Applying item response theory modeling in educational research

Dai-Trang Le
Iowa State University

Follow this and additional works at: http://lib.dr.iastate.edu/etd
Part of the Instructional Media Design Commons, and the Statistics and Probability Commons

Recommended Citation
Le, Dai-Trang, 'Applying item response theory modeling in educational research' (2013). Graduate Theses and Dissertations. 13410. http://lib.dr.iastate.edu/etd/13410

This Dissertation is brought to you for free and open access by the Graduate College at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Applying item response theory modeling in educational research

by

Dai-Trang Le

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Co-majors: Statistics; Education (Curriculum and Instructional Technology)

Program of Study Committee:
Mack C. Shelley, Co-Major Professor
Ana-Paula Correia, Co-Major Professor
Kenneth Koehler
Heike Hofmann
Ann D. Thompson

Iowa State University
Ames, Iowa
2013
Copyright © Dai-Trang Le, 2013. All rights reserved.
DEDICATION

I would like to dedicate this thesis to my family.
## TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>viii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>xi</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>xii</td>
</tr>
<tr>
<td>CHAPTER 1. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>CHAPTER 2. ANALYZING ATTITUDES TOWARD STUDENT LEARNING: AN APPLICATION OF ITEM RESPONSE THEORY</td>
<td>11</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>12</td>
</tr>
<tr>
<td>2.2 General Overview of IRT</td>
<td>13</td>
</tr>
<tr>
<td>2.2.1 General Form of IRT models</td>
<td>14</td>
</tr>
<tr>
<td>2.2.2 Model Assumptions</td>
<td>17</td>
</tr>
<tr>
<td>2.3 The National School Board Member Survey</td>
<td>18</td>
</tr>
<tr>
<td>2.3.1 Data Description</td>
<td>19</td>
</tr>
<tr>
<td>2.3.2 Identifying Latent Traits</td>
<td>22</td>
</tr>
<tr>
<td>2.4 The Graded Response Model (grm)</td>
<td>24</td>
</tr>
<tr>
<td>2.4.1 Model Formulation and Analysis</td>
<td>25</td>
</tr>
<tr>
<td>2.4.2 Model-Data Fit Assessment</td>
<td>28</td>
</tr>
<tr>
<td>2.4.3 Parameter Estimations</td>
<td>30</td>
</tr>
<tr>
<td>2.5 Analysis and Results</td>
<td>32</td>
</tr>
<tr>
<td>2.5.1 Analysis of the Classical Domain</td>
<td>32</td>
</tr>
<tr>
<td>2.5.2 Analysis of the Innovative Domain</td>
<td>39</td>
</tr>
<tr>
<td>2.6 Discussion and Conclusion</td>
<td>45</td>
</tr>
</tbody>
</table>
CHAPTER 3. STUDENT LEARNING EXPERIENCES RELATED TO USING TECHNOLOGY IN A STATISTICS COURSE

3.1 Introduction

3.1.1 Research Purpose

3.1.2 Research Questions

3.1.3 Research Contributions

3.2 Review of Literature

3.2.1 Teaching and Learning Statistics

3.2.2 Data Visualization Technology in Teaching Statistics

3.2.3 Students’ Attitudes Toward Statistics

3.2.4 Existing Attitude Assessment Instruments

3.3 Study Setting and Methods

3.3.1 Study Setting/Description of the Course

3.3.2 Profile of Study Participants

3.3.3 The Study Instruments

3.3.4 Data Sources and Procedures

3.3.5 Study Variable Measures

3.3.6 Analysis Methods

3.4 Data Analysis

3.4.1 Quantitative Data Analysis

3.4.2 Qualitative Data Analysis of Open-Ended Questions

3.5 Discussion and Conclusion

3.5.1 Students’ Change in Attitude

3.5.2 Limitations

3.5.3 Conclusion

CHAPTER 4. BRINGING DATA TO LIFE INTO AN INTRODUCTORY STATISTICS COURSE WITH GAPMINDER

4.1 Introduction

4.2 Brief History of Gapminder and its co-founder Hans Rosling
4.3 Conclusion ............................................................................................................. 136

