. R. (1998b). Interactive-engagement vs traditional methods: A six-thousand student survey of mechanics test data for introductory physics courses. *American Journal of Physics, 66*(1), 64-75.

Halloun, I., & Hestenes, D. (1985a). Common sense concepts about motion. *American Journal of Physics, 53*(11), 1056-1065.

Halloun, I., & Hestenes, D. (1985b). The initial knowledge state of college physics students. *American Journal of Physics, 53*(11), 1043-1055.

Halloun, I. A., & Hestenes, D. (1985c). The initial knowledge state of college physics students. *American Journal of Physics, 53*(11), 1043-1055.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*: Kluwer Nijhoff.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Herman, W. E. (1997). Statistical content errors for students in an educational psychology course, *Paper presented at the Annual Meeting of the American Psychological Association, Chicago*.

Hestenes, D. (1987). Toward a modeling theory of physics instruction. *American Journal of Physics, 55*(5), 440-454.

Hestenes, D., & Halloun, I. (1995). Interpreting the force concept inventory: A response to March 1995 critique by Huffman and Heller. *Physics Teacher, 33*(8), 502,504-506.

Hestenes, D., & Wells, M. (1992). A mechanics baseline test. *The Physics Teacher, 30*, 159-166.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher, 30*(3), 141-158.

Jacobi, A., Martin, J., Mitchell, J., & Newell, T. (2003). A concept inventory for heat transfer, *33rd ASEE/IEEE Frontiers in Education Conference*. Boulder, CO.

Jacobi, A., Martin, J., Mitchell, J., & Newell, T. (2004). Work in progress: A concept inventory for heat transfer, *34th ASEE/IEEE Frontiers in Education Conference*. Savannah, GA.

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

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*(3), 430-453.

Kahneman, D., & Tversky, A. (1982). Subjective probability: A judgment of representativeness. In D. Kahneman, P. slovic & A. Tversky (Eds.), *Judgment under*

uncertainty: Heuristics and biases (pp. 32-47). Cambridge: Cambridge University Press.

Kelly, T. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology, 30*, 17-24.

Kirisci, L., Hsu, T.-c., and Yu, L. (2001), "Robustness of Item Parameter Estimation Programs to Assumptions of Unidimensionality and Normality," *Applied Psychological Measurement*, 25(2), 146-162.

Knight, R. D. (2002). *Five easy lessons: Strategies for successful physics teaching*. San Francisco: Pearson Education.

Kolan, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices*. New York: Springer-Verlag.

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* (Vol. 3).

Konold, C., Pollatsek, A., Well, A., Lohmeier, J., & Lipson, A. (1993). Inconsistencies in students' reasoning about probability. *Journal for Research in Mathematics Education, 24*(5), 392-414.

Krause, S., Birk, J., Bauer, R., Jenkins, B., & Pavelich, M. (2004). *Development, testing, and application of a chemistry concept inventory.* Paper presented at the ASEE/IEEE Frontiers in Education Conference, Savannah, GA.

Krause, S., Decker, J. C., & Griffin, R. (2003, November 5-8). *Using a materials concept inventory to assess conceptual gain in introductory materials engineering courses.*

Paper presented at the 33rd ASEE/IEEE Frontiers in Education Conference, Boulder, CO.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*(3), 151-160.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3* (Monograph Supp 9), 635-694.

Loftsgaarden, D. O., & Watkins, A. E. (1998). Statistics teaching in colleges and universities: Courses, instructors, and degrees in fall 1995. *The American Statistician, 52*(4), 308-314.

Lord, F. M. (1953). An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika, 18*, 57-75.

Maloney, D. P., O'Kuma, T. L., Hieggelke, C. J., & Alan, V. H. (2001). Surveying students' conceptual knowledge of electricity and magnetism. *Physics Education Research, American Journal of Physics Supplement, 69*(7-S1), S12-S23.

Martin, J., Mitchell, J., & Newell, T. (2003, November 5-8). *Development of a concept inventory for fluid mechanics.* Paper presented at the 33rd ASEE/IEEE Frontiers in Education Conference, Boulder, CO.

Martin, J., Mitchell, J., & Newell, T. (2004). *Work in progress: Analysis of reliability of the fluid mechanics concept inventory.* Paper presented at the 34th ASEE/IEEE Frontiers in Education Conference, Savannah, GA.

Mazur, E. (1997). *Peer instruction*. Upper Saddle River, NJ: Prentice Hall.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American College on Education/Macmillan.

Mokros, J., & Russell, S. J. (1995). Children's concepts of average and representativeness. *Journal for Research in Mathematics Education, 26*(1), 20-39.

Moore, D. S. (1997). New pedagogy and new content: The case of statistics. *International Statistical Review, 65*(2), 123-137.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: National Council of Teachers of Mathematics.

National Science Foundation. (2005). US NSF - funding - research experiences for undergraduates. Retrieved July 12, 2005, 2005, from http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=5517&org=NSF

Northern Oklahoma College. (2005). Course catalog: Course descriptions. from http://www.north-ok.edu/2002/admissions/catalog_main.htm

Notaros, B. M. (2002). *Concept inventory assessment instruments in electromagnetics education.* Paper presented at the IEEE Antennas and Propagation Society International Symposium, San Antonio, TX.

Nunnally, J. C. (1967). *Psychometric theory*. New York: McGraw-Hill.

Olds, B. M., Streveler, R., & Miller, R. L. (2004, June). *Preliminary results from the development of a concept inventory in thermal and transport science.* Paper presented at the American Society for Engineering Education Annual Conference and Exposition, Salt Lake City, UT.

Pollatsek, A., Lima, S., & Well, A. D. (1981). Concept or computation: Students' understanding of the mean. *Educational Studies in Mathematics, 12*, 191-204.

Reckase, M. D. (1979), "Unifactor Latent Trait Models Applied to Multifactor Tests: Results and Implications," *Journal of Educational Statistics*, 4(3), 207-230.

Roedel, R. J., El-Ghazaly, S., Rhoads, T. R., & El-Sharawy, E. (1998, November). *The wave concepts inventory - an assessment tool for courses in electromagnetic engineering.* Paper presented at the Frontiers in Education Conference, Tempe, AZ.

Saldanha, L., & Thompson, P. (2002). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics, 51*(3), 257-270.

Saul, J. M. (1998). *Beyond problem solving: Evaluating introductory physics courses through the hidden curriculum.* Unpublished Dissertation, University of Maryland.

Schaeffer, R. L., & Stasny, E. A. (2004). The state of undergraduate education in statistics: A report from the CBMS 2000. *The American Statistician, 58*(4), 265-271.

Sedlmeier, P., & Gigerenzer, G. (1997). Intuitions about sample size: The empirical law of large numbers. *Journal of Behavioral Decision Making, 10*, 33-51.

Skromme, B. J. (2005). Spreadsheets to promote interactive engagement in semiconductor device courses. Retrieved September 20, 2005, from http://cresmet.asu.edu/proj_res/spreadsheets/Index.htm

Steif, P. (2004). *Initial data from a statics concept inventory.* Paper presented at the 2004 American Society for Engineering Education Annual Conference and Exposition.

Steif, P., Dollar, A., & Dantzler, J. A. (2005). *Results from a statics concept inventory and their relationship to other measures of performance in statics.* Paper presented at the 35th ASEE/IEEE Frontiers in Education Conference, Indianapolis, IN.

Stone, A., Allen, K., Rhoads, T. R., Murphy, T. J., Shehab, R. L., &Saha, C. (2003), *The Statistics Concept Inventory: A Pilot Study*, Paper presented at the 2003 ASEE/IEEE Frontiers in Education Conference, Boulder, CO.

Strauss, S., & Bichler, E. (1988). The development of children's concepts of the arithmetic average. *Journal for Research in Mathematics Education, 19*(1), 64-80.

Sundre, D. L. (2003). Assessment of quantitative reasoning to enhance educational quality, *Paper presented at the 2003 meeting of the American Educational Research Association, Chicago*.

Thissen, D. (2003). MULTILOG (Version 7.0.2327.3): Scientific Software International, Inc.

Thornton, R. K., & Sokoloff, D. R. (1998). Assessing student learning of Newton's laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula. *American Journal of Physics, 66*(4), 338-352.

Tobias, S., & Hake, R. R. (1988). Professors as physics students: What can they teach us. *American Journal of Physics, 56*(9), 786-794.

Tversky, A., & Kahneman, D. (1982a). Belief in the law of small numbers. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 24-31). Cambridge: Cambridge University Press.

Tversky, A., & Kahneman, D. (1982b). Judgment under uncertainty: Heuristics and biases. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 3-20). Cambridge: Cambridge University Press.

University of Oklahoma. (2005). Course catalog: Course descriptions. from http://www.ou.edu/bulletins/courses/courses.htm

University of Pittsburg. (2005). Industrial engineering: University of Pittsburgh school of engineering: Course listing. from http://www.engr.pitt.edu/industrial/pages/undergrad_courses.html

University of Virgina. (2005). Course offering directory. from http://etg08.itc.virginia.edu/cod.pages/20053/ENF/APMA.html

Wage, K., Buck, J. R., Wright, C. H. G., & Welch, T. B. (2005). The signals and systems concept inventory. *IEEE Transactions on Education, 48*(3), 448-461.

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

Watson, J. M., & Moritz, J. B. (1999). The development of concepts of average. *Focus on Learning Problems in Mathematics, 21*(4), 15-39.

Well, A. D., Pollatsek, A., & Boyce, S. J. (1990). Understanding the effects of sample size on the variability of the mean. *Organizational Behavior and Human Decision Processes, 47*(2), 289.

Wood, R. (2003). TESTFACT (Version 4.0.2328.4): Scientific Software International, Inc.

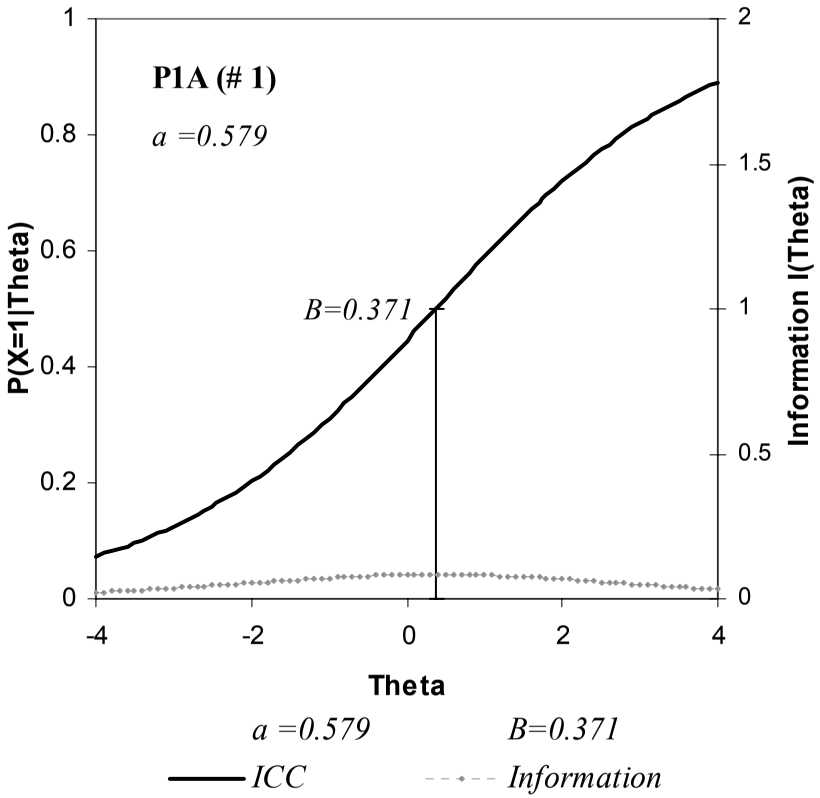
Zawojewski, J. S., & Shaughnessy, J. M. (2000). Mean and median: Are they really so easy? *Mathematics Teaching in the Middle School, 5*(7), 436-440.

Zimowsky, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). BILOG-MG (Version 3.0): Scientific Software International.

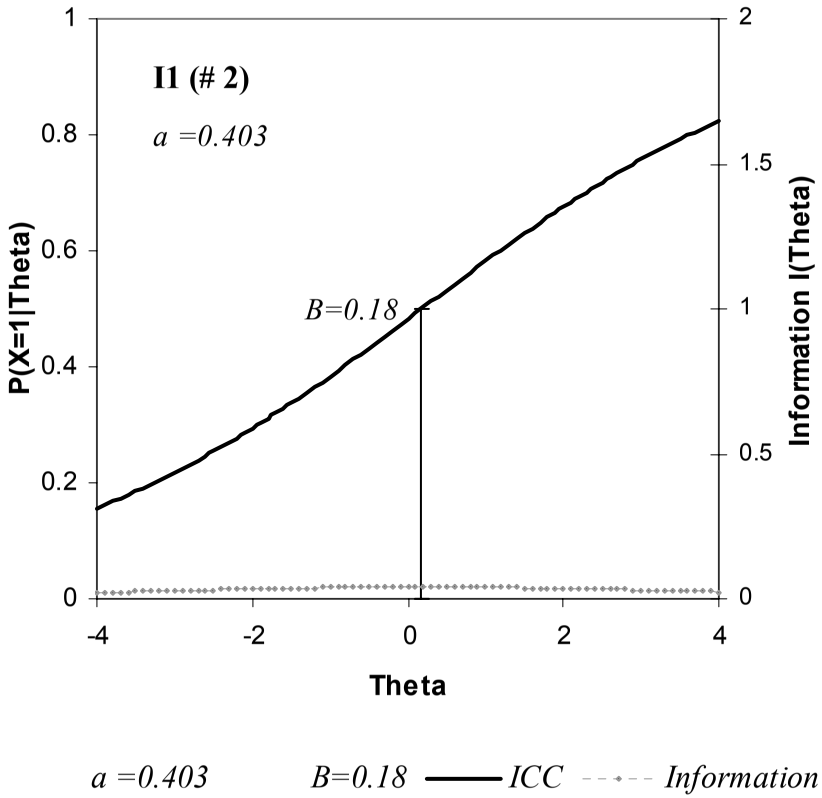
Appendix A: 2PL Model Item Characteristic Curves

The two parameter logistic model Item Characteristic Curves (ICC) and Item Information Curves for the 38 items on the current version of the SCI are included in this appendix. The numbers in parentheses on the graphs correspond to the item numbers used on the SCI and those used in chapter 3. The labels preceding the numbers (such as P1, I10a, etc) are the master numbers assigned to the items to track them historically. The initial letter in the master number assigns the question to a topic group: probability (P), descriptive statistics (D), inferential statistics (I), and graphical (G). Some master numbers include a second letter which designates the version of the question, such as P1a. Chapters 3 and 4 provide discussion on the item versions.

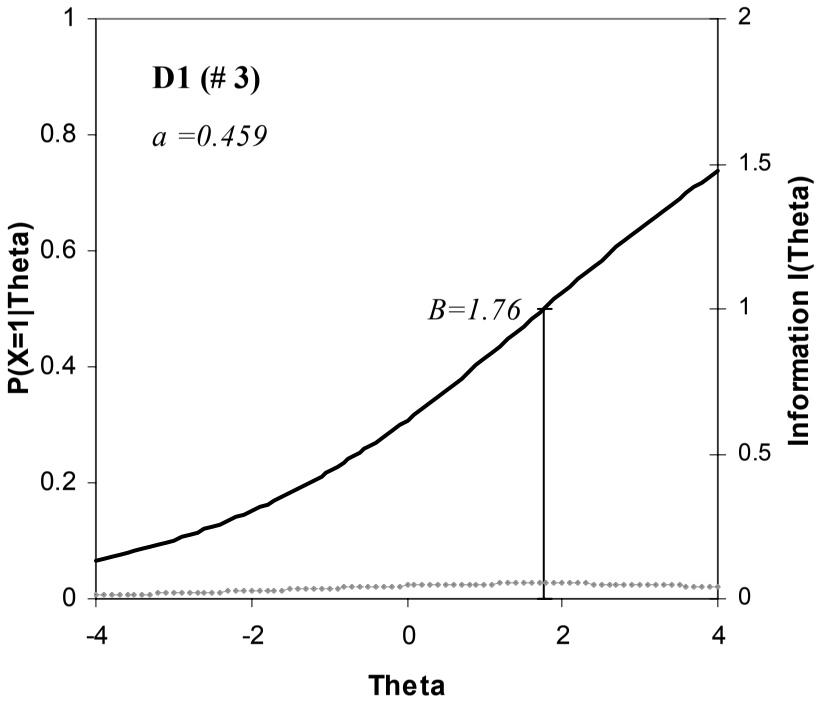
The latent trait, Θ, is assumed to be conceptual understanding of statistics and is plotted on the horizontal axis. It is assumed to have a normal distribution in the population with mean zero and standard deviation one. The ICC represents the probability of a correct response to the item for the given theta value. The probability is shown on the left vertical axis. The information function attains its maximum at $\Theta = B$ and its scale is on the right vertical axis. Higher information corresponds to less error in the estimation of theta. The item parameter estimates obtained from the two parameter logistic model are indicated on the graphs. Parameter a is the slope or discrimination parameter. Parameter B is the threshold parameter and is the value of the latent trait, Θ, for which the probability of a correct response is 0.5.