CHAPTER 5. SYNTHESIS AND CONCLUSION ................................................. 138
  5.1 Synthesis and Conclusion ............................................................................. 138
  5.2 Limitations and Suggested Resolutions ...................................................... 139
  5.3 Implications for Future Research ................................................................. 141

APPENDIX A. UNIDIMENSIONAL IRT MODELS ........................................... 143
  A.1 Models for Items with Two Response Categories .................................... 143
     A.1.1 One-Parameter Logistic Model (Rasch model) .................................. 143
     A.1.2 Two-Parameter Logistic Model (2PL model) .................................... 145
     A.1.3 Three-Parameter Logistic Model (3PL model) .................................. 147
  A.2 Models for Items with More Than Two Response Categories .................... 149
     A.2.1 The graded response model (grm) .................................................. 150
  A.3 Other Descriptive Statistics for Items and Test/Instrument ....................... 154
     A.3.1 Expected Function - The Test Characteristic Curve ......................... 154
     A.3.2 Item Information and Test Information Functions ............................. 155
     A.3.3 Standard Errors of Measurement ................................................... 158

APPENDIX B. PARAMETER ESTIMATION FOR IRT MODELS ................ 161
  B.1 Parameter Estimations for Dichotomous Items ......................................... 161
  B.2 Parameter Estimations for Polytomous Items ........................................... 164

APPENDIX C. SURVEY WEBSITES ................................................................. 169
  C.1 Pre-Survey On Attitude and Engagement in Statistics for Statistics 401D-XW . 169
  C.2 Post-Survey (Six-Week) survey on Attitude and Engagement in Statistics for
     Statistics 401D-XW ....................................................................................... 179
  C.3 Survey of Attitude Toward Statistics (SATS) .......................................... 186
  C.4 National Survey of Student Engagement (NSSE) ................................... 186

APPENDIX D. GAPMINDER PROJECT ACTIVITIES ..................................... 187
  D.1 Gapminder Project Website ....................................................................... 187
LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Response frequencies for question 21</td>
<td>21</td>
</tr>
<tr>
<td>2.2</td>
<td>Proportions of the responses to the 11 items in question 21</td>
<td>21</td>
</tr>
<tr>
<td>2.3</td>
<td>Significance tests of Kendall’s correlation</td>
<td>22</td>
</tr>
<tr>
<td>2.4</td>
<td>Factor loadings from ordinal factor analysis</td>
<td>24</td>
</tr>
<tr>
<td>2.5</td>
<td>Table of item category thresholds</td>
<td>28</td>
</tr>
<tr>
<td>2.6</td>
<td>Summary of estimated item parameter for the <em>classical</em> domain</td>
<td>34</td>
</tr>
<tr>
<td>2.7</td>
<td>Summary of item parameter estimates for the <em>Innovative</em> domain</td>
<td>43</td>
</tr>
<tr>
<td>3.1</td>
<td>Descriptions of the six subscales of the <em>attitude</em> scale</td>
<td>70</td>
</tr>
<tr>
<td>3.2</td>
<td>Demographic questions in the pre- and post-surveys</td>
<td>71</td>
</tr>
<tr>
<td>3.3</td>
<td>Questions on <em>attitude</em> in the pre-survey</td>
<td>73</td>
</tr>
<tr>
<td>3.4</td>
<td>Questions on <em>attitude</em> in the post-survey</td>
<td>74</td>
</tr>
<tr>
<td>3.5</td>
<td>Questions on technology in the pre-survey</td>
<td>75</td>
</tr>
<tr>
<td>3.6</td>
<td>Questions on technology in the post-survey</td>
<td>76</td>
</tr>
<tr>
<td>3.7</td>
<td>Pre- and post-survey <em>perfectly-matched</em> question set</td>
<td>79</td>
</tr>
<tr>
<td>3.8</td>
<td>Table of nine post-survey items in <em>question set 2</em></td>
<td>79</td>
</tr>
<tr>
<td>3.9</td>
<td>Estimates of item parameters (category thresholds, item locations, and unique discriminations or slopes) for pre-survey</td>
<td>84</td>
</tr>
<tr>
<td>3.10</td>
<td>Estimates of item parameters for the post-survey</td>
<td>85</td>
</tr>
<tr>
<td>3.11</td>
<td>Table of proportions of the categories for two items on <em>difficulty</em> level</td>
<td>95</td>
</tr>
<tr>
<td>3.12</td>
<td>Table of proportions of responses for items in question set 2</td>
<td>95</td>
</tr>
<tr>
<td>3.13</td>
<td>Table of descriptive statistics for question set 2</td>
<td>97</td>
</tr>
<tr>
<td>3.14</td>
<td>Table of distribution of the category threshold levels</td>
<td>99</td>
</tr>
<tr>
<td>Section</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>3.15</td>
<td>Item locations = item easiness = item difficulty</td>
<td>100</td>
</tr>
<tr>
<td>3.16</td>
<td>Content and common themes of question 10 in pre- and post-survey.</td>
<td>104</td>
</tr>
<tr>
<td>3.17</td>
<td>Contents and summary of question 11 in the pre- and post-survey</td>
<td>106</td>
</tr>
<tr>
<td>3.18</td>
<td>Summaries of attitude scores for both pre- and post-surveys</td>
<td>114</td>
</tr>
<tr>
<td>3.19</td>
<td>Item locations listed in order of easiness to difficulty</td>
<td>115</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

2.1 Plot of item and person locations ........................................ 14
2.2 Plot of item characteristic curves in an IRT model .................... 16
2.3 Bar graph of gender within four school enrollment categories. ...... 20
2.4 Scree plot of eigenvalues ....................................................... 23
2.5 The OCC and ICC curves for polytomous model ....................... 27
2.6 Item information curves. ....................................................... 29
2.7 Plot of distribution of latent scores ....................................... 32
2.8 Plot of the operating characteristic curves for the classical domain. . 35
2.9 Plot of the item characteristic curves for the classical domain. ...... 36
2.10 Plot of distribution, histogram, and QQ plot of latent attitude scores . 37
2.11 Plot of test and item information curves for the classical domain . . 38
2.12 Plot of test information and standard error curves for the Classical domain 39
2.13 Plot of ICC curves for the innovative domain .......................... 40
2.14 Plot of OCC curves for the innovative domain .......................... 41
2.15 Plot of latent attitude scores for the innovative domain ............... 44
2.16 Plot of test information and standard error curves for the innovative domain .................................................. 44
3.1 Picture of different information visualization tools. .................... 58
3.2 Schema of process of interpretation and understanding of information via data visualization. ............................................... 59
3.3 The two domains of learning, affective and cognitive. .................. 65
3.4 Display of scree plot, eigenvalues, and confirmatory factor analysis results ........................................ 81
3.5 Plot of information curves from a graded response model ................................................................. 82
3.6 Plot of the OCC curves for the four items in the pre-survey ............................................................. 86
3.7 Plot of the OCC curves for the four items in the post-survey ............................................................ 87
3.8 Plots of the category characteristic curves for the pre-survey .......................................................... 88
3.9 Plot of the category characteristic curves for the post-survey ......................................................... 89
3.10 Item characteristics curves for perfectly-matched items in the pre- and post-surveys. .......................... 90
3.11 Summary tables and distribution of latent attitude scores ................................................................. 92
3.12 Item information curves for pre- and post-surveys ............................................................................ 93
3.13 Instrument information curves versus the standard error of measurement curves ............................. 94
3.14 Scree plot, eigenvalues, and factor loadings ....................................................................................... 98
3.15 Plot of information curves for items in question set 2. ..................................................................... 101
3.16 Distribution and summary of attitude scores for question set 2. ....................................................... 101
3.17 The word cloud for the responses to question 16: list three useful items... ................................. 107
4.1 Bubble chart of breast cancer versus income per person ................................................................... 124
4.2 Picture of the video: “200 years that changed the world” ................................................................. 125
4.3 Picture of video “The Joy of Statistics” ............................................................................................... 126
4.4 Chart created with Google bubble chart ............................................................................................ 127
4.5 Plot of number of internet users versus income per person .............................................................. 128
4.6 Exploring the world with Gapminder ................................................................................................. 131
4.7 A view of Dollar Street software ....................................................................................................... 132
4.8 Gapminder World Map 2010. ............................................................................................................. 133
4.9 Plot of number of football players versus FIFA ranking ................................................................. 134
4.10 Graph of countries in the 2010 World Cup and their income level ............................................... 135
A.1 Plot of item characteristic curves (ICC) in a Rasch model .............................................................. 144
| A.2 | Plot of ICC with unequal slopes | 147 |
| A.3 | Plot of ICC with varying difficulty levels | 148 |
| A.4 | Plot of ICC with guessing parameter | 149 |
| A.5 | The OCC and ICC curves with threshold lines | 153 |
| A.6 | Plot of OCC of an with low discriminating power | 154 |
| A.7 | Plot of the expected function of test score | 155 |
| A.8 | Plot of item and test information functions | 157 |
| A.9 | Plot of test information function and standard error curve | 159 |
ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting research and the writing of this thesis. First and foremost, Dr. Shelley for his guidance, patience and support throughout this research. Dr. Kaiser for his insight, inspirational advice, comments and conversation. I have learned so much from him in a short time. He has modeled how to advise students and assist them in discovering their strengths and weaknesses which I will carry with me into my teaching career. I also sincerely express my appreciation to Dr. Andre for agreeing to substitute on behalf of Dr. Thompson. I thank my committee members, Drs. Koehler, Hofmann, and Thompson, for their time, efforts, and contributions. A special thanks to Dr. Correia for her guidance throughout the design and writing of my second research study.
ABSTRACT