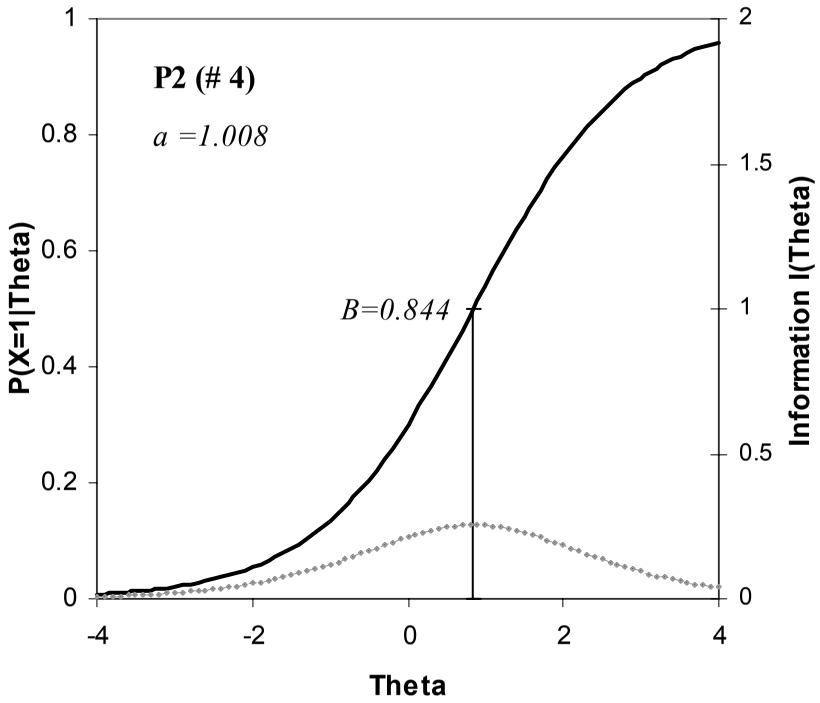
a =0.579 B=0.371 ICC Information



a =0.403 B=0.18 ICC Information

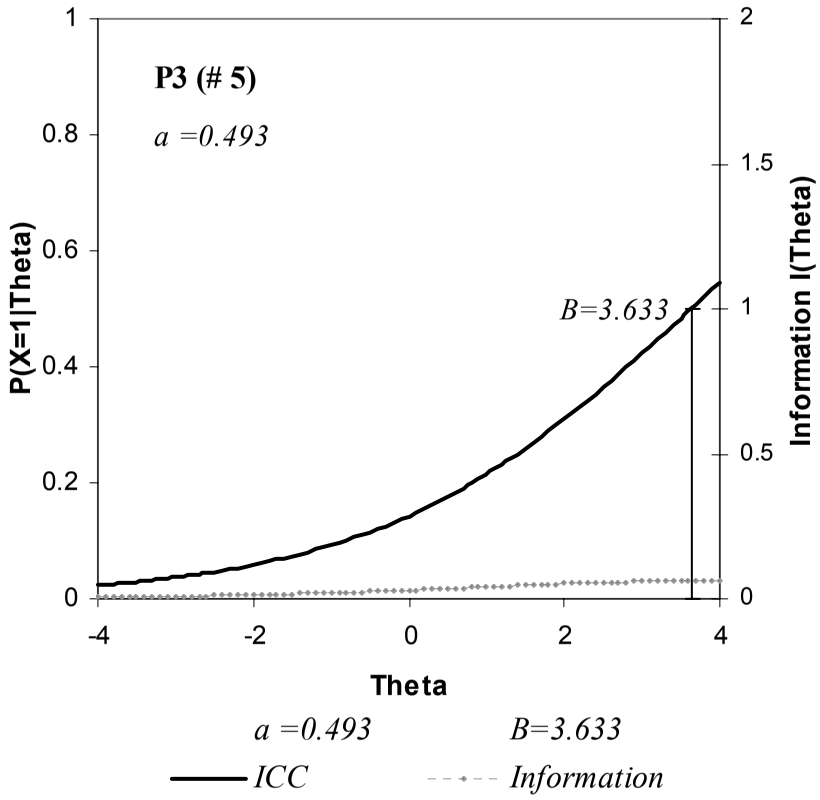


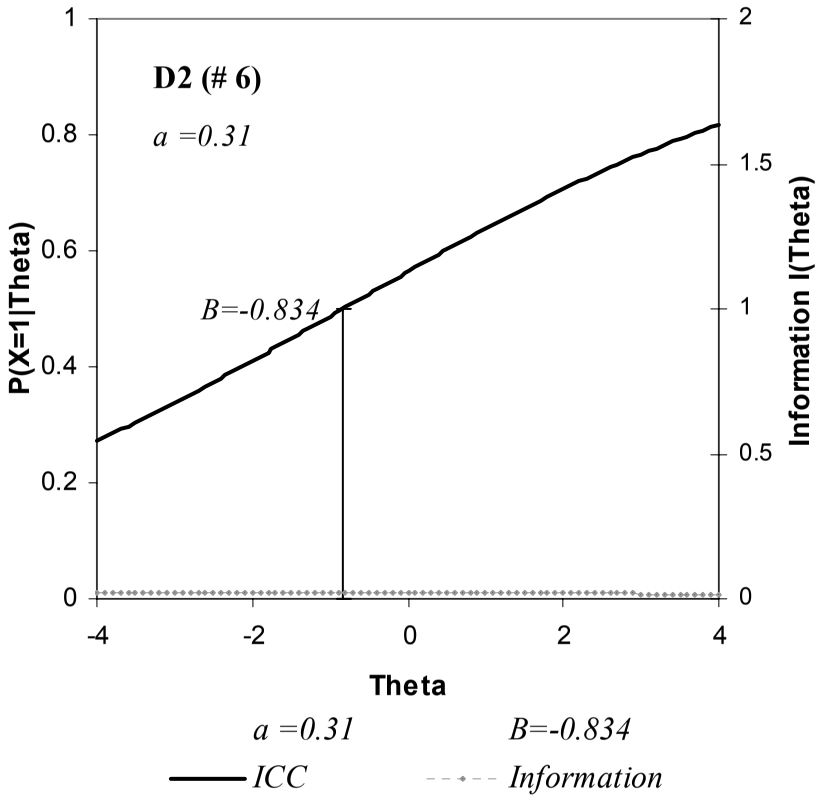
a =0.459 *B=1.76* ICC Information

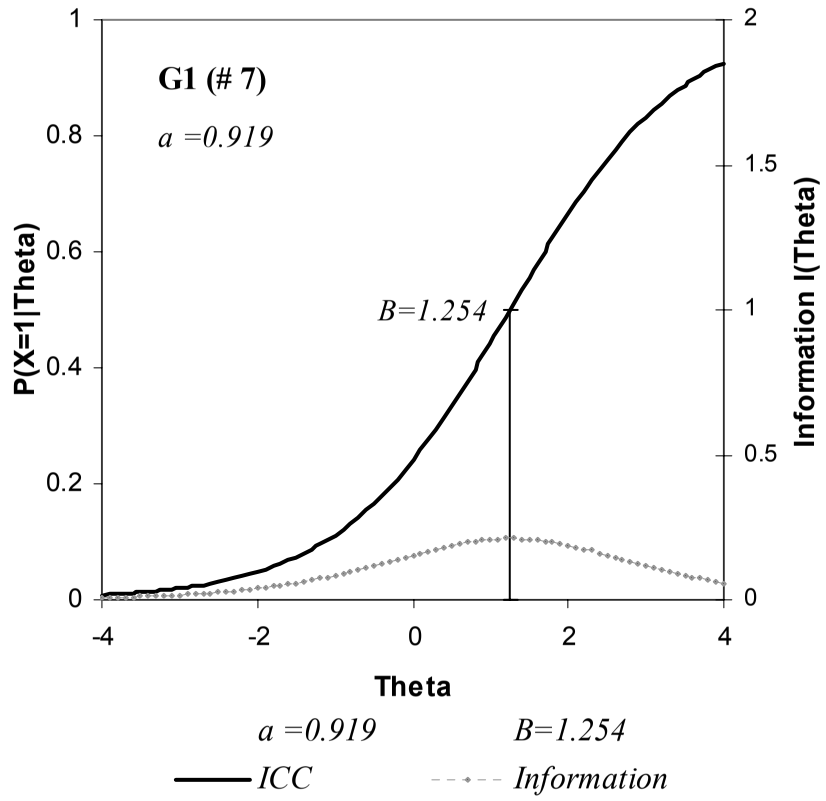


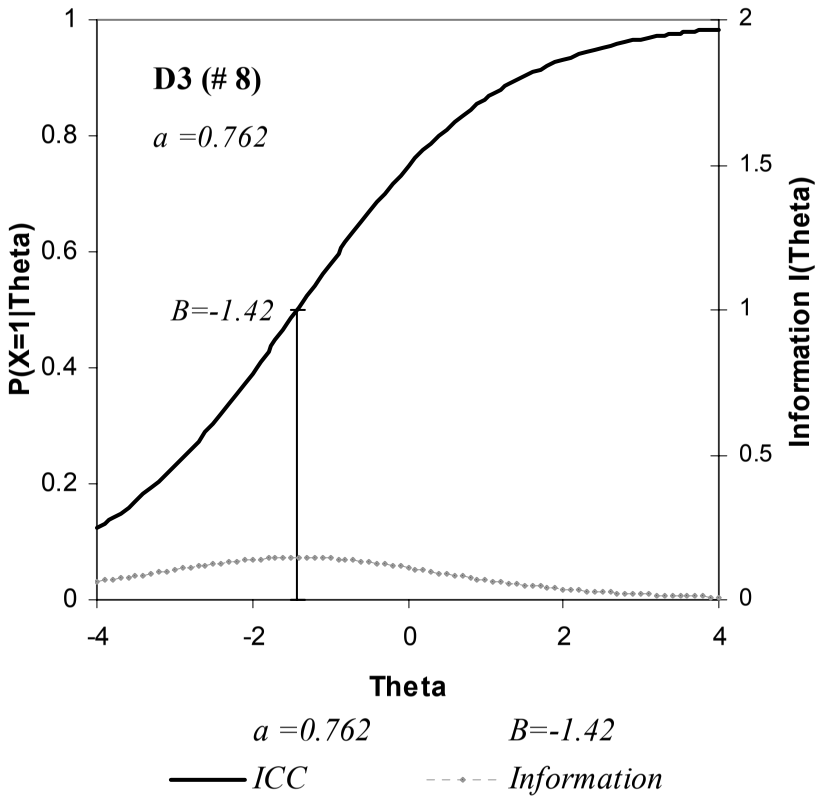
a =1.008 *B=0.844*

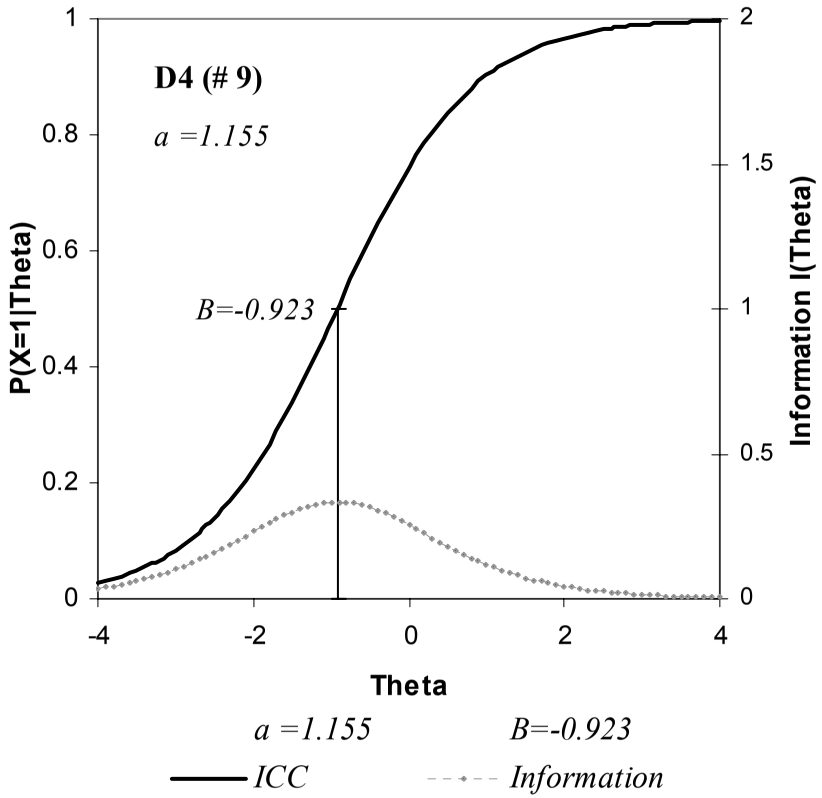
ICC Information

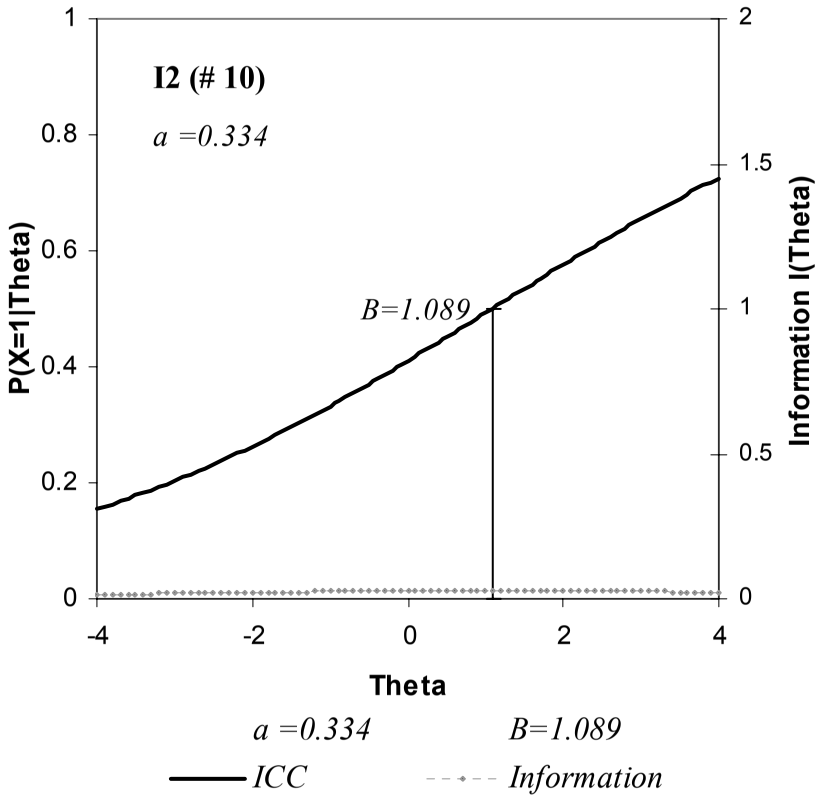

P3 (# 5)
a =0.493
B=3.633
P(X=1|Theta)
Information I(Theta)
Theta
1
0.8
0.6
0.4
0.2
0
2
1.5
1
0.5
0
-4
-2
0
2
4
a =0.493
B=3.633
ICC
Information

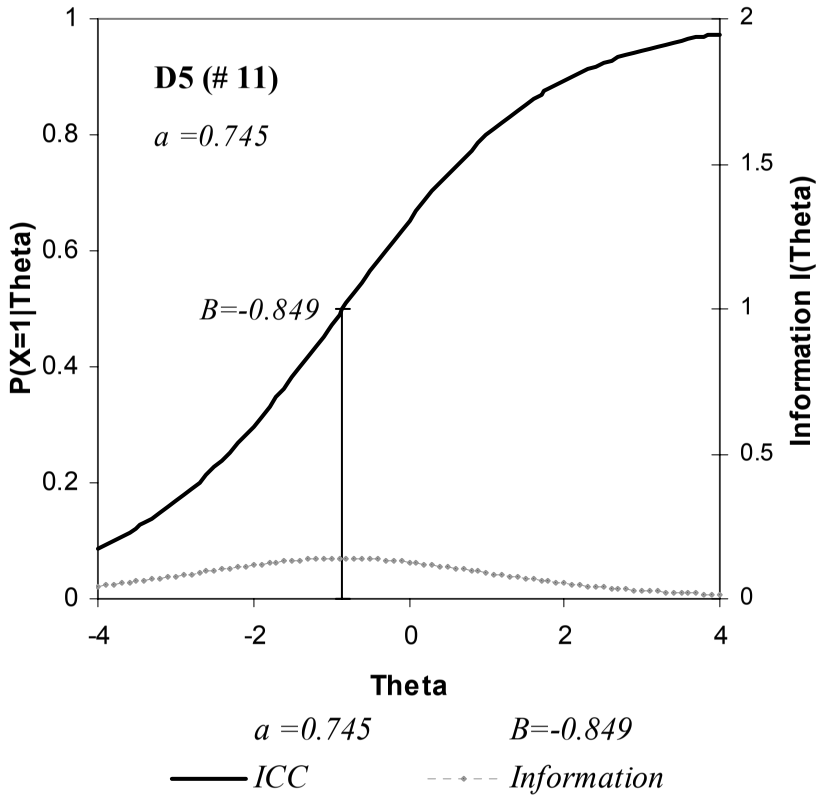

D2 (# 6)
a =0.31
B=-0.834
P(X=1|Theta)
Information I(Theta)
Theta
1
0.8
0.6
0.4
0.2
0
2
1.5
1
0.5
0
-4
-2
0
2
4
a =0.31
B=-0.834
ICC
Information

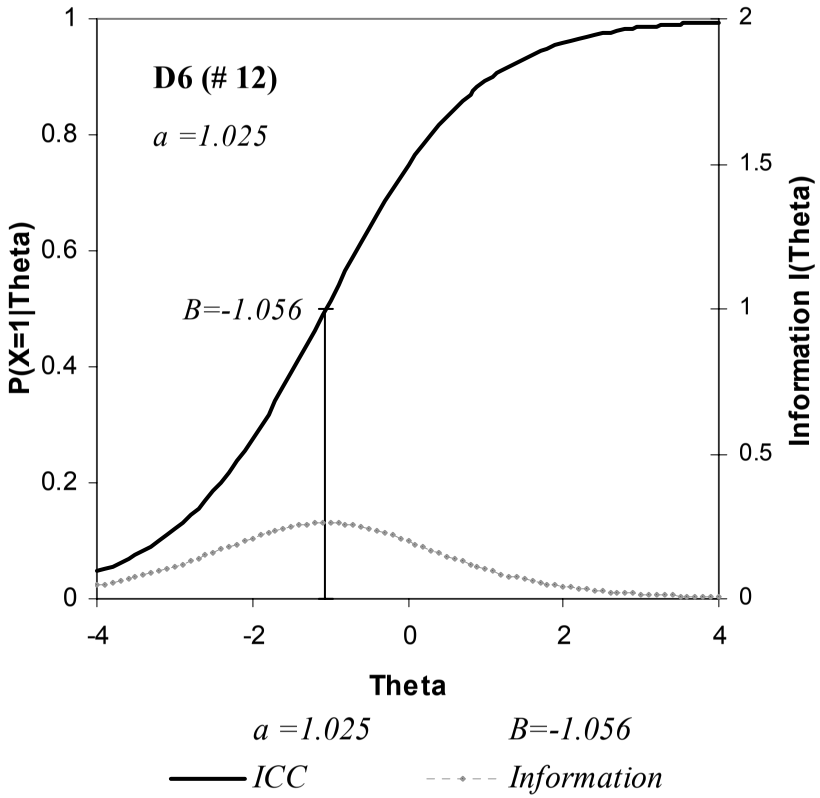

G1 (# 7)
a =0.919
B=1.254
P(X=1|Theta)
Information I(Theta)
Theta
a =0.919
B=1.254
ICC
Information

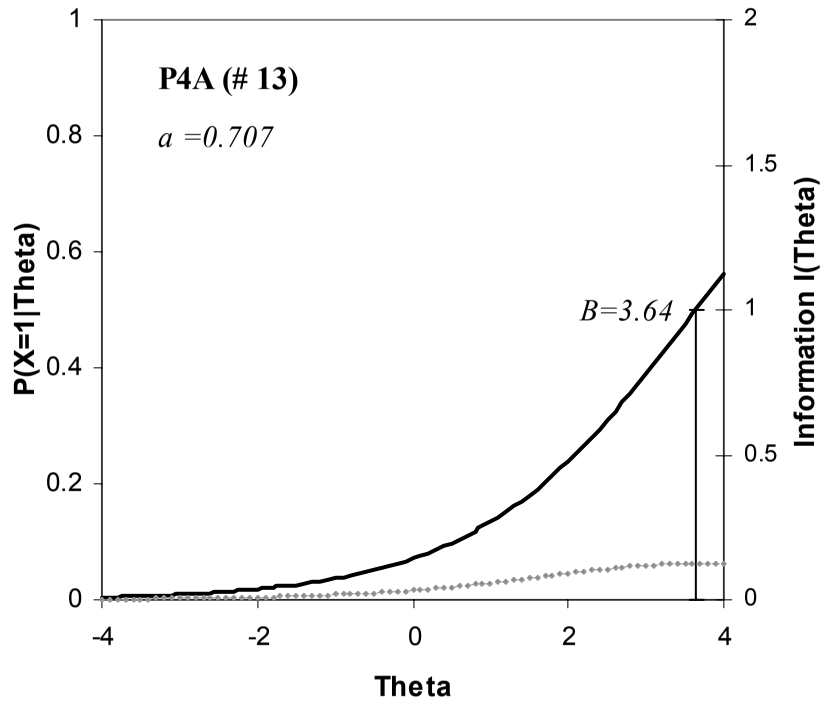

D3 (# 8)
a =0.762
B=-1.42
P(X=1|Theta)
Information I(Theta)
Theta
a =0.762
B=-1.42
ICC
Information


D4 (# 9)
a =1.155
B=-0.923
P(X=1|Theta)
Information I(Theta)
Theta
1
0.8
0.6
0.4
0.2
0
2
1.5
1
0.5
0
-4
-2
0
2
4
a =1.155
B=-0.923
ICC
Information

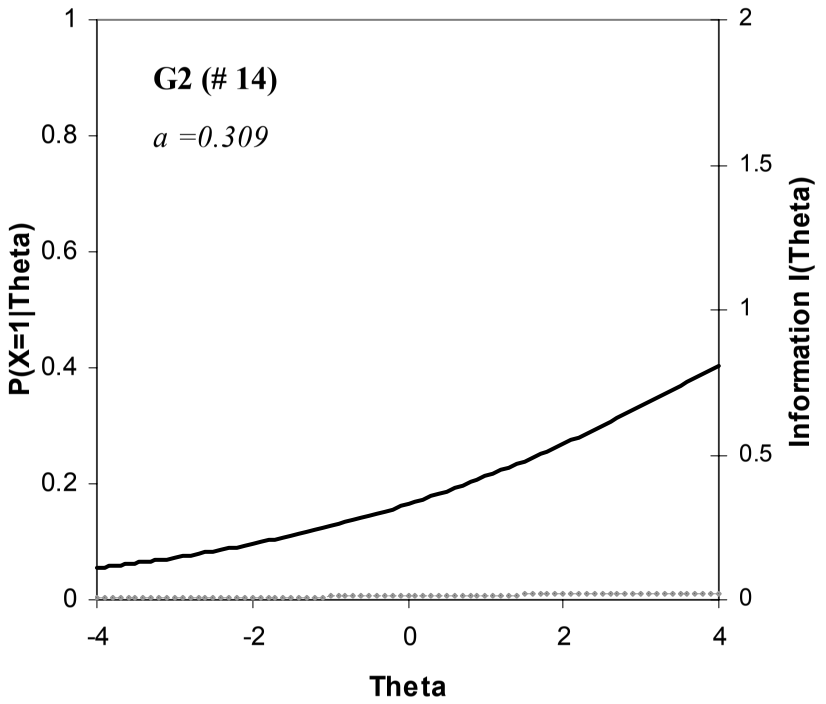

I2 (# 10)
a =0.334
B=1.089
P(X=1|Theta)
Information I(Theta)
Theta
1
0.8
0.6
0.4
0.2
0
2
1.5
1
0.5
0
-4
-2
0
2
4
a =0.334
B=1.089
ICC
Information


D5 (# 11)
a =0.745
B=-0.849
P(X=1|Theta)
Information I(Theta)
Theta
1
0.8
0.6
0.4
0.2
0
2
1.5
1
0.5
0
-4
-2
0
2
4
a =0.745
B=-0.849
ICC
Information


D6 (# 12)
a =1.025
B=-1.056
P(X=1|Theta)
Information I(Theta)
Theta
1
0.8
0.6
0.4
0.2
0
2
1.5
1
0.5
0
-4
-2
0
2
4
a =1.025
B=-1.056
ICC
Information

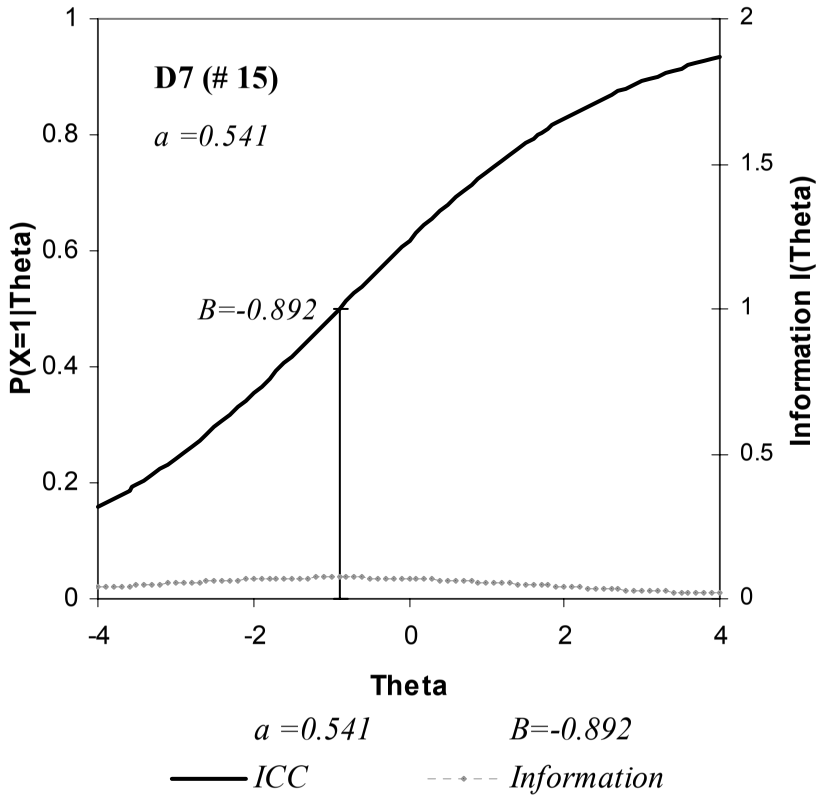


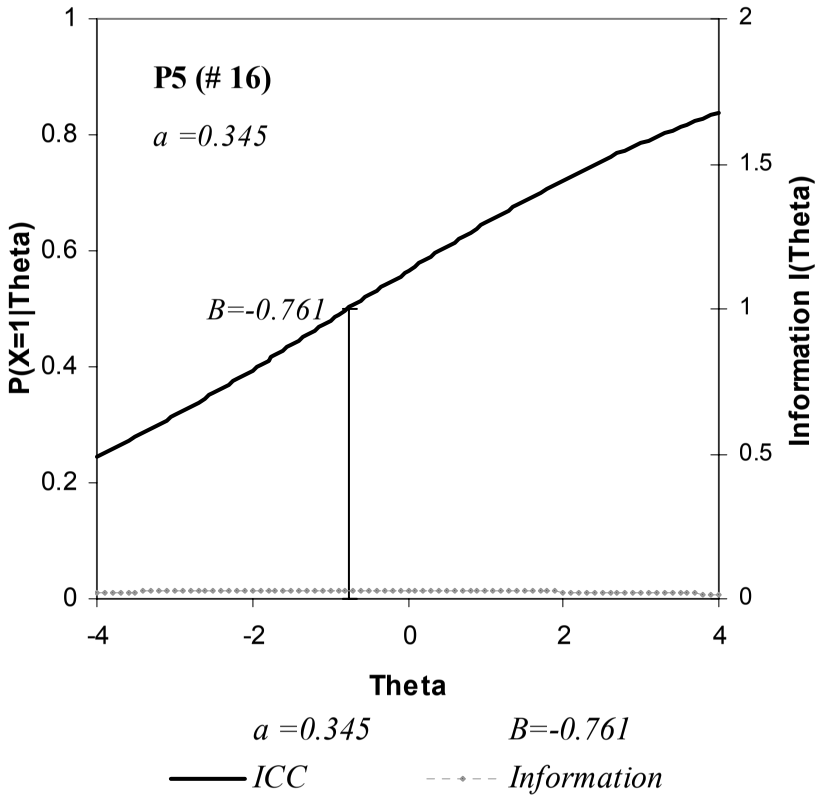
a =0.707 *B=3.64* ICC *Information*

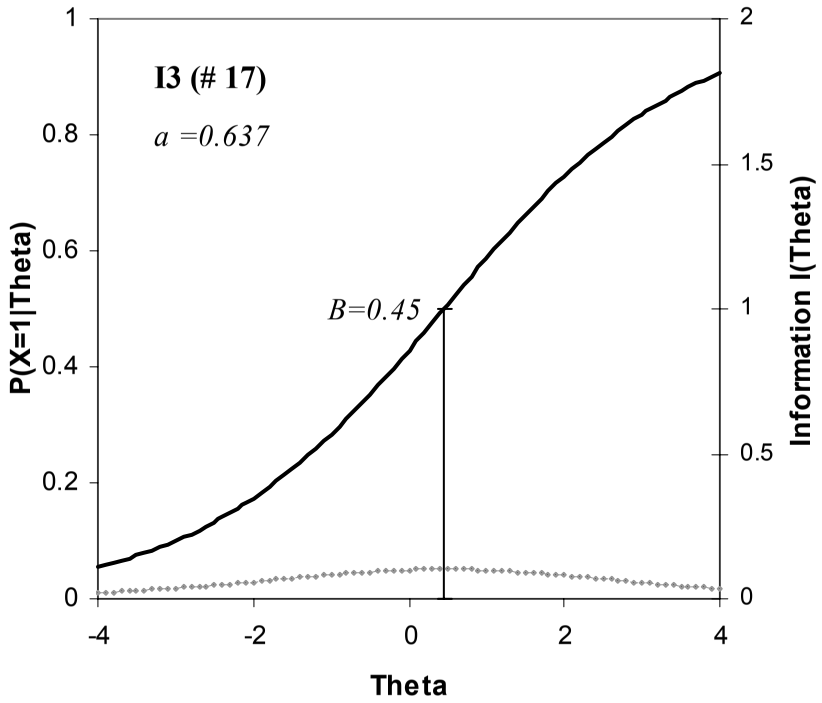


a =0.309 *B=5.243*

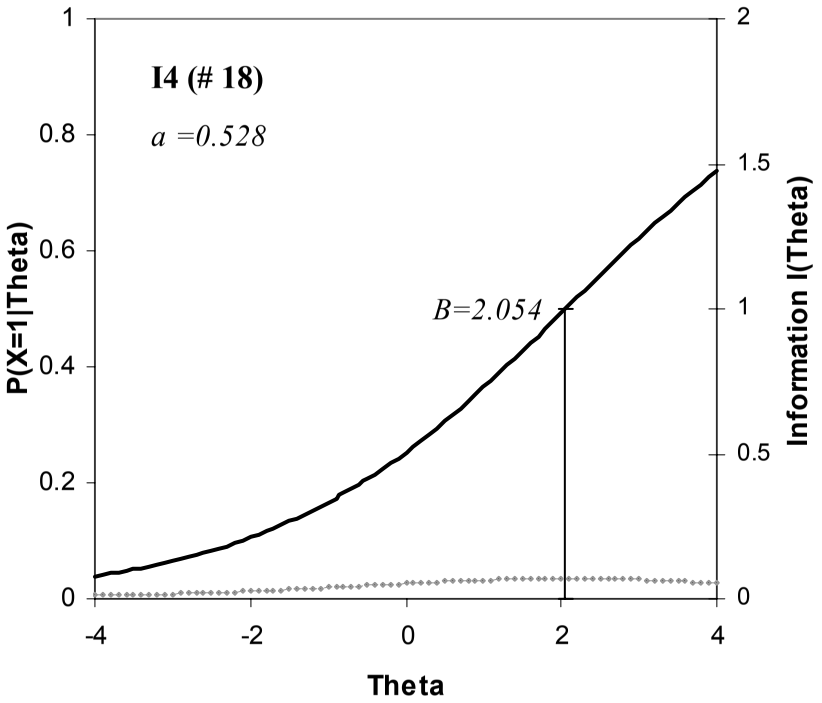
ICC *Information*


D7 (# 15)
a =0.541
B=-0.892
P(X=1|Theta)
Information I(Theta)
Theta
1
0.8
0.6
0.4
0.2
0
2
1.5
1
0.5
0
-4
-2
0
2
4
a =0.541
B=-0.892
ICC
Information