The analysis of test-generated data or survey responses often utilizes means, total scores, or latent scores from factor analysis as outcome variables in a regression-based procedure. These scoring methods, however, do not take into account the characteristics of the items in the test or survey. Item response theory is different in this sense. It models the relationship between a respondent’s trait level (ability, attitude,...) and the pattern of item responses. Thus estimation of individual latent traits can differ even for two individuals with the same total scores. IRT scores can yield additional benefits that will be discussed in detail. In this thesis, we illustrate the use of Unidimensional IRT to analyze polytomous and dichotomous item response data respectively. Applications of IRT were illustrated through three research studies which include (1) Analyzing data toward students learning; (2) Student Learning Experience Related To Using Technology in Statistics Course, and (3) Bringin Data to Life in an Introductor Statistics Course with Gapminder.
CHAPTER 1. INTRODUCTION

This chapter presents a brief history of the development and applications of Item Response Theory (IRT). It also includes a comparison between two frameworks, Classical Test Theory (CTT) and the so-called modern test theory, IRT. The last section provides the organization of the remaining chapters of my dissertation and the implications of the research.

IRT Development and Applications: Past, Present, and Future

Item response theory (IRT) is a set of latent variable techniques especially designed to model the interaction between a subject’s “ability” and the item level stimuli (difficulty, guessing, etc.) (Chalmers, 2012). The focus is on the pattern of responses rather than on composite or total score variables and linear regression theory. The IRT framework emphasizes how responses can be thought of in probabilistic terms. In IRT the item responses are considered the outcome (dependent) variables, and the examinee’s ability and the items’ characteristics are the latent predictor (independent) variables.

Classical test theory (CTT) was the dominant approach until 1953 when Frederic Lord published his doctoral dissertation on Latent Trait Theory. While CTT models test outcomes based on the linear relationship between true and observed score (Observed score = True Score + Error), IRT models the probability of a response pattern of an examinee as a function of the person’s ability and the characteristics of the items in a test or survey.