P5 (# 16)
a =0.345
B=-0.761
P(X=1|Theta)
Information I(Theta)
Theta
1
0.8
0.6
0.4
0.2
0
2
1.5
1
0.5
0
-4
-2
0
2
4
a =0.345
B=-0.761
ICC
Information

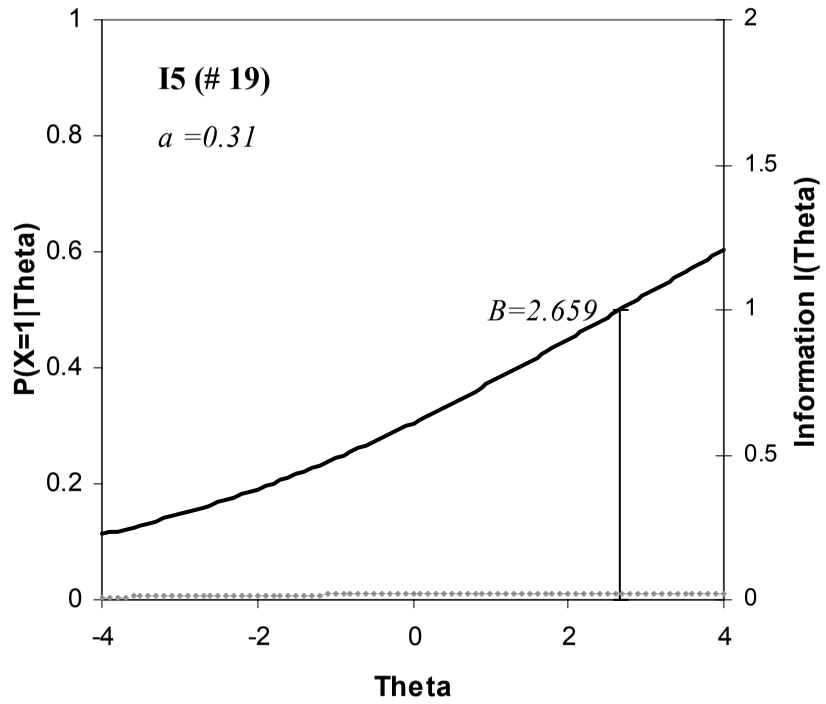


a =0.637 *B=0.45* ICC *Information*

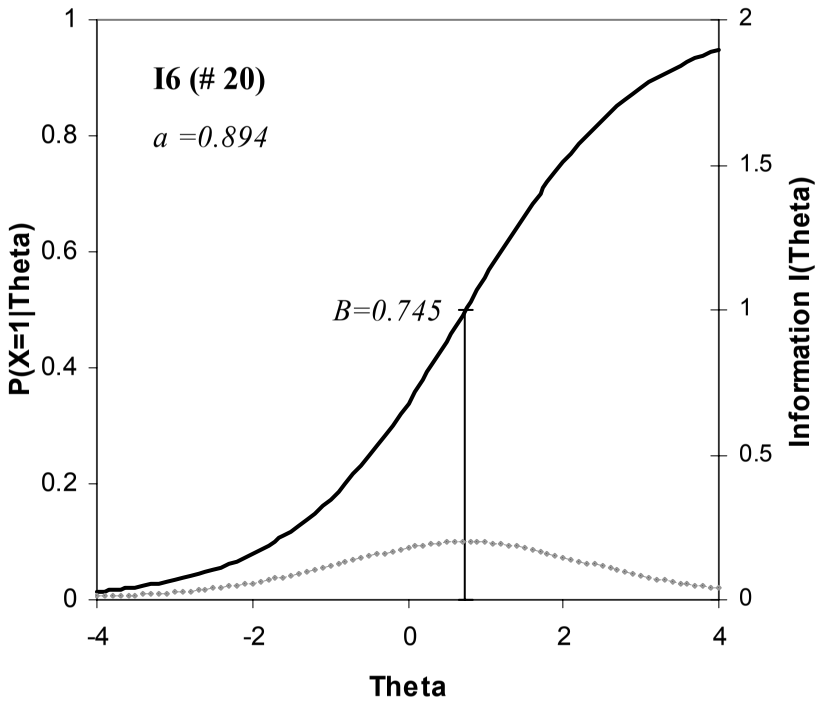


a =0.528 *B=2.054*

ICC *Information*

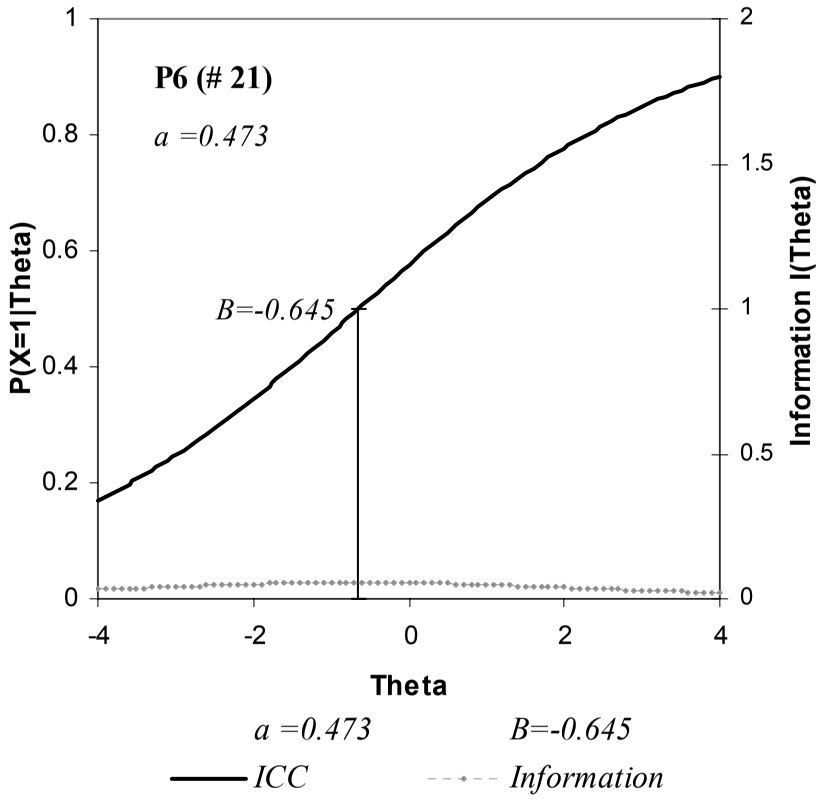


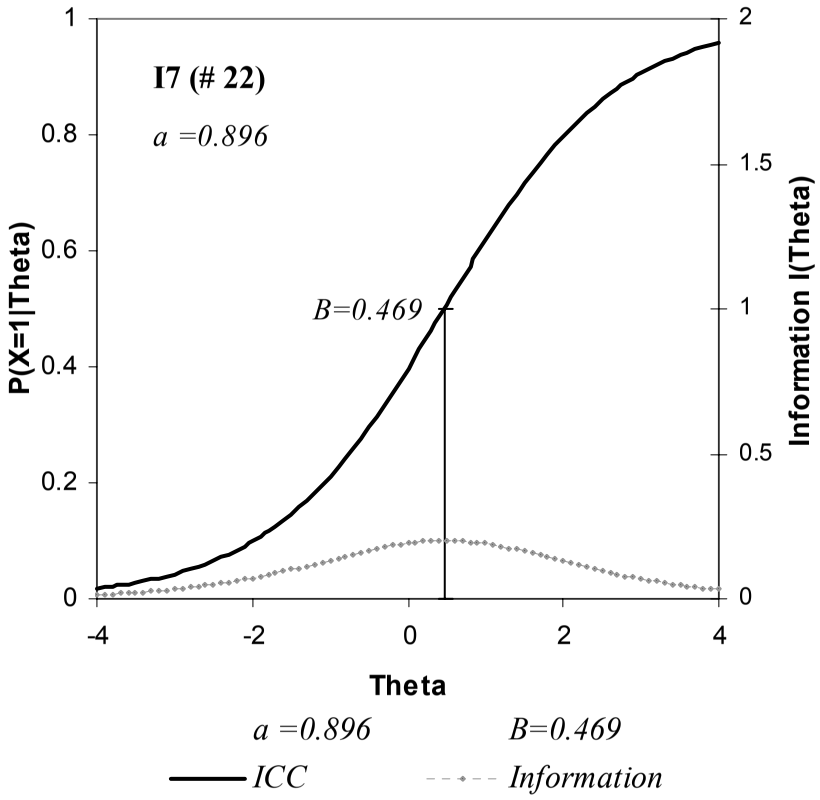
a =0.31 *B=2.659* ICC *Information*

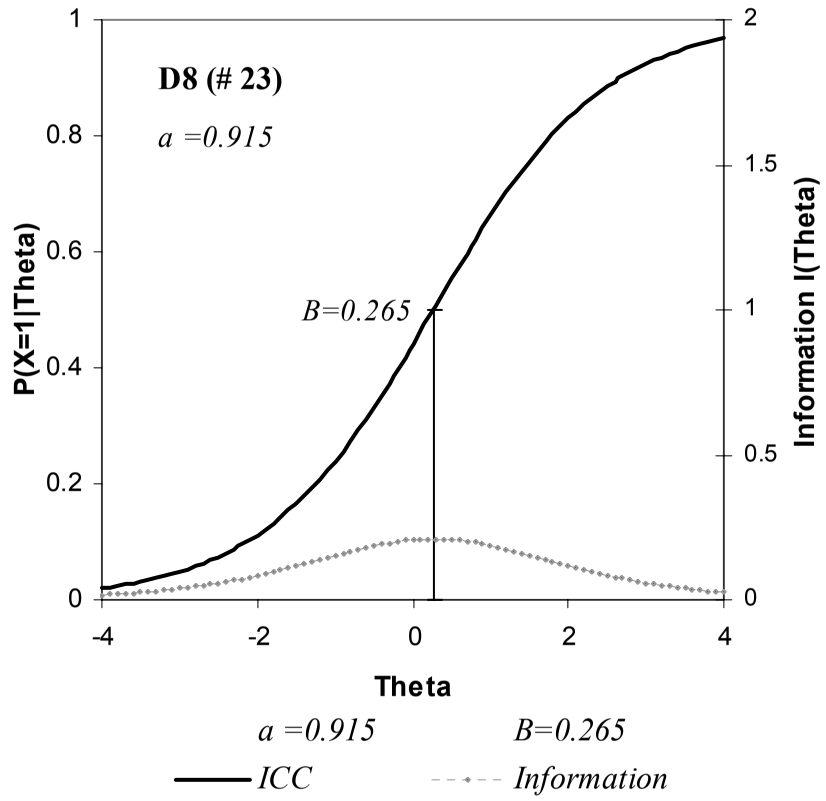


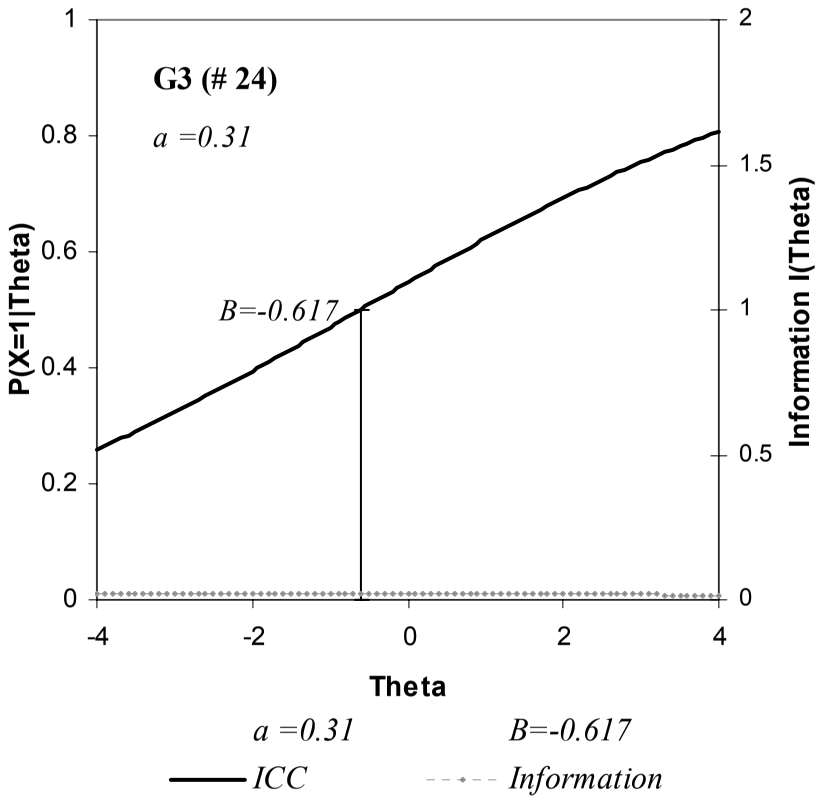
a =0.894 *B=0.745*

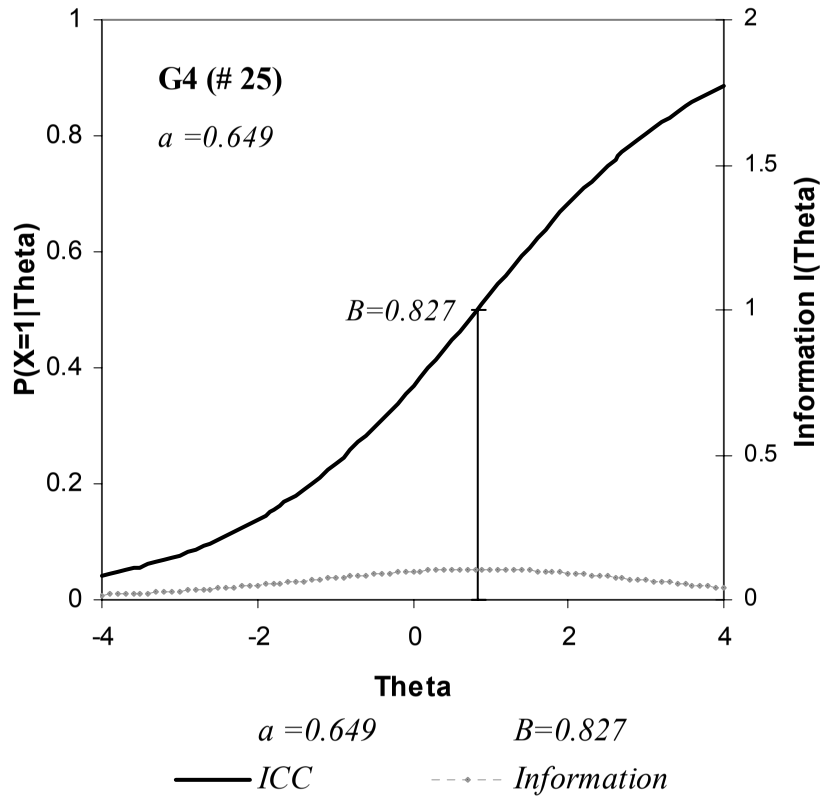
ICC *Information*

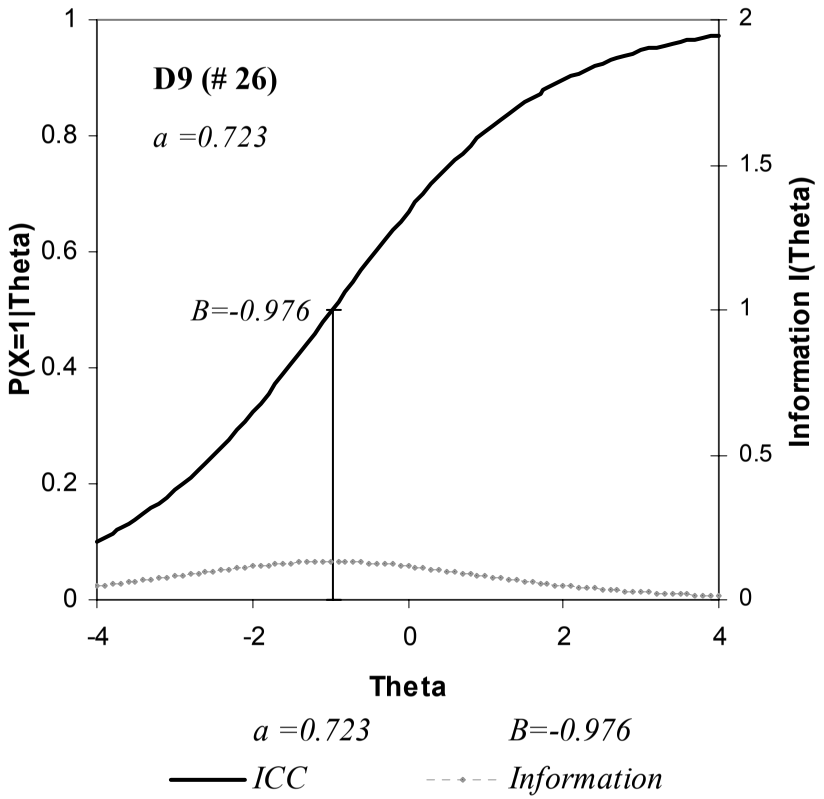

P6 (# 21)
a =0.473
B=-0.645
P(X=1|Theta)
Information I(Theta)
Theta
a =0.473
B=-0.645
ICC
Information

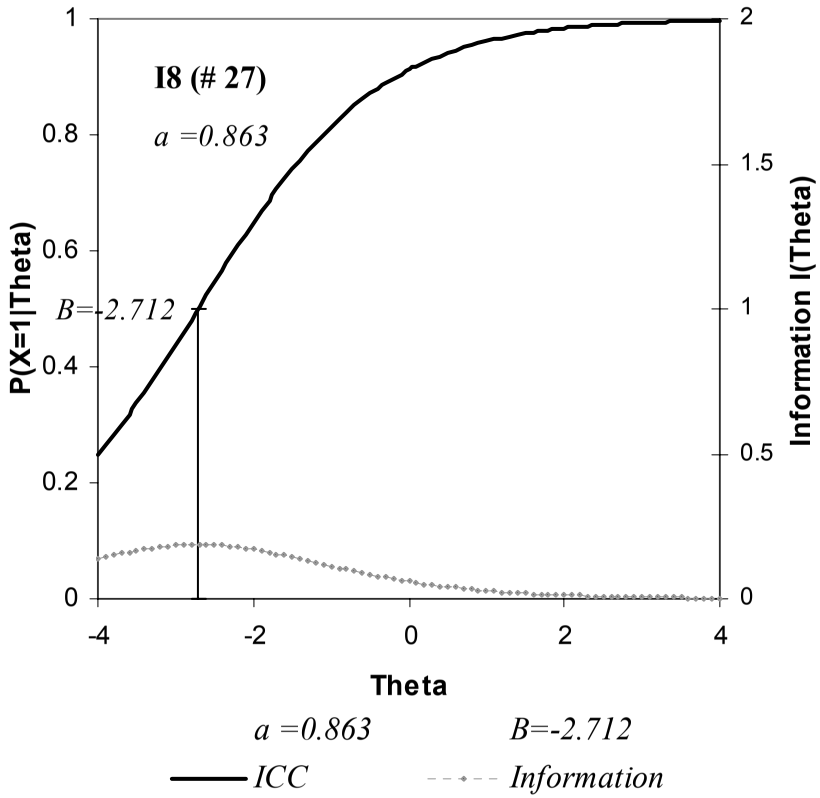

I7 (# 22)
a =0.896
B=0.469
P(X=1|Theta)
Information I(Theta)
Theta
a =0.896
B=0.469
ICC
Information

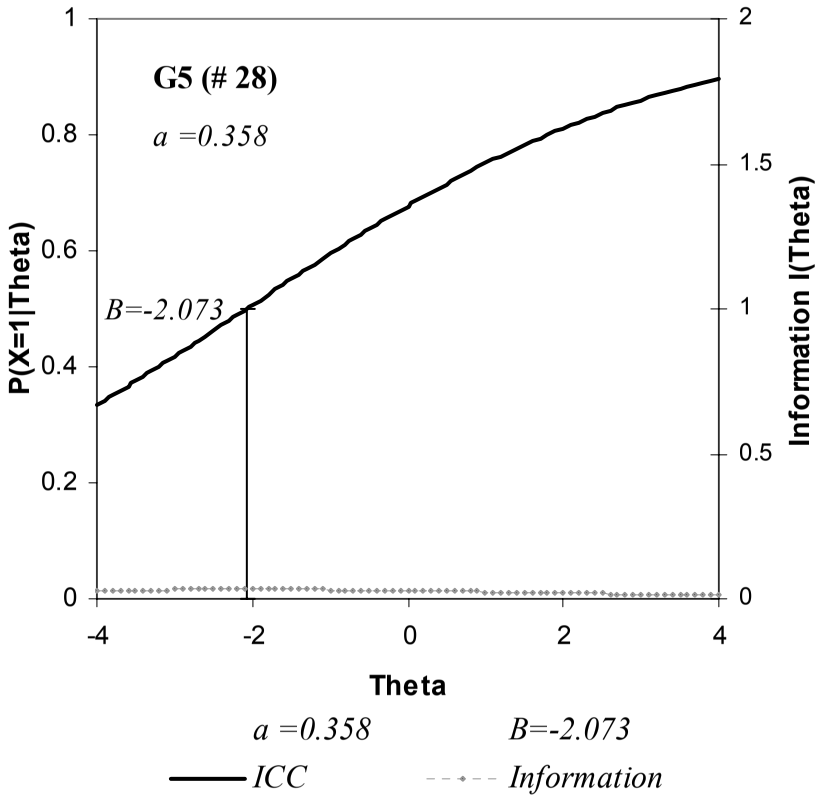

D8 (# 23)
a =0.915
B=0.265
P(X=1|Theta)
Information I(Theta)
Theta
a =0.915
B=0.265
ICC
Information

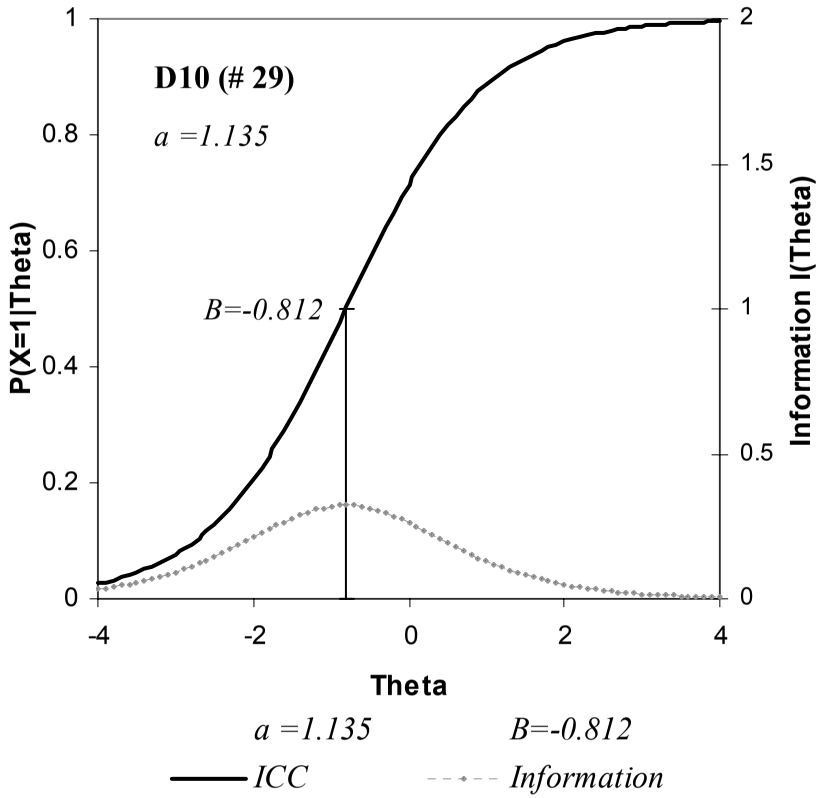

G3 (# 24)
a =0.31
B=-0.617
P(X=1|Theta)
Information I(Theta)
Theta
a =0.31
B=-0.617
ICC
Information

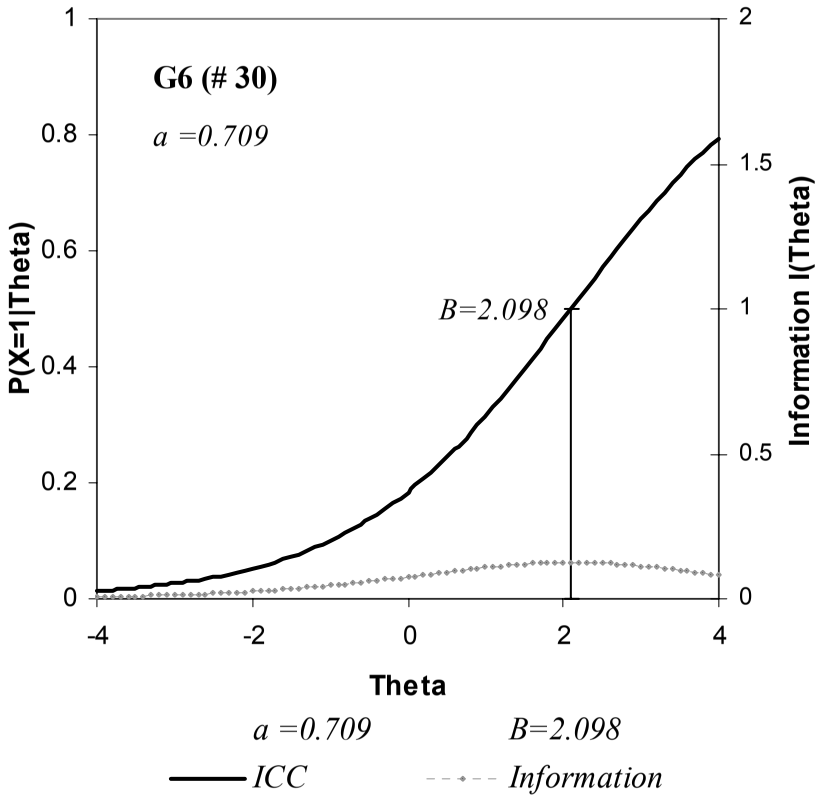

G4 (# 25)
a =0.649
B=0.827
P(X=1|Theta)
Information I(Theta)
Theta
1
0.8
0.6
0.4
0.2
0
2
1.5
1
0.5
0
-4
-2
0
2
4
a =0.649
B=0.827
ICC
Information

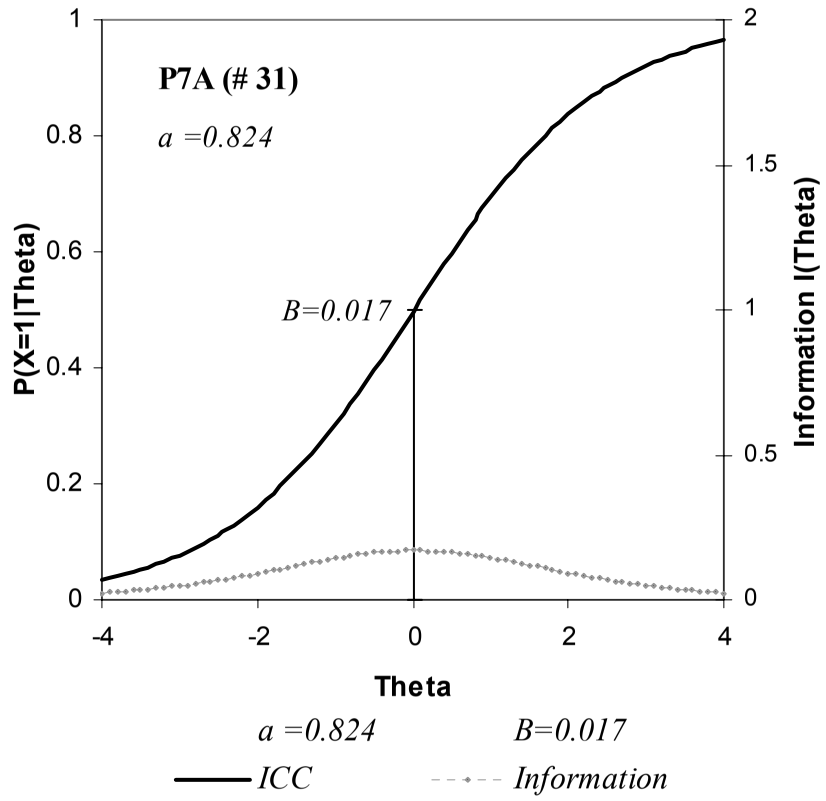

D9 (# 26)
a =0.723
B=-0.976
P(X=1|Theta)
Information I(Theta)
Theta
1
0.8
0.6
0.4
0.2
0
2
1.5
1
0.5
0
-4
-2
0
2
4
a =0.723
B=-0.976
ICC
Information

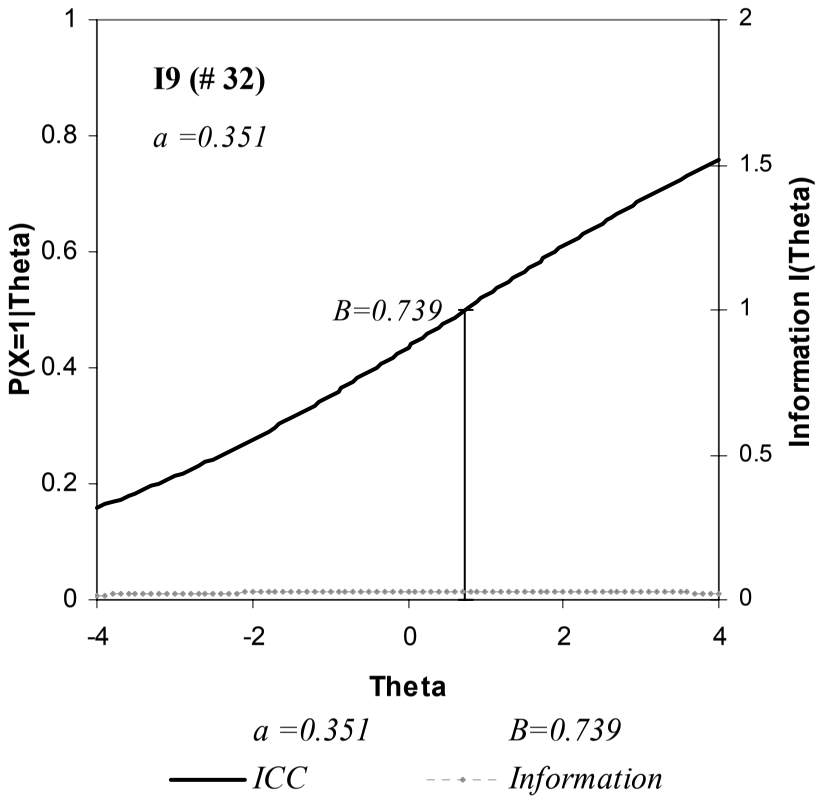

I8 (# 27)
a =0.863
B=-2.712
P(X=1|Theta)
Information I(Theta)
Theta
a =0.863
B=-2.712
ICC
Information

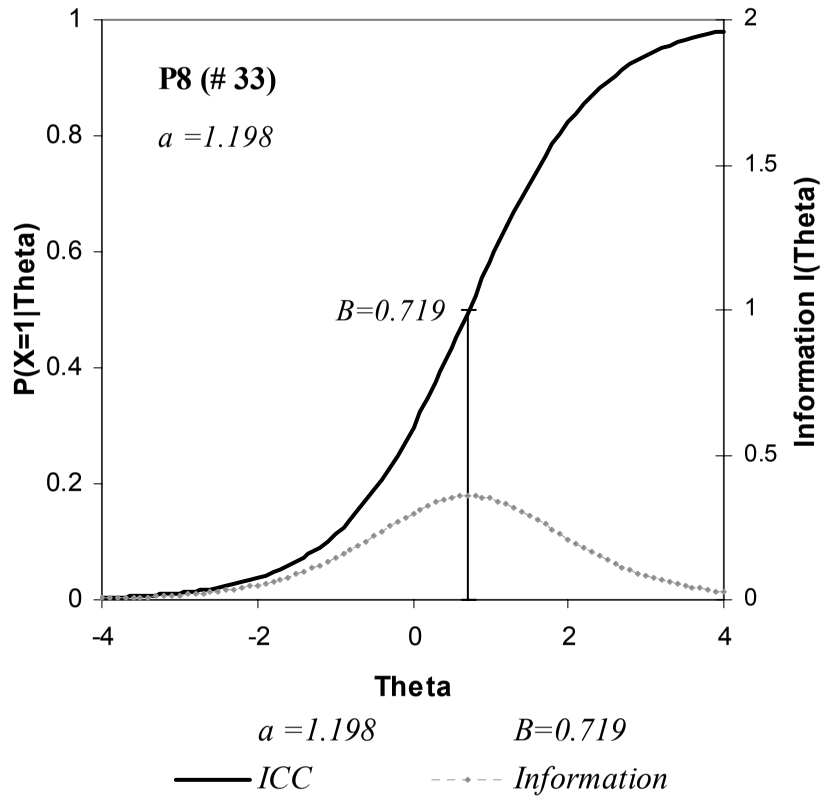

G5 (# 28)
a =0.358
B=-2.073
P(X=1|Theta)
Information I(Theta)
Theta
a =0.358
B=-2.073
ICC
Information


D10 (# 29)
a =1.135
B=-0.812
P(X=1|Theta)
Information I(Theta)
Theta
a =1.135
B=-0.812
ICC
Information


G6 (# 30)
a =0.709
B=2.098
P(X=1|Theta)
Information I(Theta)
Theta
a =0.709
B=2.098
ICC
Information

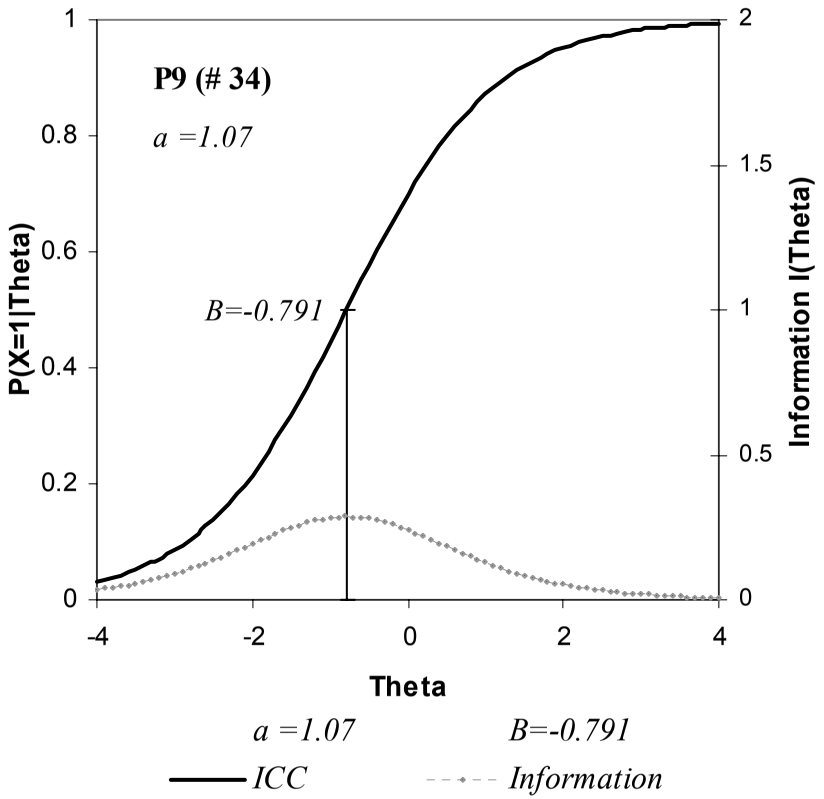

P7A (# 31)
a =0.824
B=0.017
P(X=1|Theta)
Information I(Theta)
Theta
1
0.8
0.6
0.4
0.2
0
2
1.5
1
0.5
0
-4
-2
0
2
4
a =0.824
B=0.017
ICC
Information


I9 (# 32)
a =0.351
B=0.739
P(X=1|Theta)
Information I(Theta)
Theta
1
0.8
0.6
0.4
0.2
0
2
1.5
1
0.5
0
-4
-2
0
2
4
a =0.351
B=0.739
ICC
Information



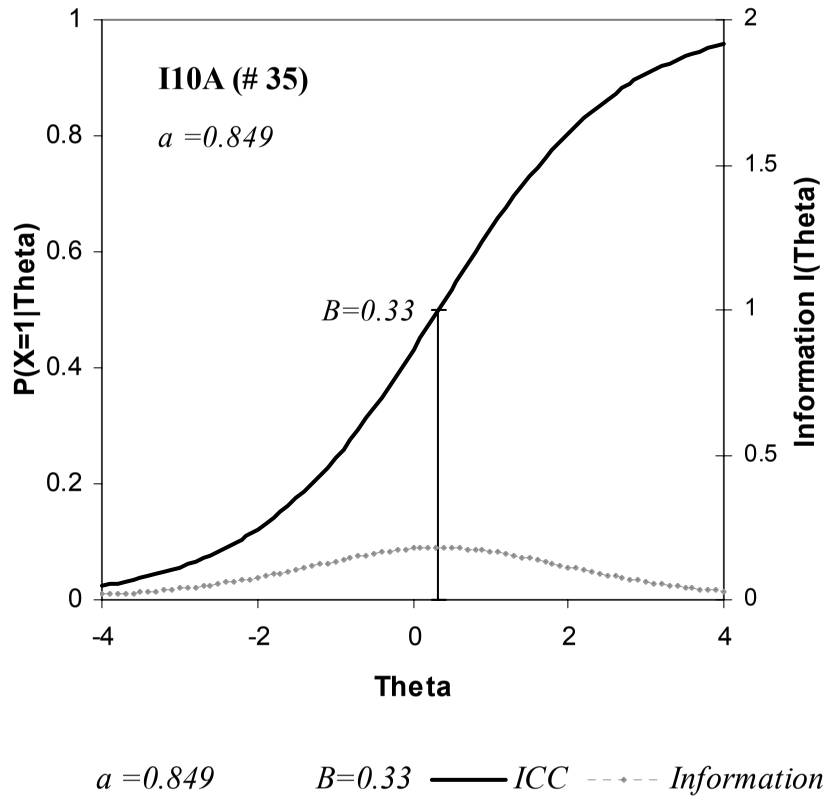
a =1.198 *B=0.719*

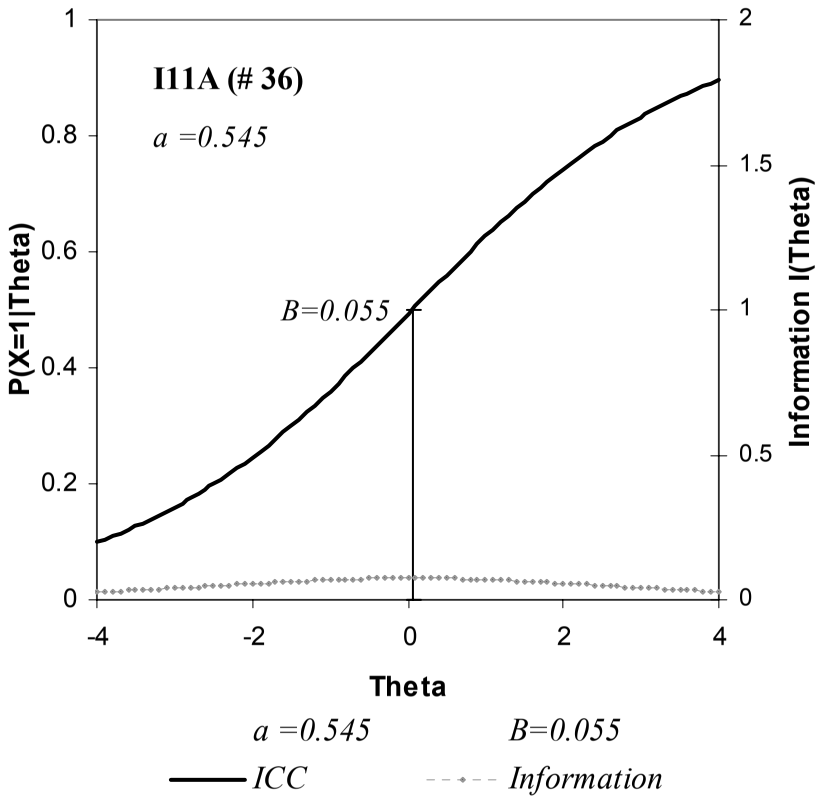
ICC *Information*

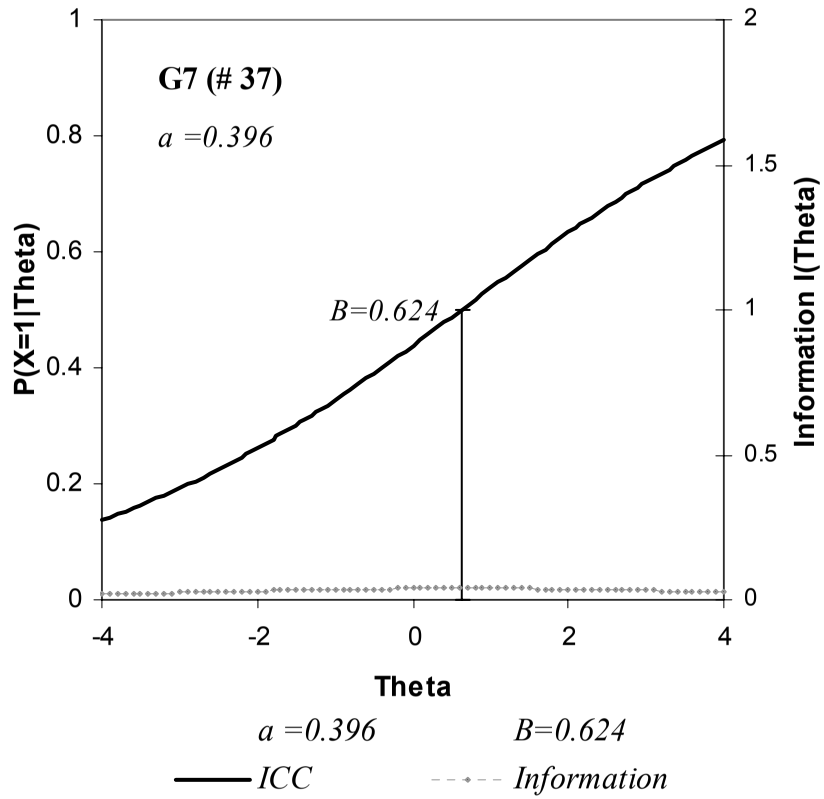


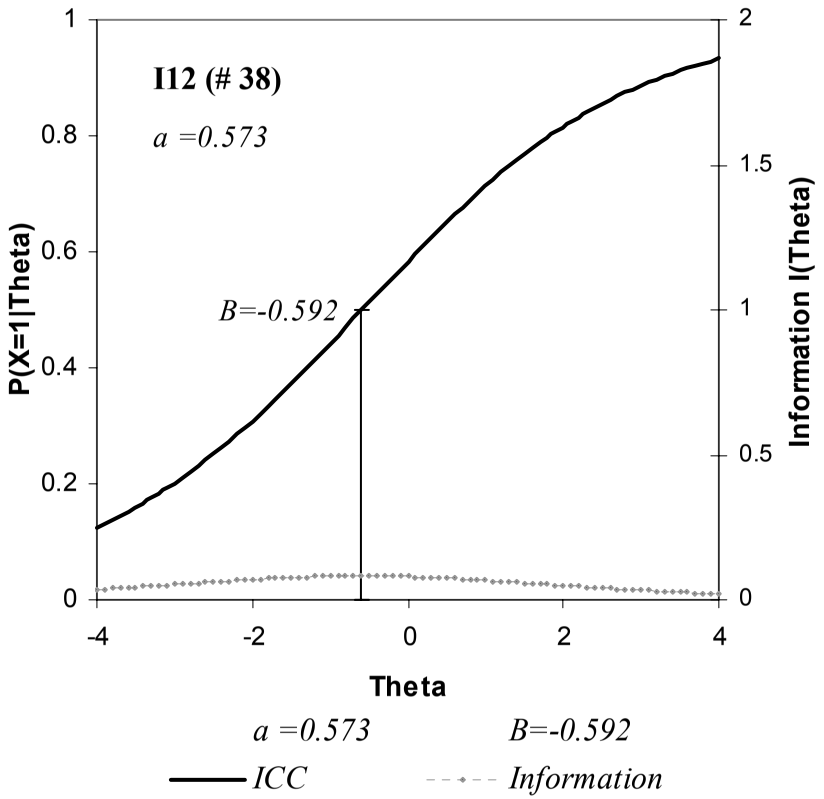
a =1.07 *B=-0.791*

ICC *Information*


I10A (# 35)
a =0.849
B=0.33
P(X=1|Theta)
Information I(Theta)
Theta
1
0.8
0.6
0.4
0.2
0
2
1.5
1
0.5
0
-4
-2
0
2
4
a =0.849
B=0.33
ICC
Information


I11A (# 36)
a =0.545
B=0.055
P(X=1|Theta)
Information I(Theta)
Theta
1
0.8
0.6
0.4
0.2
0
2
1.5
1
0.5
0
-4
-2
0
2
4
a =0.545
B=0.055
ICC
Information

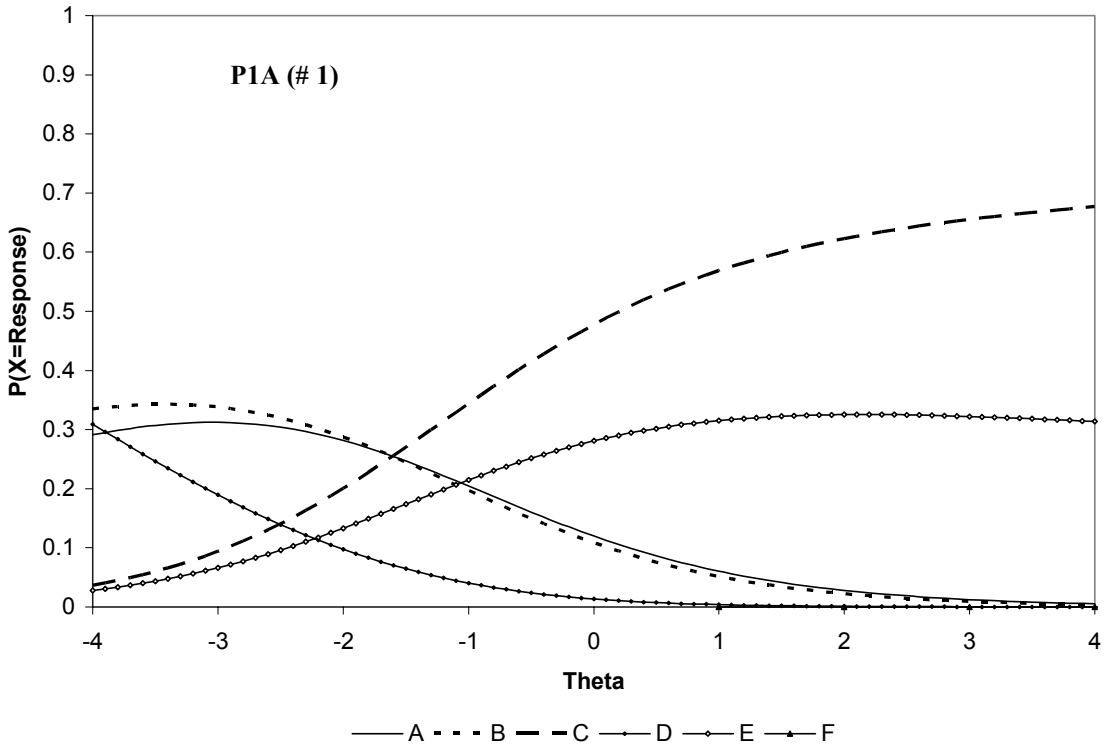

G7 (# 37)
a =0.396
B=0.624
P(X=1|Theta)
Information I(Theta)
Theta
1
0.8
0.6
0.4
0.2
0
2
1.5
1
0.5
0
-4
-2
0
2
4
a =0.396
B=0.624
ICC
Information

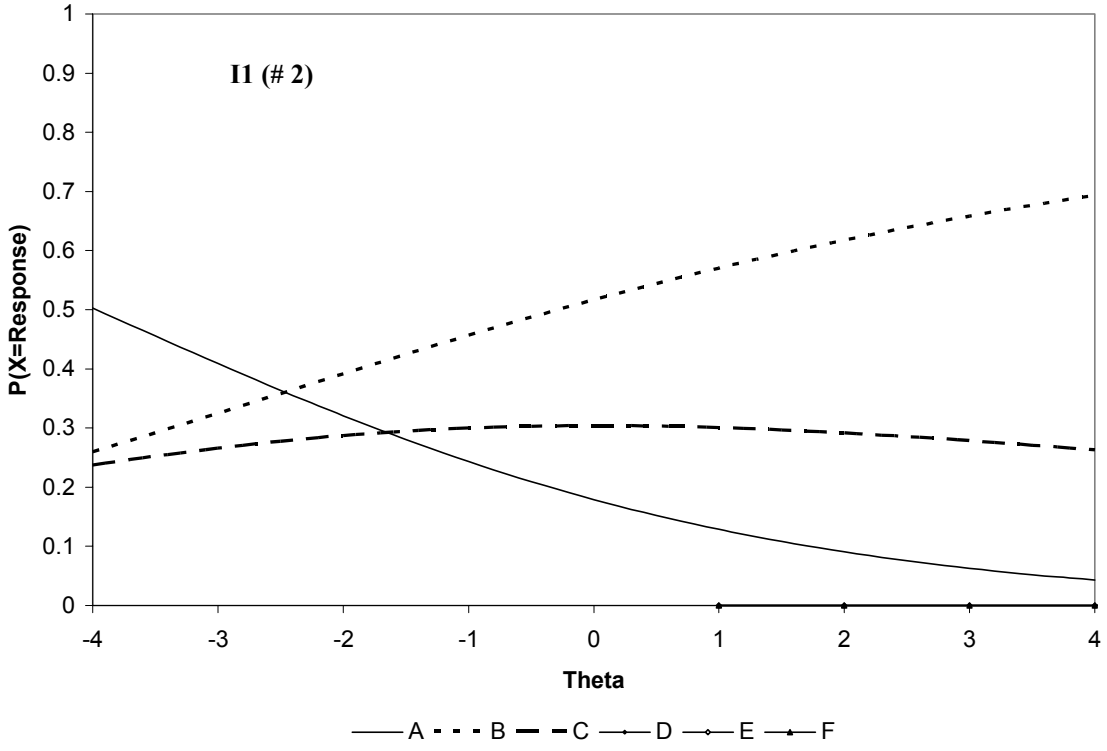

I12 (# 38)
a =0.573
B=-0.592
P(X=1|Theta)
Information I(Theta)
Theta
1
0.8
0.6
0.4
0.2
0
2
1.5
1
0.5
0
-4
-2
0
2
4
a =0.573
B=-0.592
ICC
Information