Interest in Lord’s work (1953) spread quickly, as evidenced in the increase in publications in this area. Allen Birnbaum wrote a series of technical reports on logistic test models and model parameter estimation in 1957 and 1958. George Rasch (1960) published his book proposing several models for item responses. In the 1960s more work in this area was contributed by
Baker (1961) on empirical comparison between logistic and normal ogive functions, while Lord and Novick (1968), and Wright (1968) worked on dichotomous models. Samejima proposed the graded response model for polytomous in 1969. This group of scholars brought considerable attention to the field. Through the 1970 and 1980s, a new group of scholars surfaced including Aldrich (1978), Anderson (1977, 1980), Hambleton and Swaminathan (1986), Wright and Stone (1979), Swaminathan and Rogers (1981), and Harris (1989). Journals such as *Applied Psychological Measurement* and *The Journal of Educational Measurement* were filled with technical advances and applications of IRT. These scholars continued to make significant contributions to the field.

In the 2000s the IRT field was promoted by a new wave of researchers who not only expanded the technical aspects of the framework (estimation, model identification, and goodness of fit), but also advanced its computational aspects. The extensive study of IRT during the past 50 years was manifested in a rise in the number of software packages designed for analyzing item response data from surveys or tests. Various IRT commercial software was also created such as BILOG, MULTILOG, WINSTEPS, IRTPRO, MPLUS, and HLM, to name just a few. More importantly, a number of IRT packages developed in the open source R software to estimate various IRT models also appeared and gained recognition. These included the packages *ltm* for unidimensional IRT (Rizopoulos, 2006), *eRm* for extended Rasch models (Mair & Hatzinger, 2007), *mirt* for multilevel and Bayesian estimation of some IRT models (Fox, 2007), *gpcm* (Johnson, 2007) for a Bayesian estimation of the generalized partial credit model, *MCMCpack* for Bayesian IRT (Martin, Quinn, & Park, 2011), and *mirt* for multidimensional IRT (Chalmers, 2012). De Boeck (2008) and Wilson (2008) made use of the general statistics package *lme4* and incorporated Rasch models under the generalized linear mixed model framework. This makes it possible to use SAS PROC NLMIXED (SAS Institute Inc.) for IRT.

There are two primary branches of IRT, unidimensional and multidimensional. Unidimensional IRT (UIRT) operates under the assumption that all items in an instrument share a common strand. Under the UIRT assumption, there exists an underlying idea or message threading through the content of all items, and test or survey instruments are developed to measure this unique construct. Unlike UIRT, Multidimensional IRT (MIRT) assumes that
items are grouped or clustered in different domains, and instruments are designed to measure these multiple domain constructs. Moreover, not only are items in similar domains related but respondents are related due to being clustered in similar settings such as classrooms and schools. This hierarchical structure is integrated into MIRT to form multilevel MIRT. De Boeck and Wilson (2004)’s edited volume *Explanatory Item Response Models* discussed this integrated model from the generalized linear mixed models (GLMM) framework. They differentiated the descriptive measurement approach for scoring tests/surveys and the explanatory analysis approach for modeling. In essence, a multilevel or random effects model is estimated when the model’s first level is the item measurement model.

In review, IRT research has evolved from unidimensional modeling of item responses and the measurement of person latent traits, to multidimensional analysis and it has come to be viewed a special class of generalized linear mixed model. Research on IRT now is concentrated in multidimensional models that are capable of describing the multiple traits and multiple dimensions inherent in persons and tests, respectively. Active researchers in the field include, but are not limited to, Reckase (2009), DeBoeck and Wilson (2010), Bates (2008), Chalmers (2012), and DeMars (2010). The latest development of IRT has shifted its applications from being a tool set used exclusively by behavioral scientists and psychometricians to becoming data analysis tools for a wide set of applications used by statisticians.

So what lies ahead for IRT researchers? It has been speculated that IRT methodology will be incorporated and overlap with other frameworks such as structural equation modeling and factor analysis. There may be new development in models and estimation methods as well as computer software to accommodate these advances.

In summary, IRT has been an extremely active area of research for more than half a century. The combination of methodological advances and increasingly powerful software has increased applicability and interest. IRT is widely used in assessment and evaluation research to describe the interaction between examinees and test questions. For many years CTT remained the dominant framework used in education despite the development and progress of IRT. Currently IRT is finding widespread application in the engineering of large-scale assessments as well as on a smaller scale in sociological and psychological assessments. Other applications of IRT besides
assessment include: scaling, which involves instrument developments and refinement; equating, which involves methodology for comparison between different tests and creating test banks; computerized adaptive testing, to provide optimal measurement of a person’s true proficiency or trait; and differential item functioning, designed for group comparisons based on external or non-construct-related factors, and many more. Future development of the IRT framework will see it being overlapped and integrated with different statistical frameworks.

**Classical Test Theory versus Item Response Theory**

In comparison to classical test theory (CTT), item response theory (IRT) is considered as the standard, if not preferred, method for conducting psychometric evaluations of new and established measures (Osteen 2010). However, although IRT has been studied for the past 50 years, CTT has still been researched and applied continuously. Many testing programs still implement CTT in their design and assessment of test results. This is due to some advantages of CTT over IRT. For example, CTT describes the relationship between the true score and observed score in a linear fashion which makes CTT’s models easy to understand and apply for many researchers. It is based entirely on total scores or number-of-correct-answer scores. An examinee’s observed score is the total score obtained by each examinee and it is different from the true score by a common error score. This methodology of scoring has generated a number of advantages as well as limitations. The first advantage of CTT is that the analyses require smaller sample sizes than does IRT. Second, CTT mathematical procedures are much simpler compared to IRT, as the models in CTT are linear while IRT’s models are nonlinear. Third, model parameter estimation in CTT is conceptually straightforward and requires minimum assumptions, making models useful and widely applicable. Fourth, analyses do not require strict goodness of fit studies as in IRT.

However, CTT has a few major drawbacks. The cornerstone of many CTT analyses is the characteristics of the test items difficulty and reliability. These indices are measured by the item’s proportion, \( p \), of examinees who answer the question correctly and the item-total correlation, \( r \). However, the indices are not constant as they are entirely dependent on the sample of examinees from whom they are obtained. They cannot be used to indicate the
characteristic or quality of a test. Another drawback is that the examinees’ scores are test-dependent. That is, examinees may obtain higher scores on an easier test and lower scores on a harder test, and thus no true score can be extracted. This does not allow a basis for matching test items and ability levels.

In this sense, IRT has major benefits over CTT. In the IRT framework, item characteristics are sample-independent and a person’s latent scores are test-independent provided that the selected models fit the data well. Thus, scores that describe examinee proficiency are not dependent on test difficulty. Their scores may be lower on more difficult tests and higher on easier tests, but their ability scores remain constant over any test at the time of testing or surveying. IRT also permits calculation of the probability of a particular respondent selecting a category on a test item. Sample-independent test items facilitate design of computer adaptive tests, which allows for more accurate comparison or identification of examinees. Moreover, IRT can be used for scale refinement or development, as it is capable of the calculation of standard errors and therefore provides information on the quality of each item. This aids with making decisions in selecting items to exclude or include in a test or survey instrument. In addition, items are also selected based on their difficulty and discrimination indices, i.e., their capability of discriminating low and high latent trait groups.