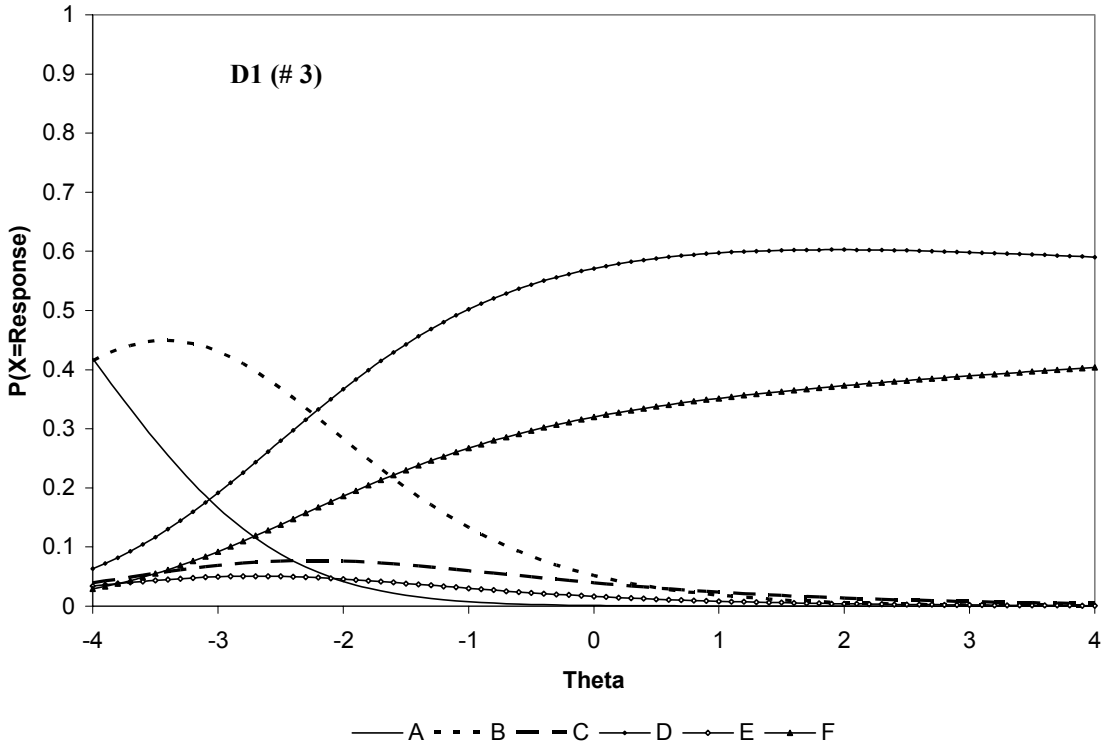
Appendix B: Nominal Response Model Item Response Curves

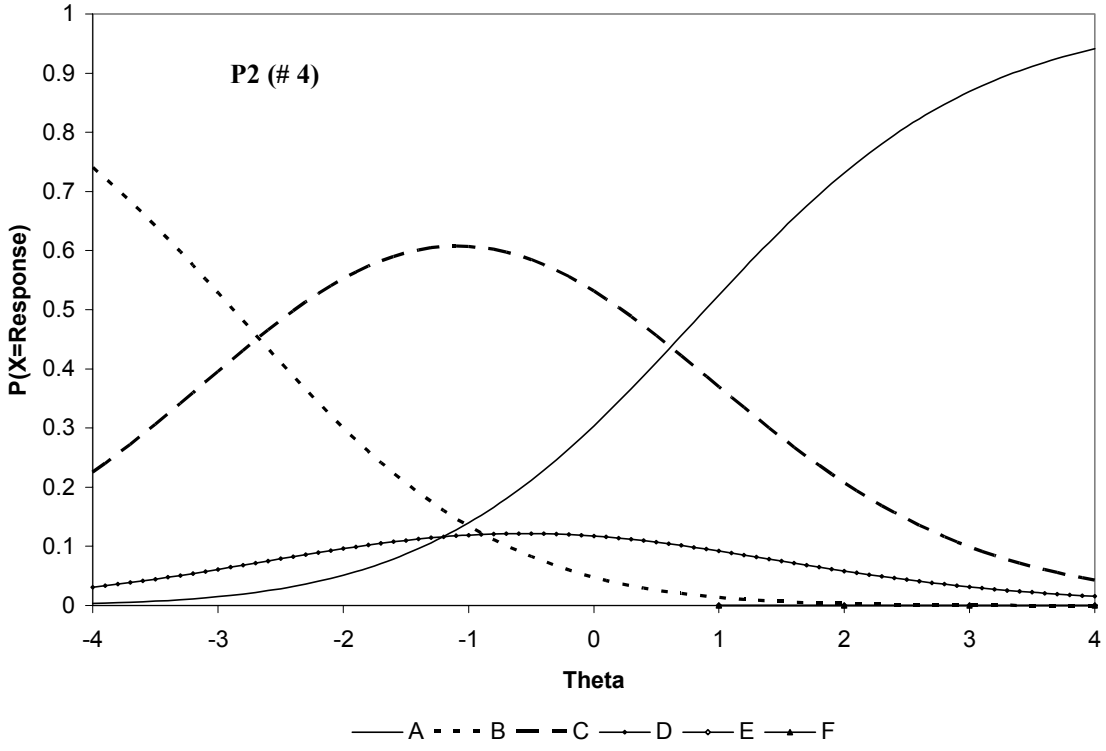
The Response Curves for the nominal model are included in this appendix. The 38 questions on the current version of the SCI are presented. The numbers in parentheses on the graphs correspond to the item numbers used on the SCI and those used in chapter 3. The labels preceding the numbers (such as P1, I10a, etc) are the master numbers assigned to the items to track them historically. The initial letter in the master number assigns the question to a topic group: probability (P), descriptive statistics (D), inferential statistics (I), and graphical (G). Some master numbers include a second letter which designates the version of the question, such as P1a. Chapters 3 and 4 provide discussion on the item versions.

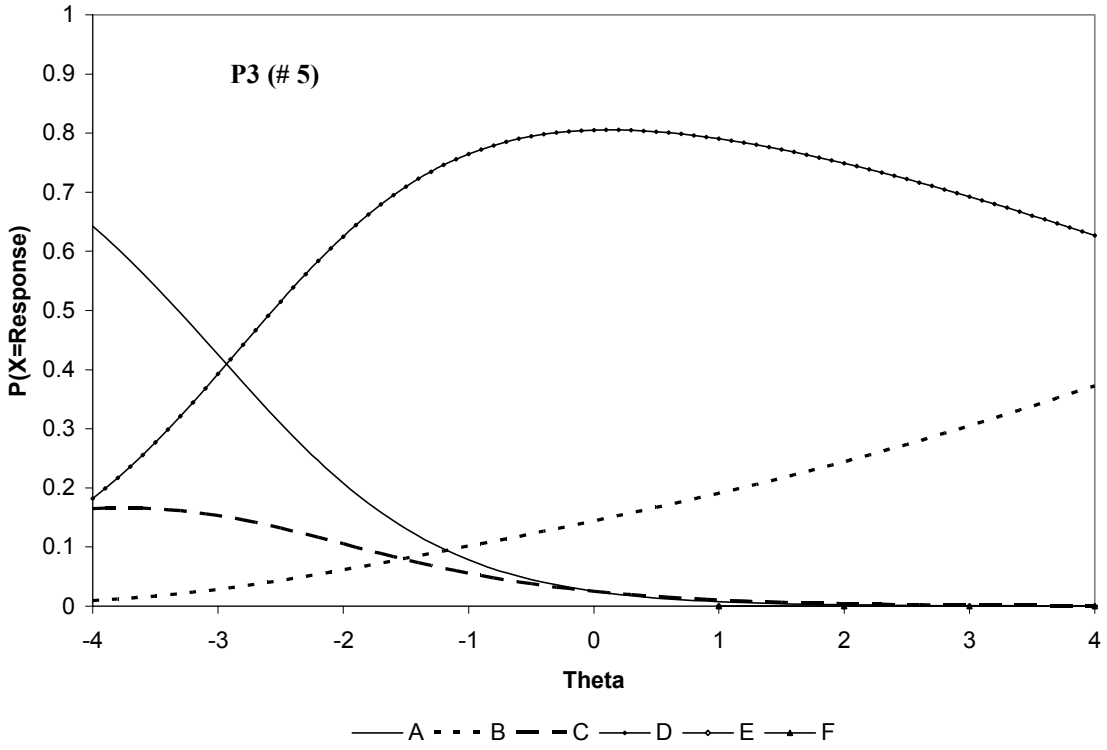
The latent trait, Θ, is assumed to be conceptual understanding of statistics and is plotted on the horizontal axis. It is assumed to have a normal distribution in the population with mean zero and standard deviation one. The response curves represent the probability of choosing response alternative *k* for the given theta value. The probability is shown on the left vertical axis. The response curves are shown with distinct line patterns which correspond to responses (a), (b), (c), (d), (e), or (f). The pattern key is shown below each graph.