Despite those advantages, IRT models have their shortcomings, too. On the technical side, the models are more complex and the parameter estimation methods often involve complicated numerical methods. Latent traits as well as item parameters can also be difficult to interpret both graphically and numerically. Rasch models (one-parameter models) are more straightforward to apply than other IRT models. However, because of the restrictions imposed by model assumptions, problems may arise with the fit of the Rasch models to testing data. Consequently, despite some advantages of IRT over CTT, the rise in popularity of IRT, and the accommodation of many computer packages, the IRT methodology presents a number of unique challenges. For one, specification of models in IRT is more complicated compared to CTT. Second, proper interpretation of results requires careful attention to several unique features of IRT graphical and numerical outputs, especially for more challenging models. Third, data processing and preparation for use with IRT software can be challenging and time consuming.
Organization of the dissertation

The rest of the dissertation is organized as a collection of four studies as follows.

The first research study is “Analyzing Attitudes Toward Student Learning: An Application of Unidimensional Item Response Theory.” The purpose of this study is to provide an overview of the IRT framework and to illustrate applications of IRT in analyzing survey data in an educational setting. General assumptions and characteristics of IRT models such as, unidimensionality, local independence, item response functions, and the general framework of estimating model parameters, are also discussed. A blend of theory and application makes the introduction of IRT both rigorous and informative. This study has been accepted as a book chapter titled *Item Response Analysis* in “Assessment & Evaluation Methods in Medical Education” in Turkey in June, 2013.

The second research study, “Student Learning Experiences Related to Using Technology in a Statistics Course,” studies the use of visualization technology in teaching statistics. The study examines the influence of technology in enhancing student learning experiences and attitudes toward statistics. A manuscript will be sent to the journal “Technology Innovation in Teaching Statistics” (TISE).

The third research study, “Bringing Data to Life into an Introductory Statistics Course with Gapminder,” is an article published in the journal *Teaching Statistics* in June 2013. The article illustrates how Gapminder can be used to turn mundane data to life and showcases the application of statistics in social development.

Implications for Research

The implications of this research are two-fold. First, it illustrates the application of IRT in analyzing survey survey data in educational settings. Second, it emphasizes the importance of implementing data visualization and web-based technology in teaching service statistics courses at the graduate level.
BIBLIOGRAPHY


CHAPTER 2. ANALYZING ATTITUDES TOWARD STUDENT LEARNING: AN APPLICATION OF ITEM RESPONSE THEORY

Dai-Trang Le, Mark Kaiser, & Mack Shelley

This study has been accepted as a book chapter, Item Response Analysis, in Assessment & Evaluation Methods in Medical Education.

Abstract

In an effort to understand how school boards in America’s K-12 school system function, a research collaboration was undertaken among four agencies: the National School Boards Association, the Thomas B. Fordham Institute, the Iowa Association of School Boards, and the Wallace Foundation. These groups joined effort to conduct research on school boards. As a result, a National Survey of School Board Members was conducted in 2009. The focus was on understanding the critical role of school board members as participants on the governance bodies of public school systems and their attitudes on issues related to student achievement.

This study provides a basic overview of item response theory (IRT) analysis for assessing a latent factor structure. The objective was to evaluate an underlying variable related to an issue in a national survey of school boards. We analyzed one question from the survey to gain an understanding of the attitudes of K-12 school board members towards factors related to students’ learning achievement. Utilizing a graded response model in IRT we estimated the underlying variable, attitude\(^1\), of the respondents based on their patterns of responses. Discussion includes methodology of IRT, analysis findings, and implications of IRT for educational research.

**Keywords:** Item Response Theory, unidimensional IRT, school boards, latent trait.

\(^1\)Attitude here is defined as an expression of favor or disfavor toward an educational approach.
2.1 Introduction

Item Response Theory (IRT) models are commonly used to model the latent traits associated with a set of items or questions in a test or survey survey. In education, testing is an inherent part of the curriculum as an assessment tool to evaluate students’ subject matter proficiency and skill development. Apart from viewing the total score as an indicator of performance one may wish to understand whether the testing instrument is adequately designed to measure particular aspects of the knowledge and skills of respondents. IRT attempts to simultaneously examine the appropriateness of the questions in terms of measuring what they are designed to measure and the proficiency of the respondents.