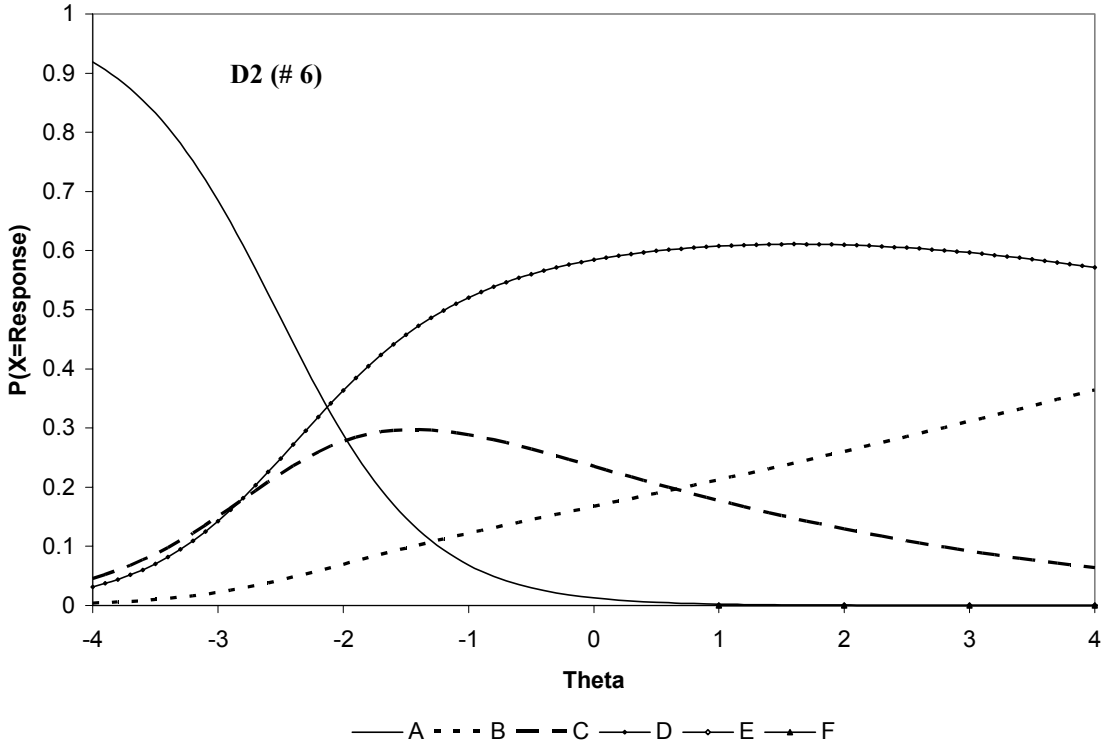

P1A (# 1)
P(X=Response)
Theta
A
B
C
D
E
F

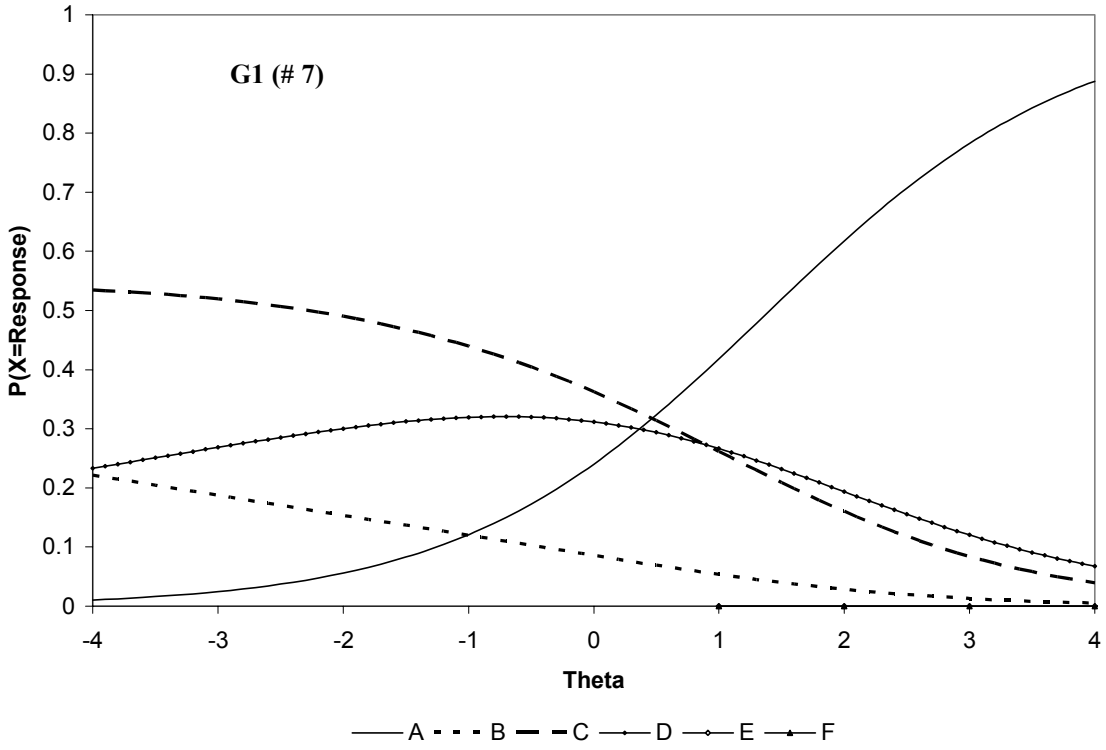

I1 (# 2)
P(X=Response)
Theta
A
B
C
D
E
F

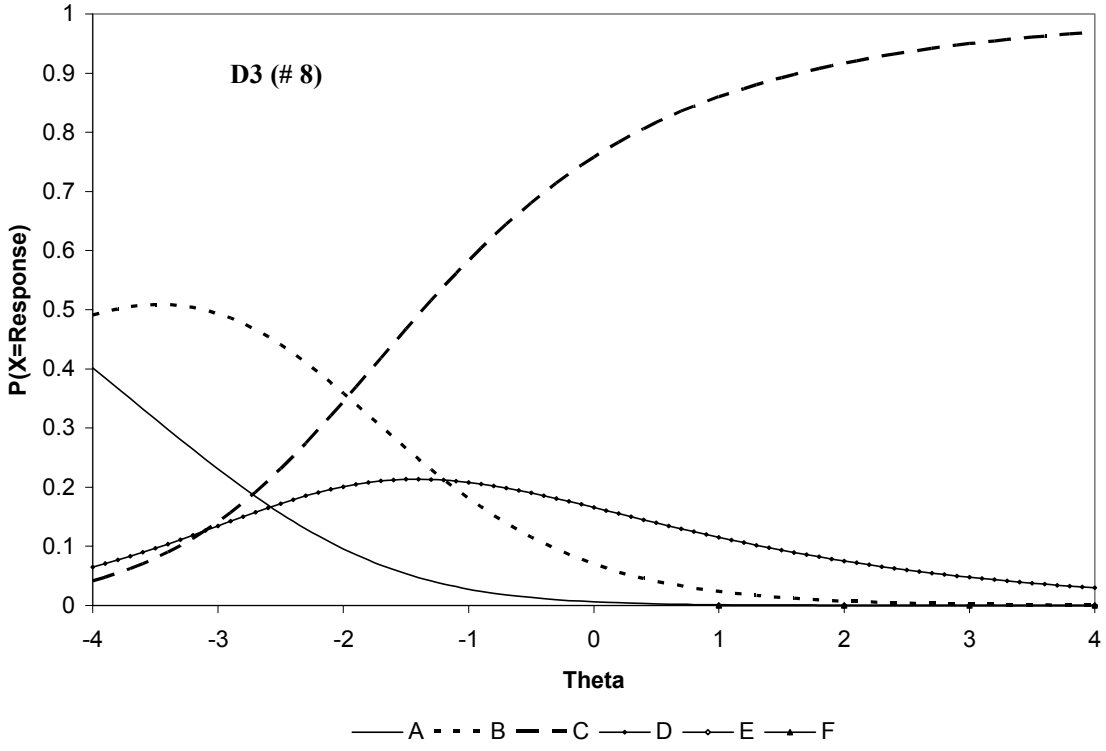

D1 (# 3)
P(X=Response)
Theta
A
B
C
D
E
F

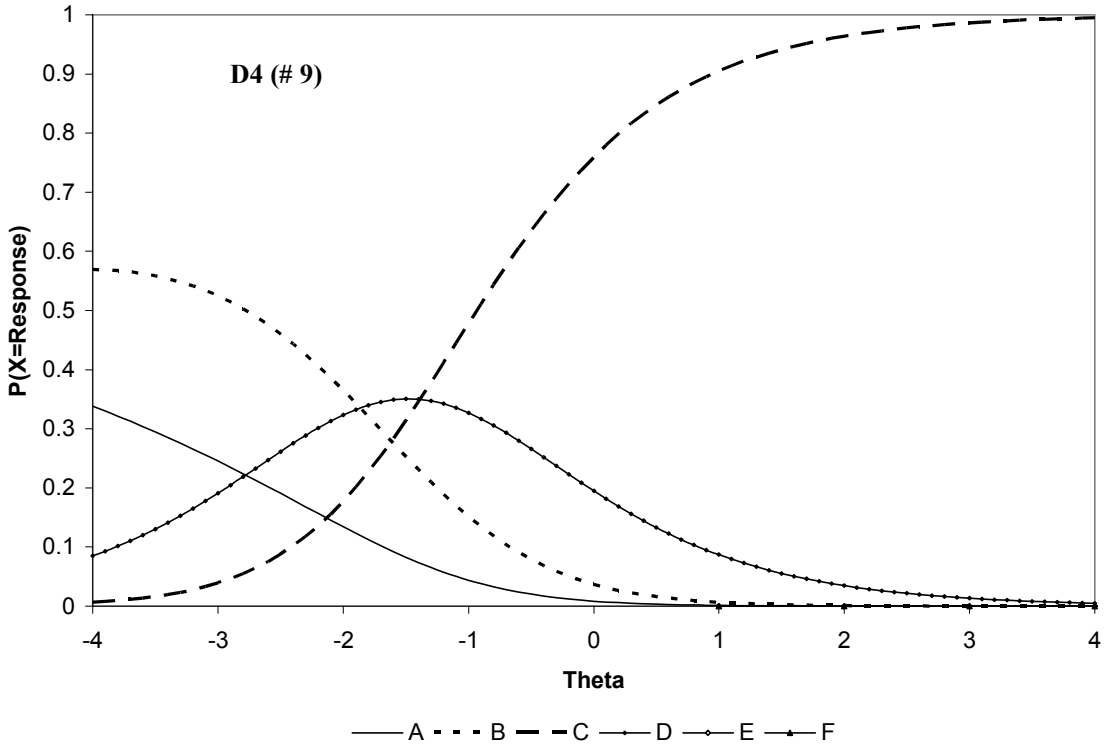

P2 (# 4)
P(X=Response)
Theta
A
B
C
D
E
F

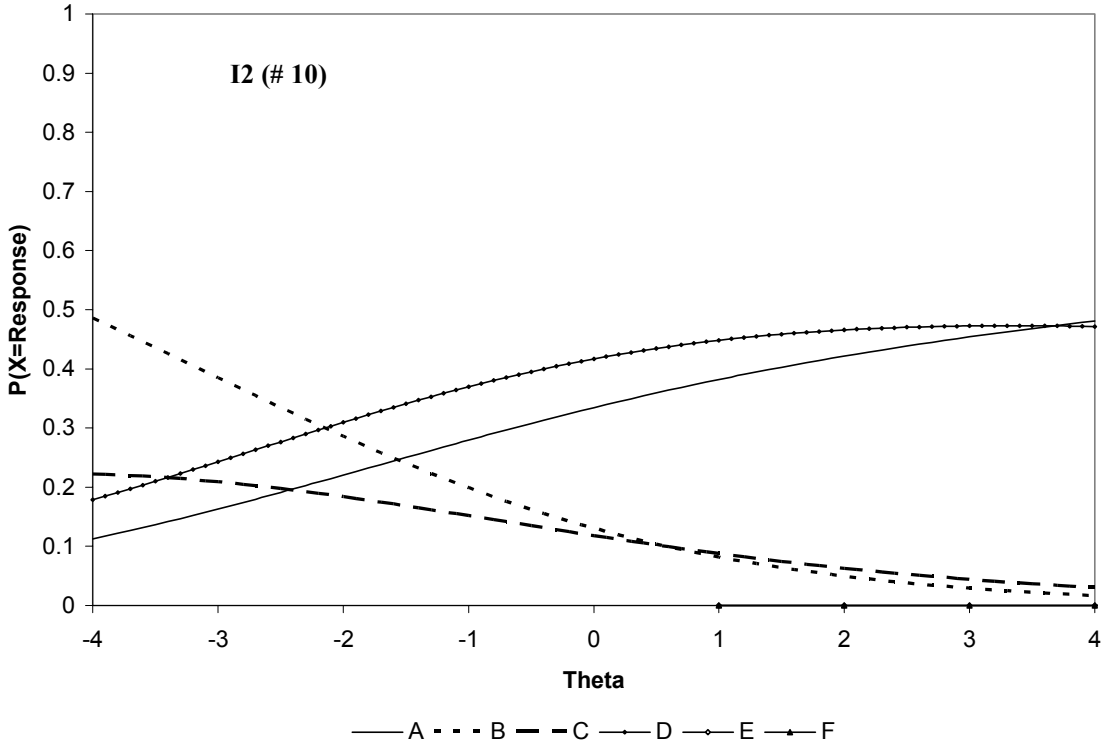

P3 (# 5)
P(X=Response)
Theta
A
B
C
D
E
F

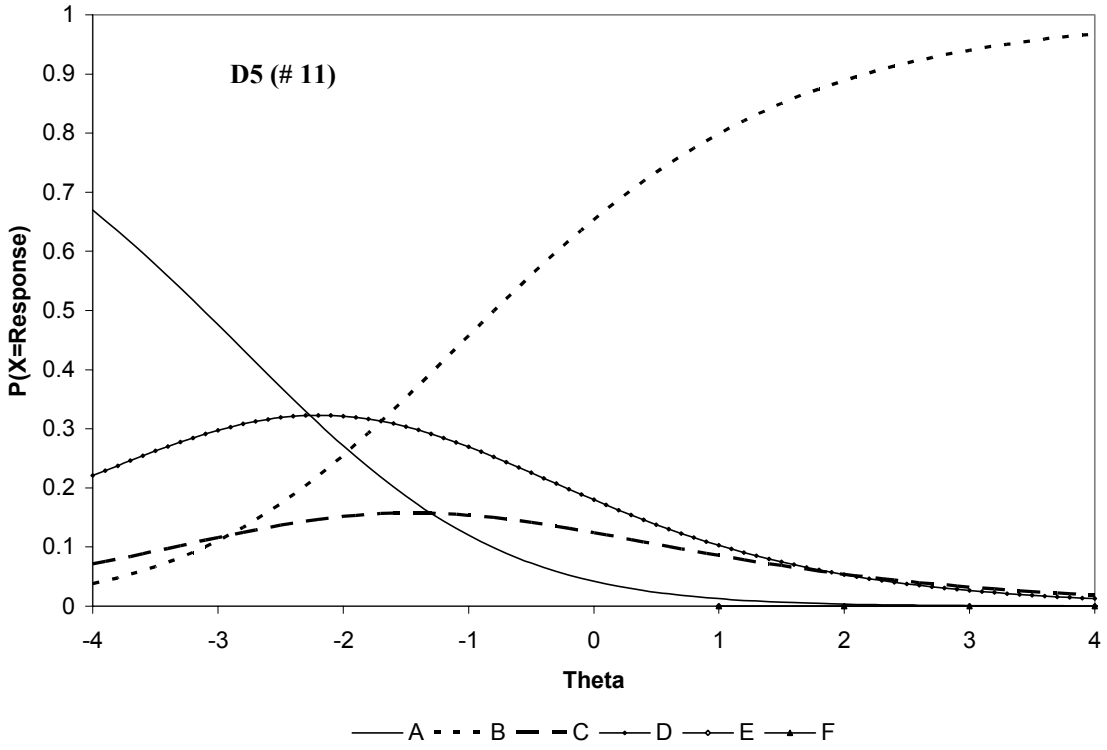

D2 (# 6)
P(X=Response)
Theta
A
B
C
D
E
F

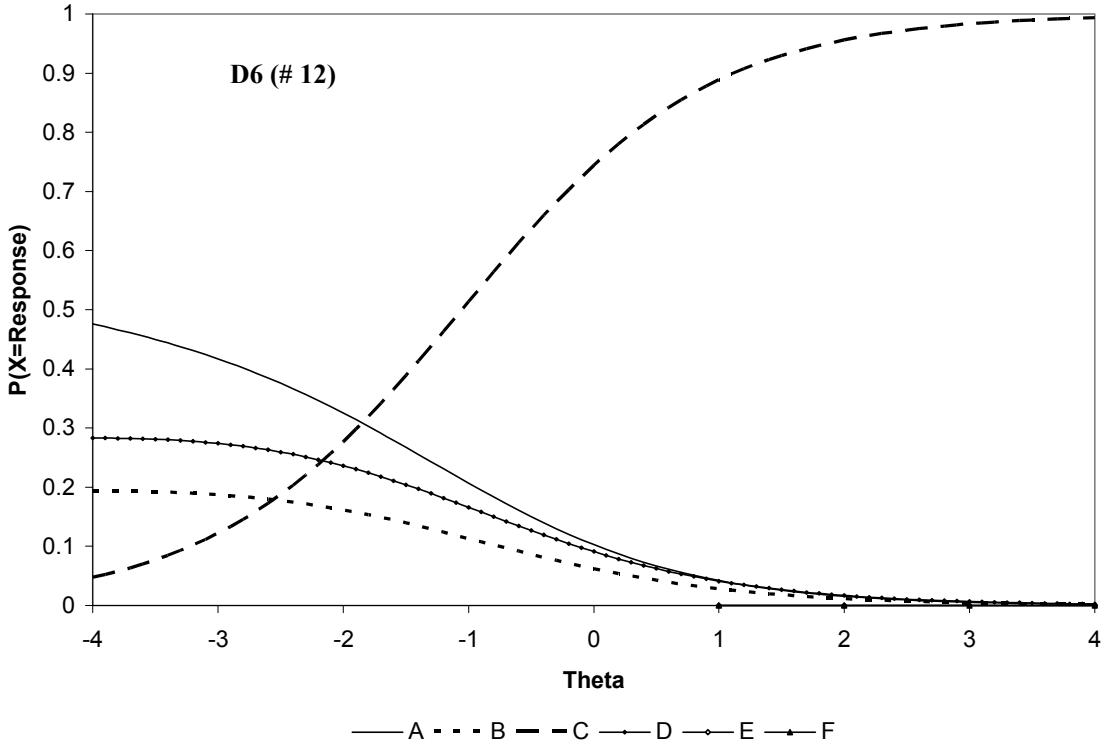

G1 (# 7)
P(X=Response)
Theta
A
B
C
D
E
F

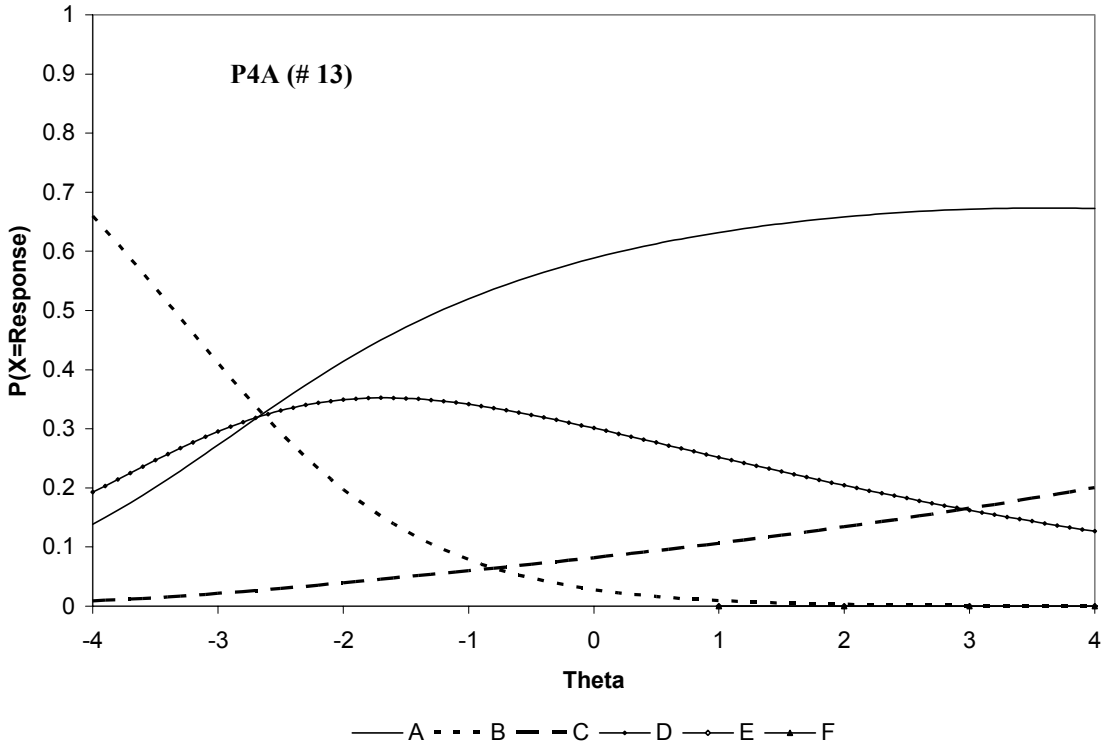

D3 (# 8)
P(X=Response)
Theta
A
B
C
D
E
F

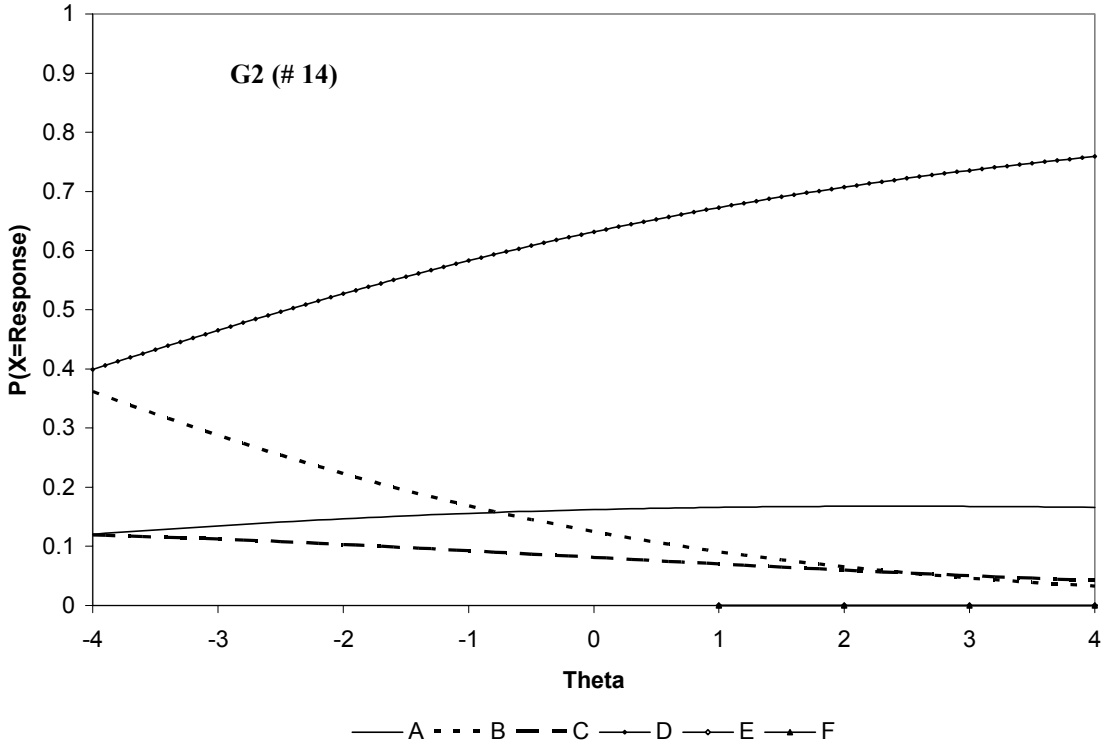

D4 (# 9)
P(X=Response)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-4
-3
-2
-1
0
1
2
3
4
Theta
A
B
C
D
E
F

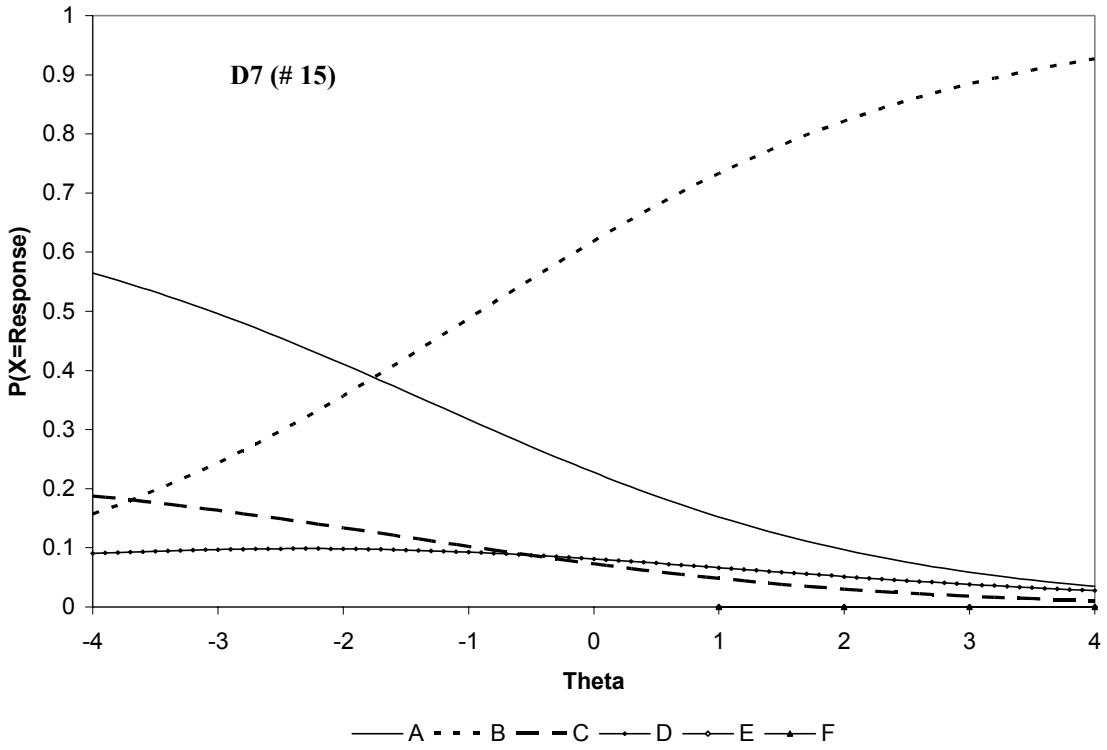

I2 (# 10)
P(X=Response)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-4
-3
-2
-1
0
1
2
3
4
Theta
A
B
C
D
E
F

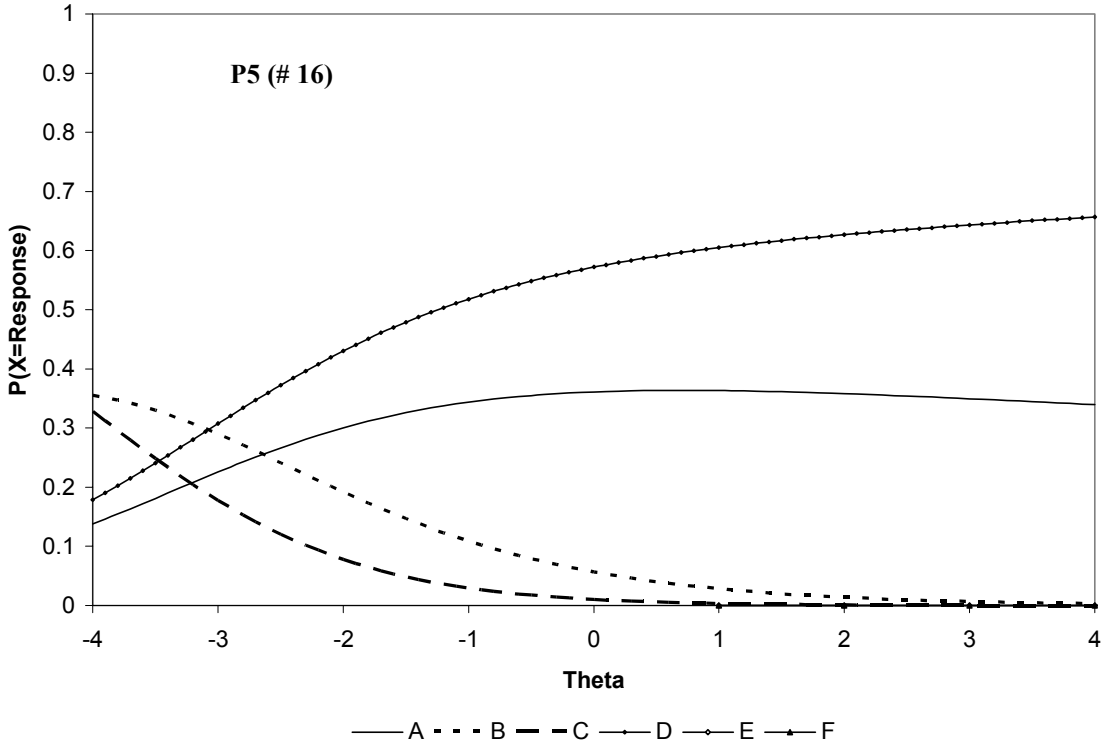

D5 (# 11)
P(X=Response)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-4
-3
-2
-1
0
1
2
3
4
Theta
A
B
C
D
E
F

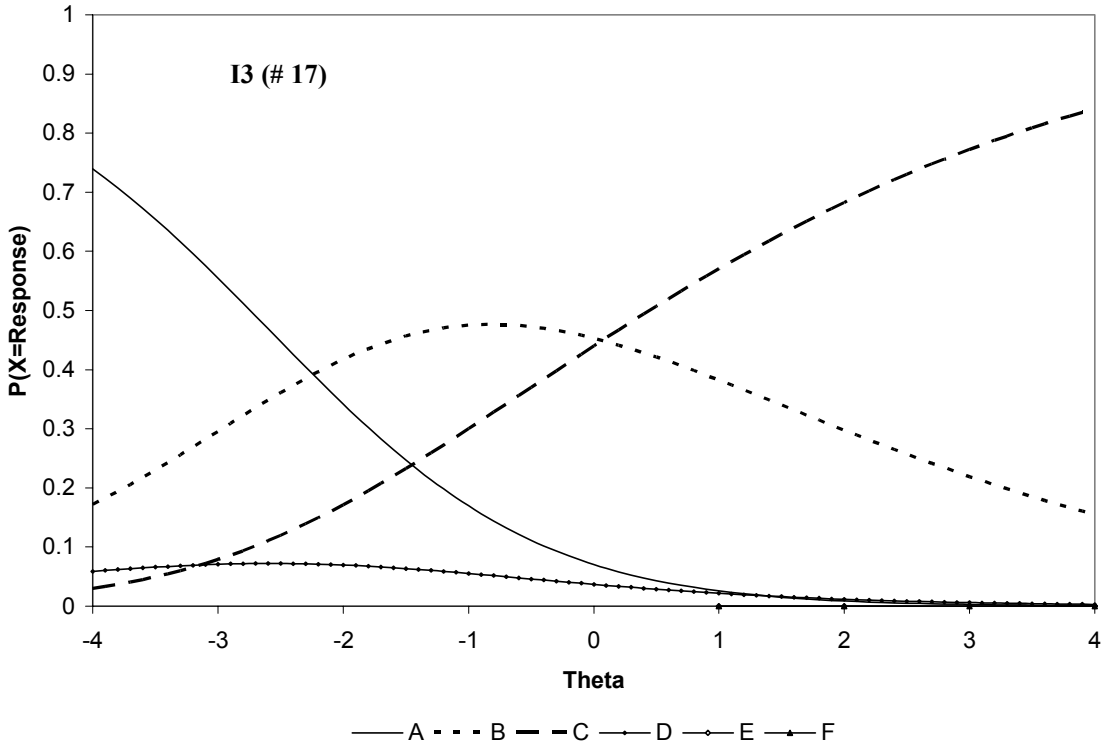

D6 (# 12)
P(X=Response)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-4
-3
-2
-1
0
1
2
3
4
Theta
A
B
C
D
E
F

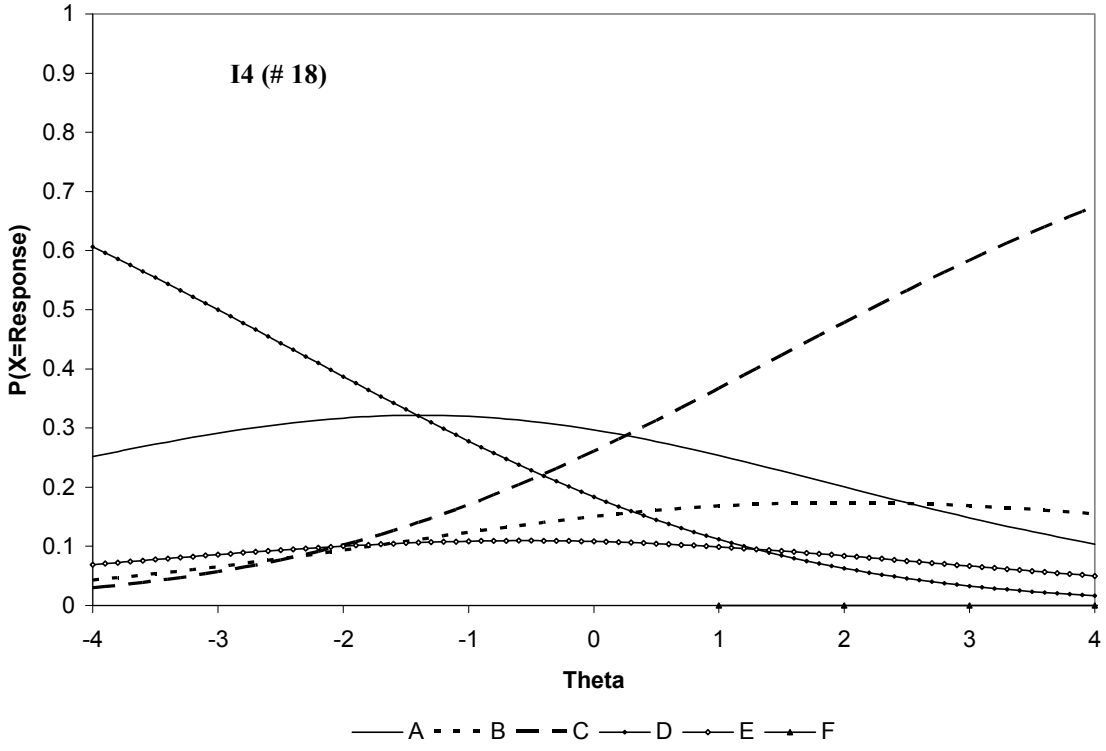

P4A (# 13)
P(X=Response)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-4
-3
-2
-1
0
1
2
3
4
Theta
A
B
C
D
E
F

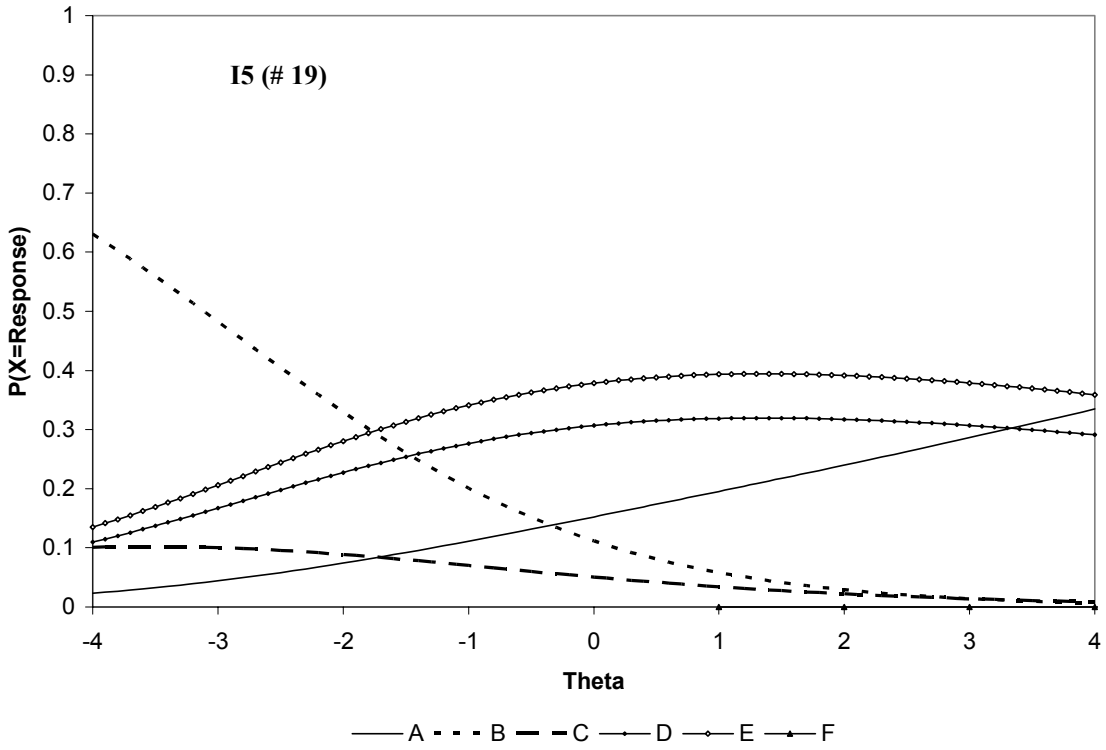

G2 (# 14)
P(X=Response)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-4
-3
-2
-1
0
1
2
3
4
Theta
A
B
C
D
E
F

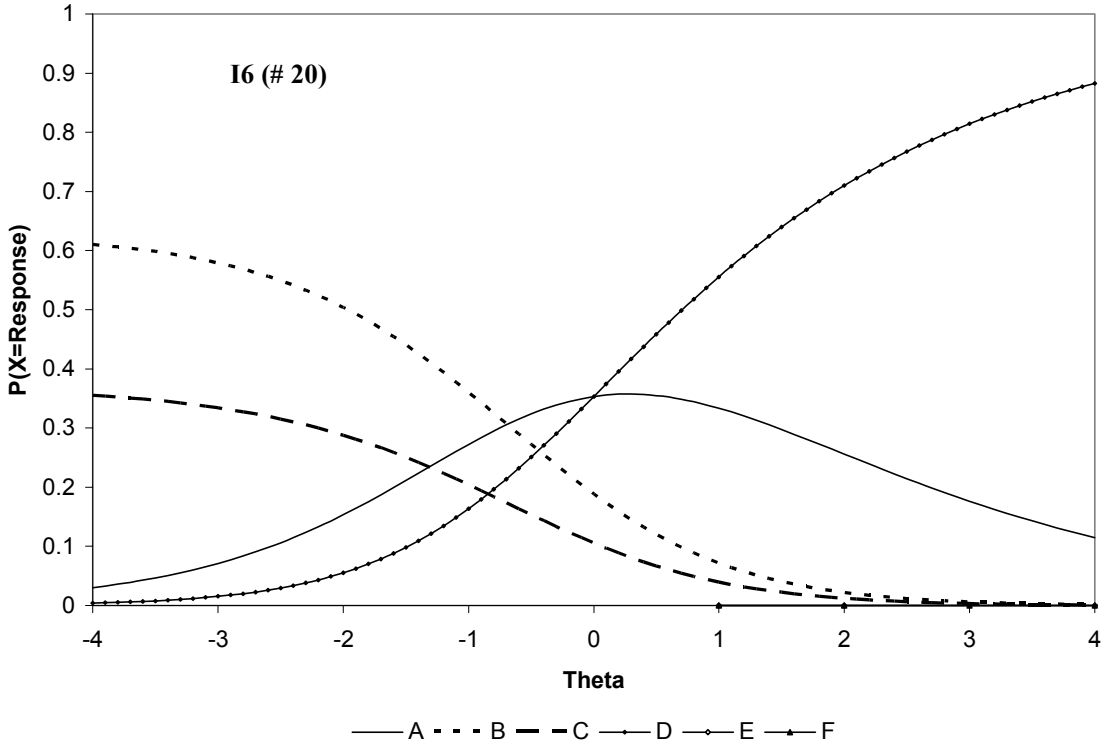

D7 (# 15)
P(X=Response)
Theta
A
B
C
D
E
F

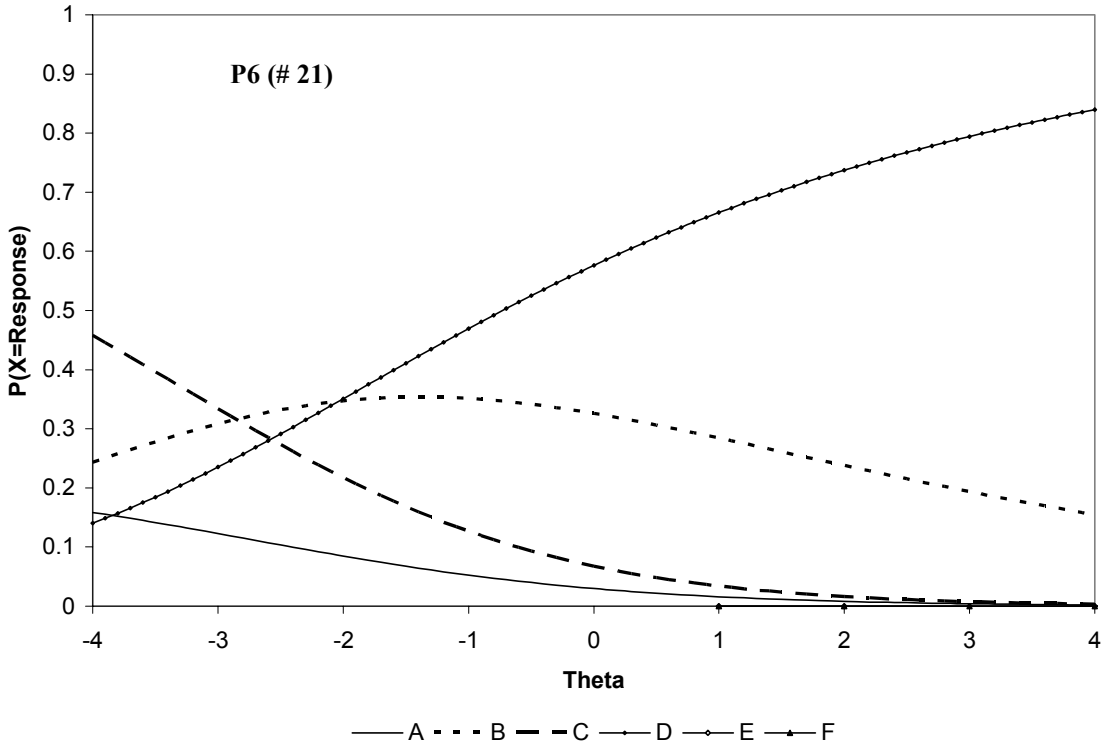

P5 (# 16)
P(X=Response)
Theta
A
B
C
D
E
F

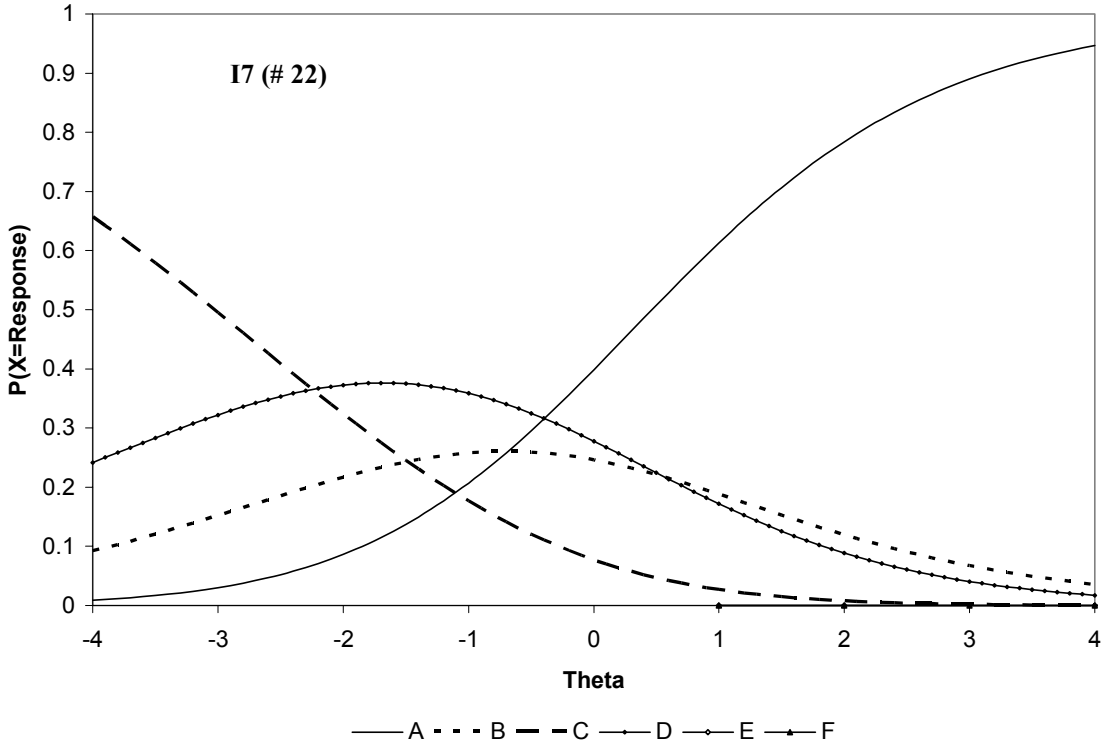

I3 (# 17)
P(X=Response)
Theta
A
B
C
D
E
F

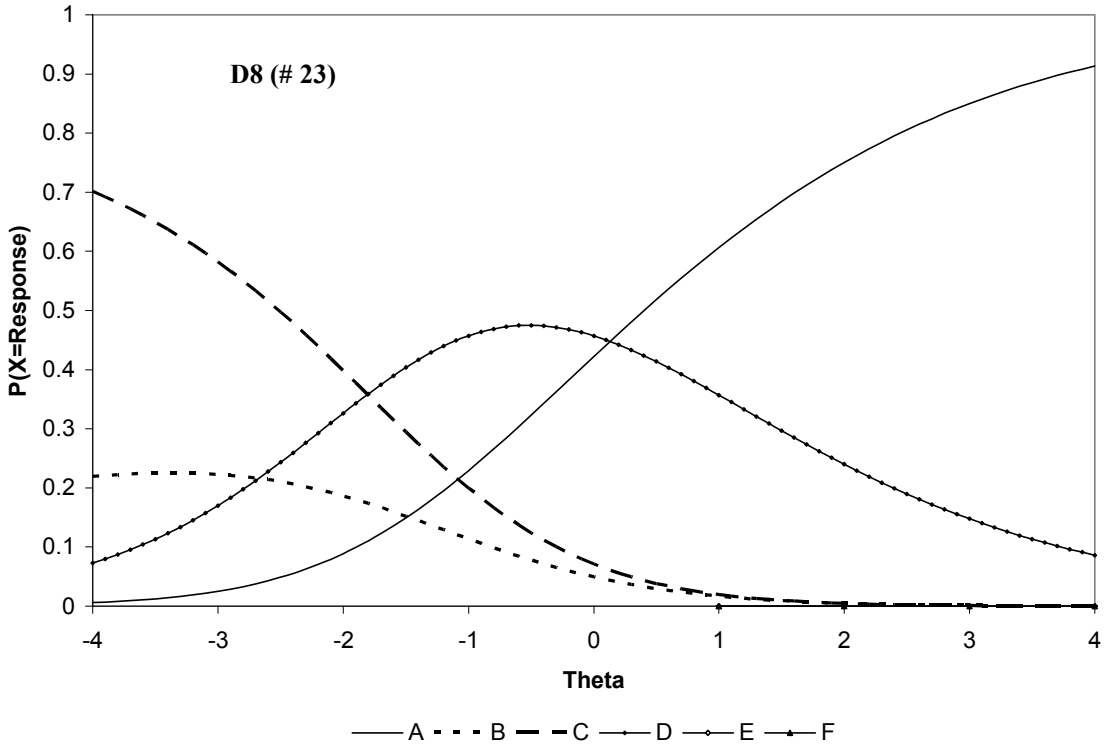

I4 (# 18)
P(X=Response)
Theta
A
B
C
D
E
F

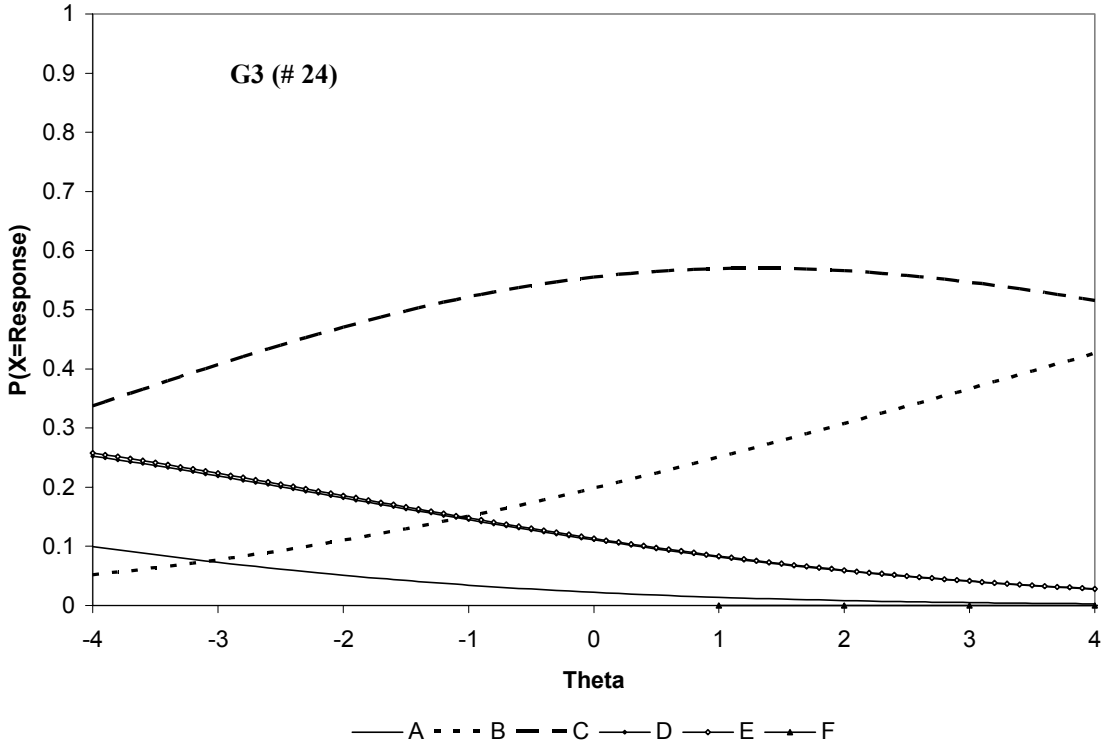

I5 (# 19)
P(X=Response)
Theta
A
B
C
D
E
F

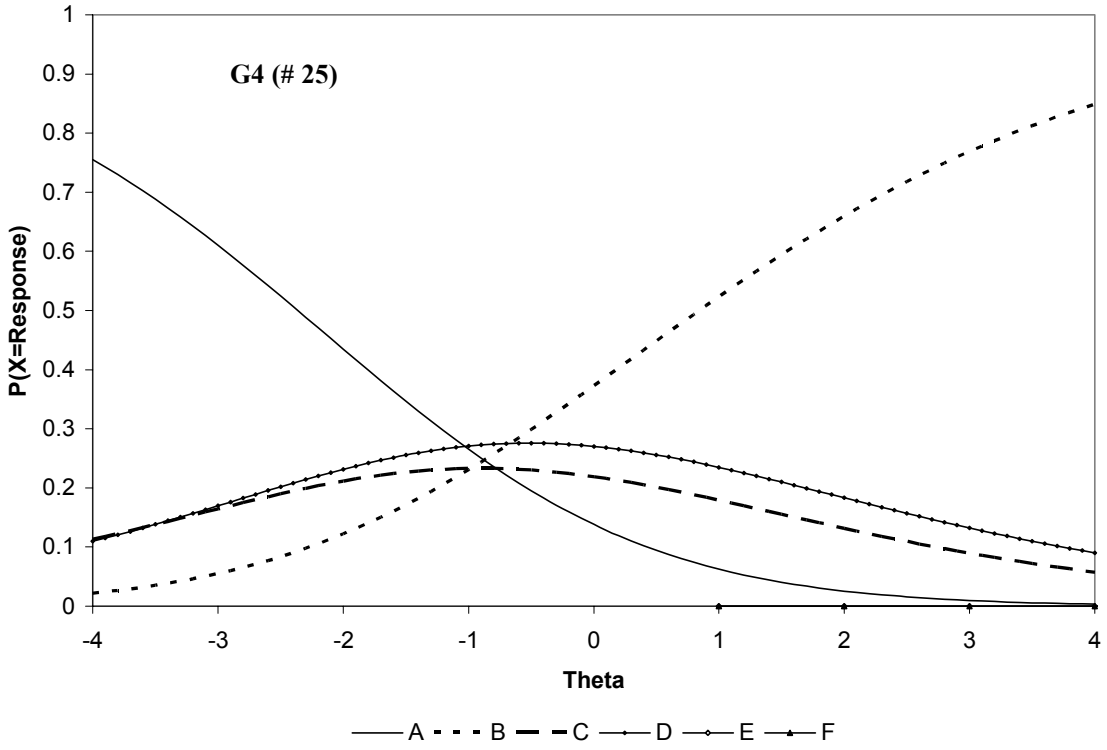

I6 (# 20)
P(X=Response)
Theta
A
B
C
D
E
F

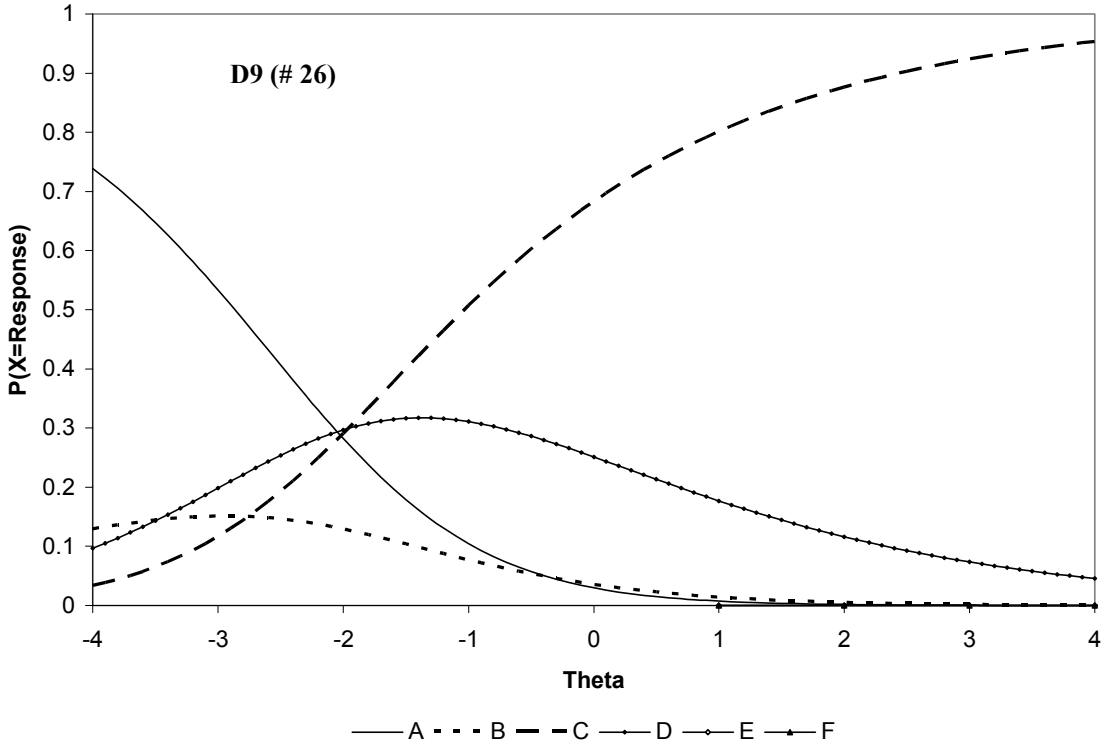

P6 (# 21)
P(X=Response)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-4
-3
-2
-1
0
1
2
3
4
Theta
A
B
C
D
E
F

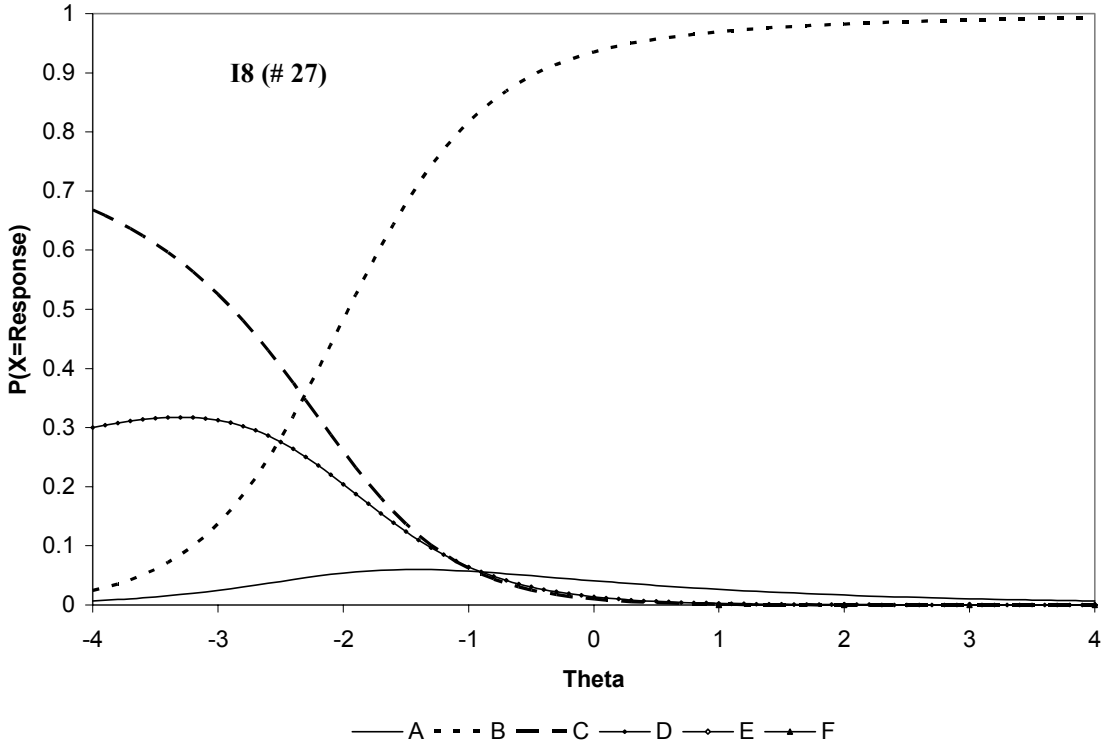

I7 (# 22)
P(X=Response)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-4
-3
-2
-1
0
1
2
3
4
Theta
A
B
C
D
E
F

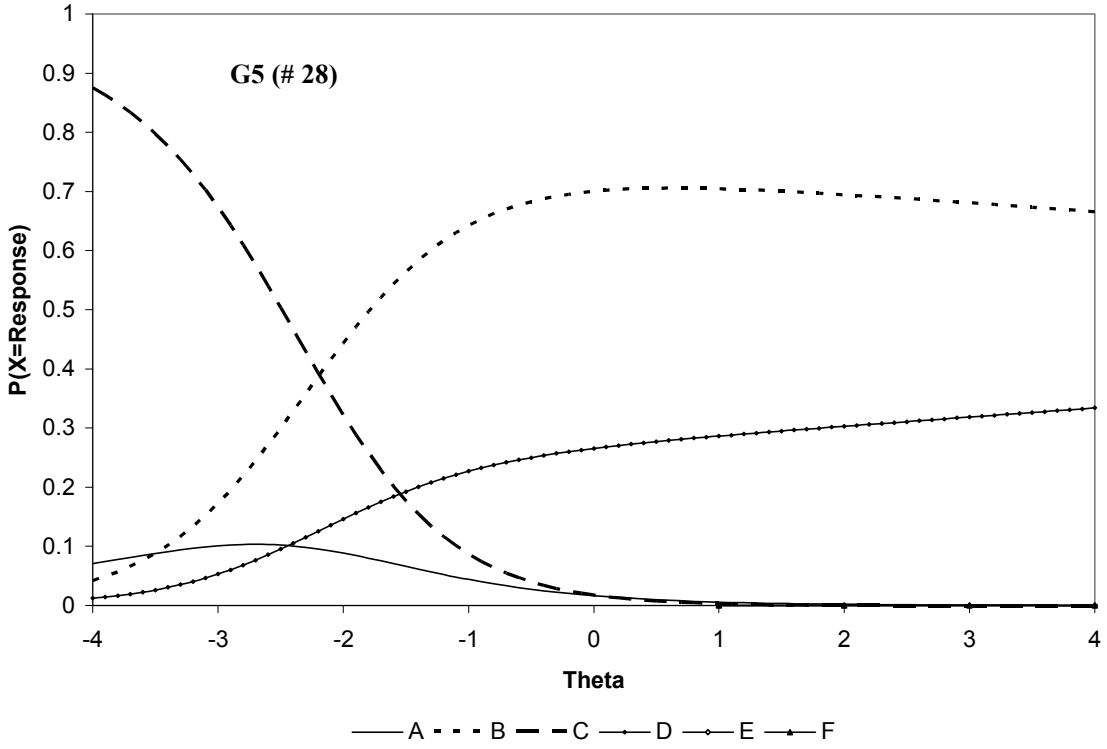

D8 (# 23)
P(X=Response)
Theta
A
B
C
D
E
F

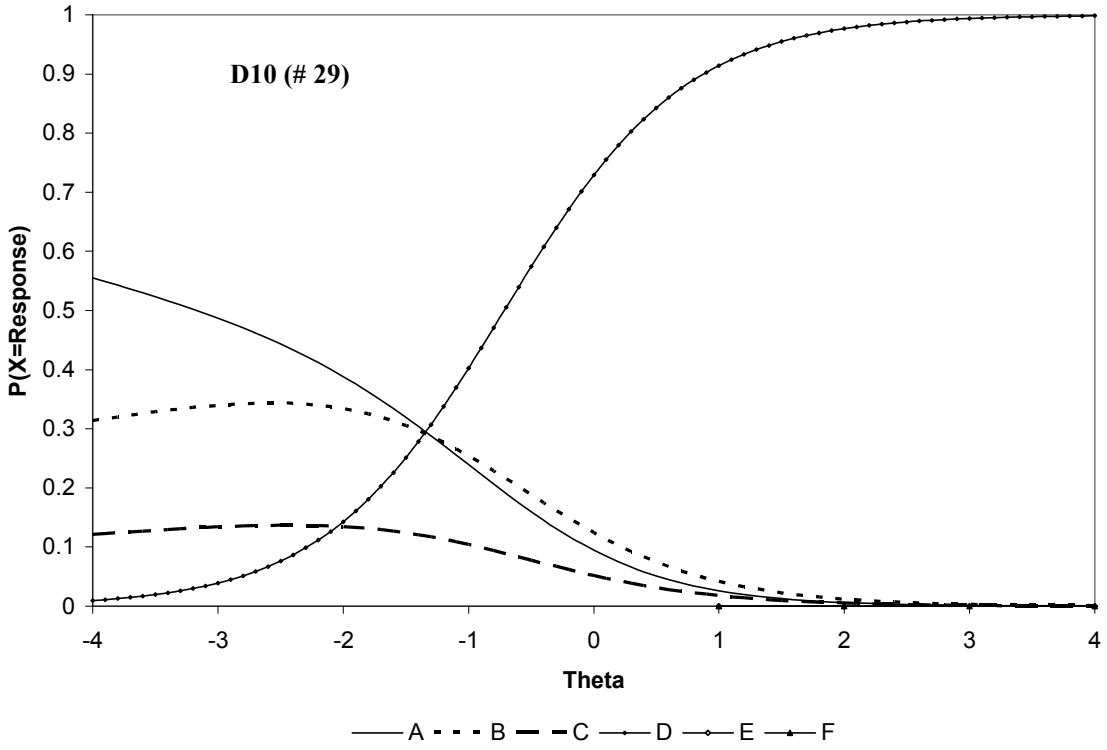

G3 (# 24)
P(X=Response)
Theta
A
B
C
D
E
F

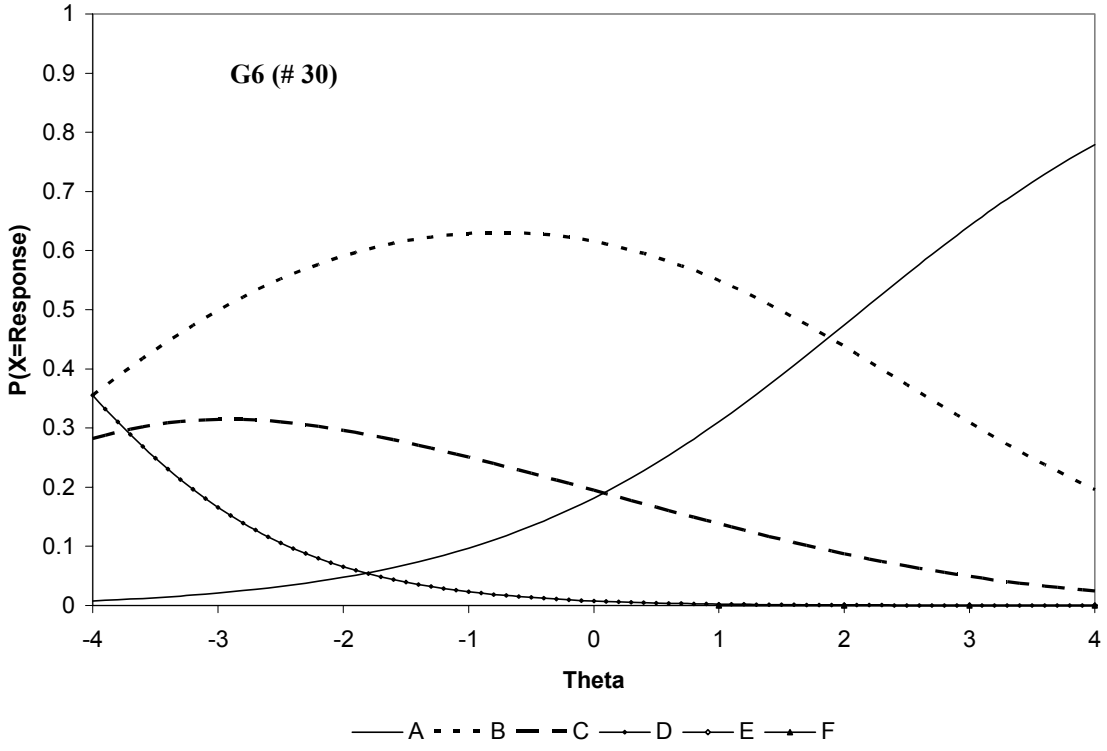

G4 (# 25)
P(X=Response)
Theta
A
B
C
D
E
F

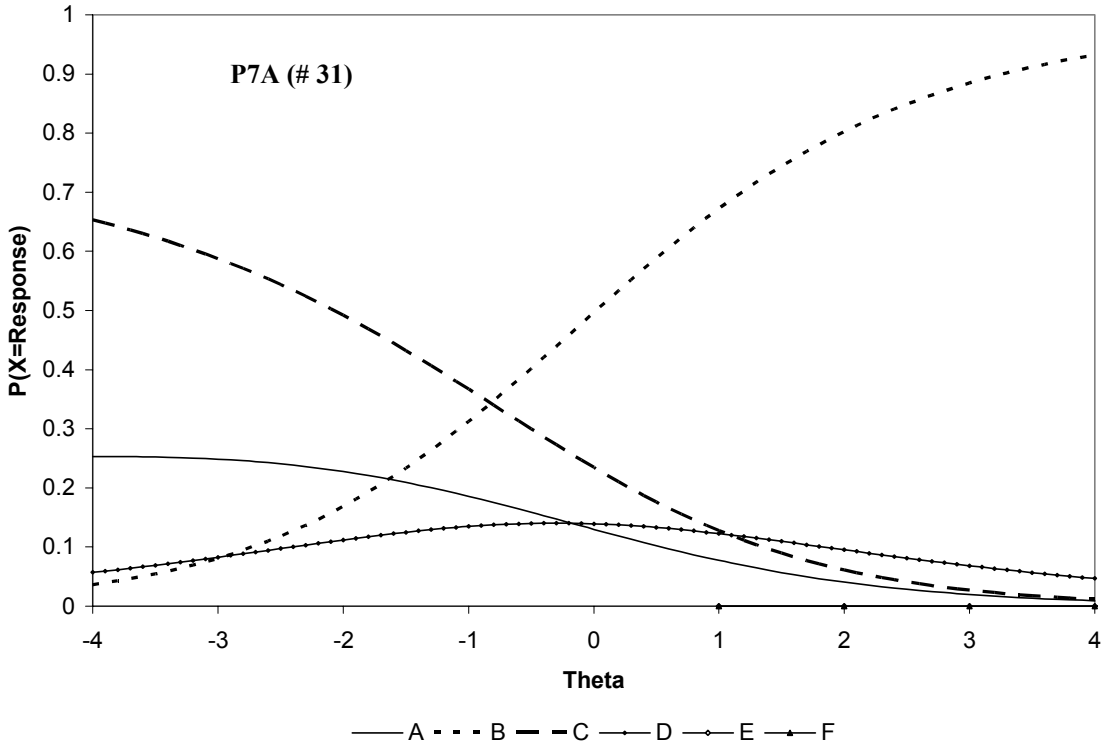

D9 (# 26)
P(X=Response)
Theta
A
B
C
D
E
F

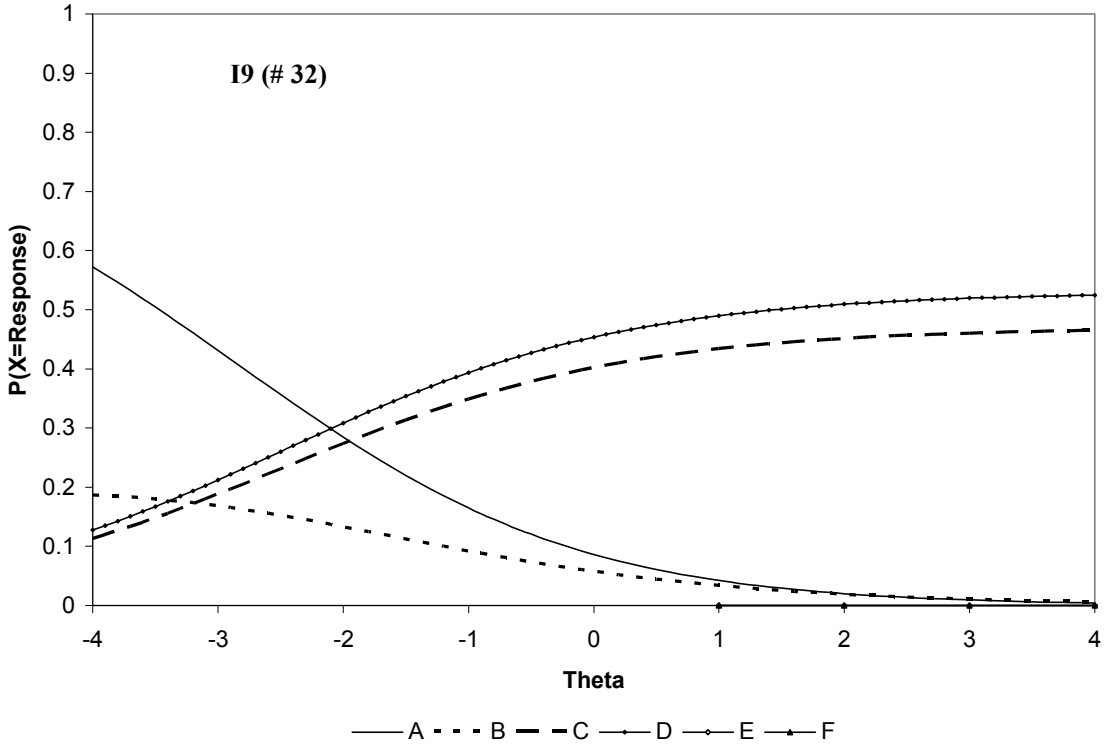

I8 (# 27)
P(X=Response)
Theta
A
B
C
D
E
F

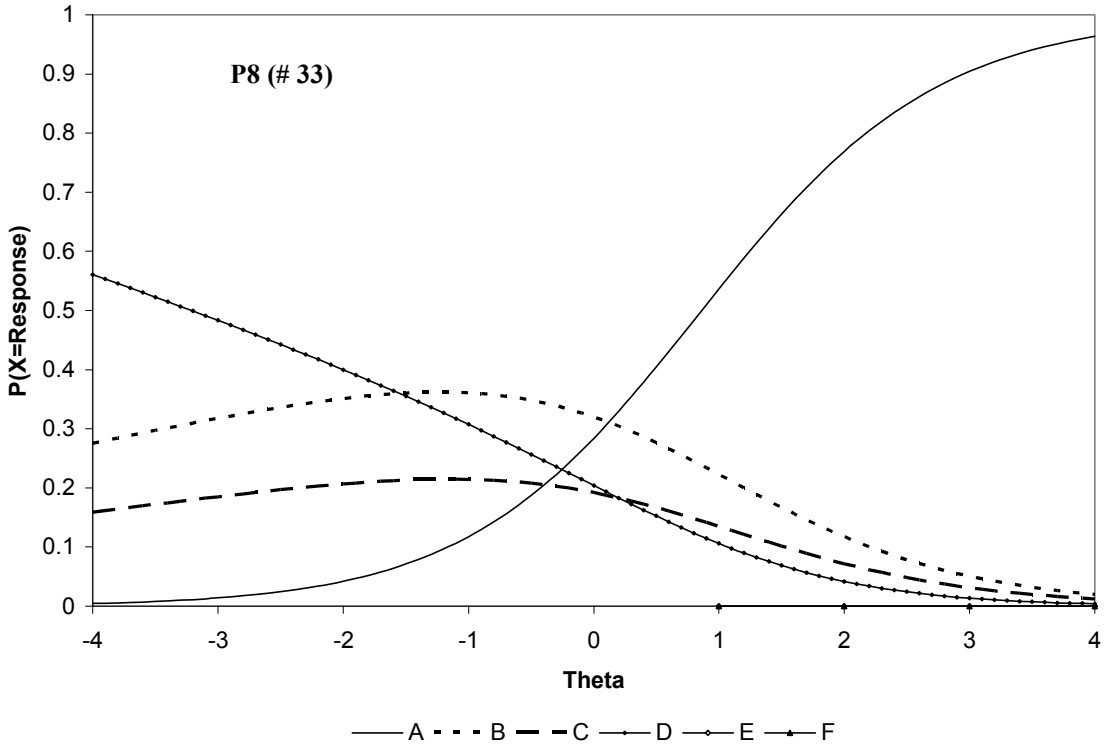

G5 (# 28)
P(X=Response)
Theta
A
B
C
D
E
F

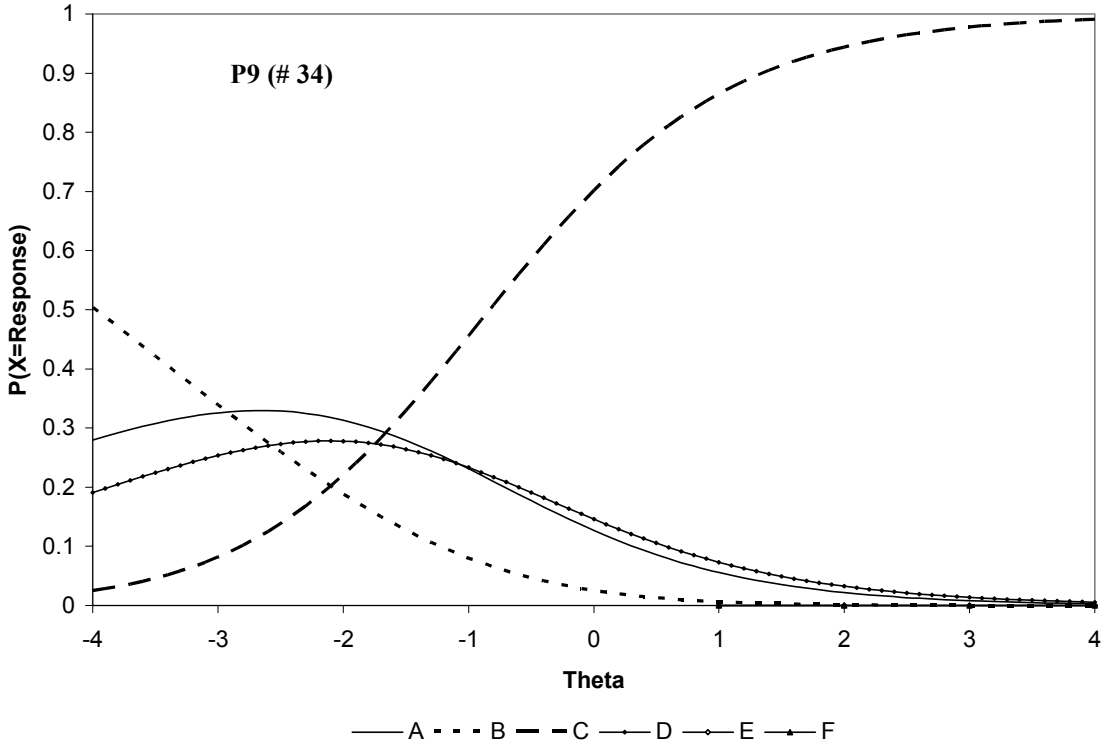

D10 (# 29)
P(X=Response)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-4
-3
-2
-1
0
1
2
3
4
Theta
A
B
C
D
E
F

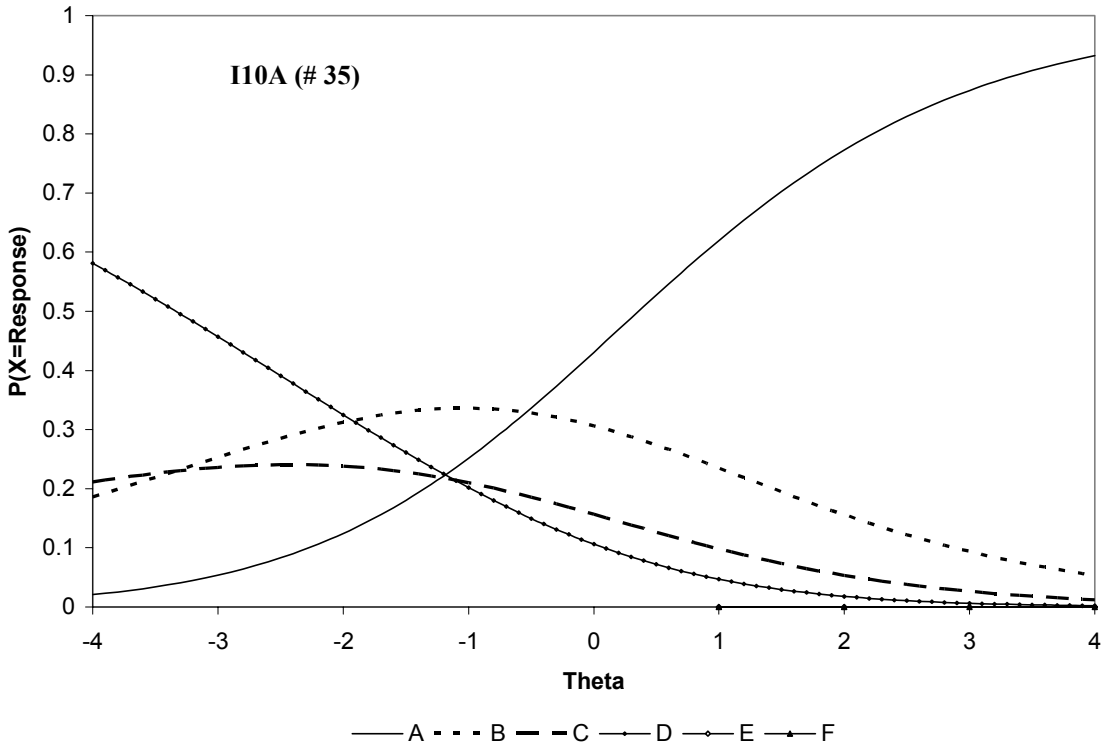

G6 (# 30)
P(X=Response)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-4
-3
-2
-1
0
1
2
3
4
Theta
A
B
C
D
E
F

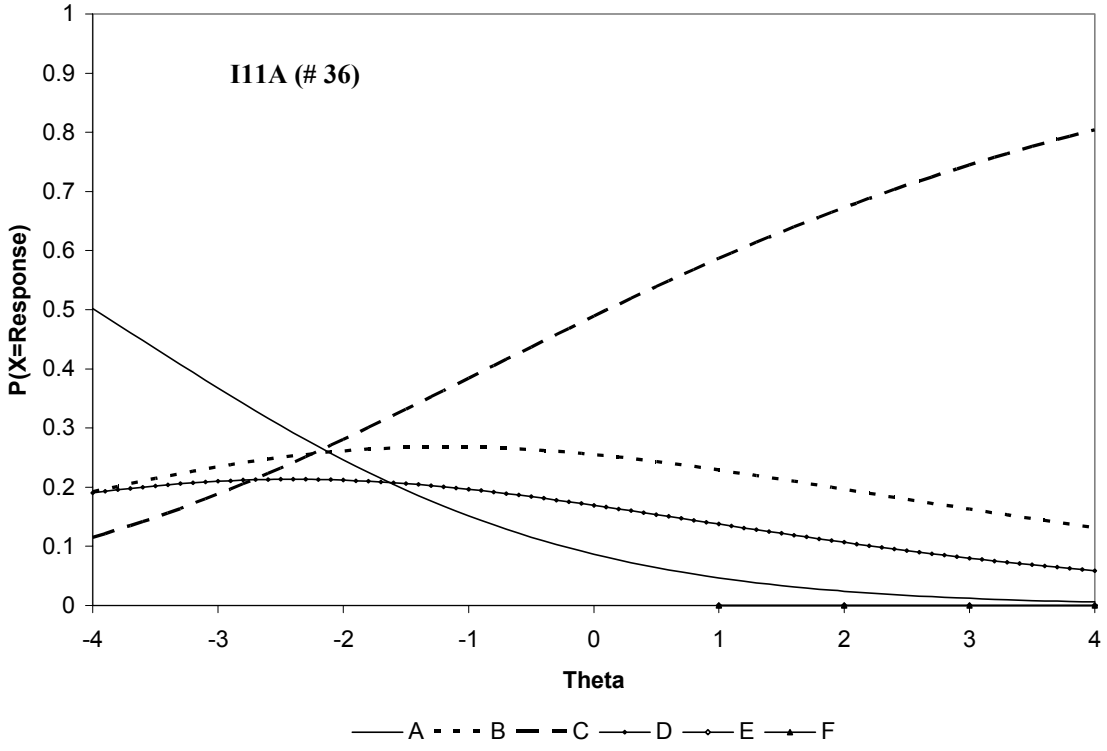

P7A (# 31)
P(X=Response)
Theta
A
B
C
D
E
F

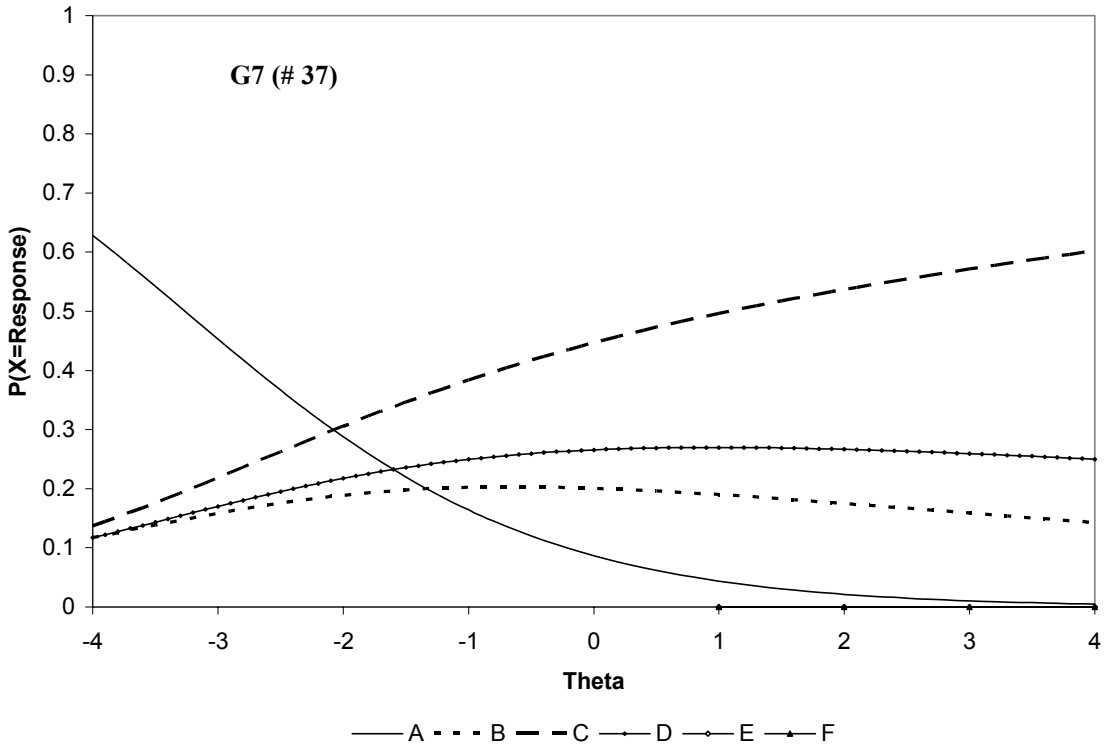

I9 (# 32)
P(X=Response)
Theta
A
B
C
D
E
F

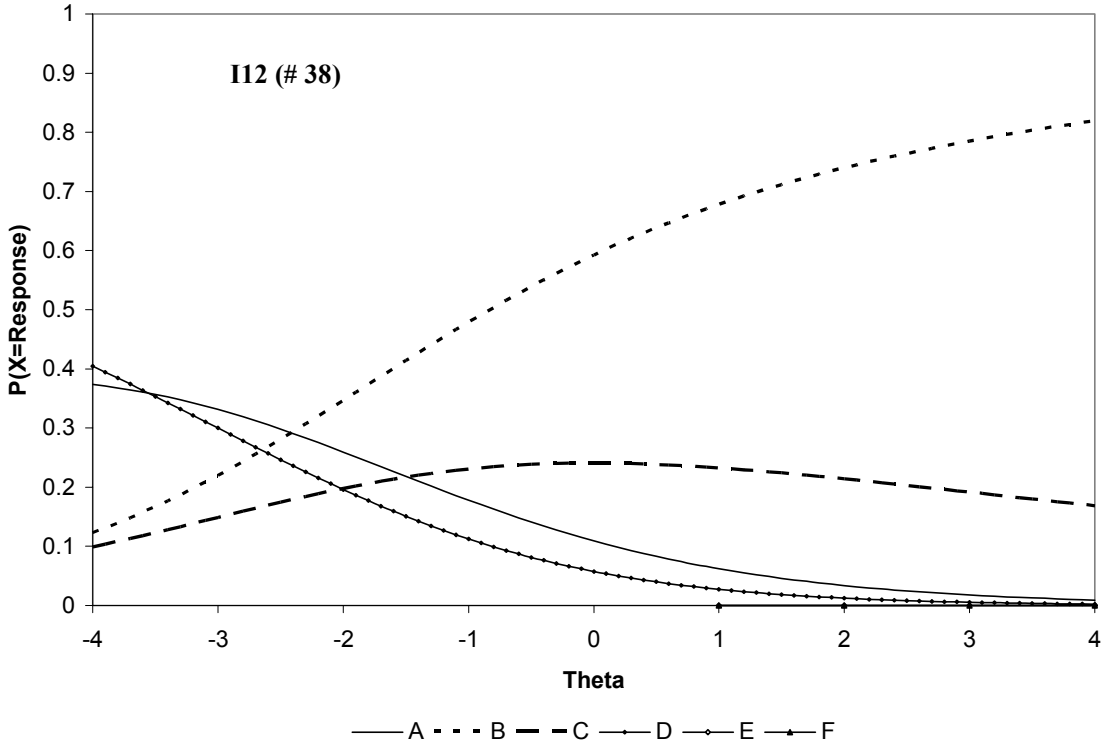

P8 (# 33)
P(X=Response)
Theta
A
B
C
D
E
F


P9 (# 34)
P(X=Response)
Theta
A
B
C
D
E
F


I10A (# 35)
P(X=Response)
Theta
A
B
C
D
E
F


I11A (# 36)
P(X=Response)
Theta
A
B
C
D
E
F


G7 (# 37)
P(X=Response)
Theta
A
B
C
D
E
F


I12 (# 38)
P(X=Response)
Theta
A
B
C
D
E
F