IRT models describe the interactions of persons and test items (Reckarse, 2009). Hence, IRT is a general framework for specifying mathematical functions that characterize the relationship between a person’s ability or trait as measured by an instrument and the person’s responses to the separate items in the instrument (DeMars, 2010). In educational testing, IRT offers an alternative to classical test theory, which depends on total scores or number correct as outcome variables. IRT models have become a popular framework in many fields including psychology, nursing, and public health. But IRT-based analyses are still scarce in the social sciences and in the educational research literature.

The focus of this study is on the application of IRT to a problem that involves the assessment of educators’ attitudes about factors involved in improving student learning. The data for this analysis come from the National Survey of School Board Members conducted in 2009. The key research question was: How were the attitudes of the board members related to different approaches in education?

The rest of the study is organized as follows. Section 2 presents an overview of the IRT models within the context of an educational testing situation in which the objective is to assess both individual ability and question difficulty. This is the classical setting for application of IRT methods. Section 3 introduces the National Survey of School Board Members and contains a discussion of how this problem can be considered within the IRT framework. In Section 4 the specific model used is developed, and issues involved with estimation and model assessment
(goodness of fit) are discussed. Results of the analysis are presented in Section 5. Section 5 contains concluding remarks. More technical details on the formulation of IRT models and parameter estimation are presented in Appendices A and B.

### 2.2 General Overview of IRT

The development of IRT modeling has a long history and extensive literature. In this section, we provide a brief overview of some popular IRT models and their assumptions. Thorough discussion of IRT can be found in Baker and Kim (2004), Bock (1997), and van der Linden and Hambleton (1997).

IRT is an approach to modern educational and psychological measurement which addresses the measurement of a hypothetical latent construct such as ability or attitude. These latent traits cannot be measured directly on individuals and must be quantified via responses to items or questions in a test or survey.

IRT methods are commonly used to obtain latent scores for individual respondents on qualities such as trait, ability, proficiency, or attitude in a test or survey. IRT is perhaps most easily understood in terms of the latent trait ability in a testing situation. In fact, the first applications of IRT were in educational testing. The IRT scoring process takes into account respondents’ latent variable and items’ characteristics such as difficulty and discrimination. Applications of IRT are used in many fields such as psychometrics, educational sciences, sociology, health professional fields, and computer adaptive testing (CAT). Additionally, IRT can be used in test or instrument development because the IRT models utilize the information on the items’ characteristics to evaluate and refine an instrument.

IRT models, in contrast to classical test theory (CTT), do not rely on sums or number-correct scores to evaluate a person’s performance, nor do they assume equal contribution of the items (questions) to the overall scores. Since items vary in their difficulty and persons vary in their trait level, this method may result in a more accurate assessment of respondents’ latent traits because respondents with the same sum score may differ in their trait measurement. IRT methods also use the same metric to measure the latent variable and the items’ difficulty levels, thereby facilitating comparison and estimation of parameters. Figure 2.1 displays the
placement of item location and person trait level on the same scale. If the difference between a person’s location (ability) and an item’s location (difficulty level) is positive, the person has a high chance (greater than 50%) of answering that item correctly, or endorsing that item with a positive score. Otherwise, if the difference is negative then the person’s chance of getting an incorrect answer is higher.

Figure 2.1 Figure depicts the locations of the items and individuals on the same continuum of the latent trait. Person’s location is the measure of the person’s latent trait and item’s location is the measure of the item’s difficulty. When the two locations coincide, the person is expected to answer the item correctly or positively with a 50:50 chance.

Estimation of items’ and respondents’ characteristics can be performed separately or simultaneously, depending on the method of estimation used. Fox (2010) states that “estimates of persons’ characteristic from different sets of items measuring the same underlying constructs are comparable and different only due to measurement error” and that “estimates of item characteristics from responses from different groups of respondents in the same population are comparable and differ only due to sampling error.” While this may be an overly optimistic view of what IRT (and statistical models in general) can accomplish, it does represent the ideal underlying model formulation.

2.2.1 General Form of IRT models

IRT includes a set of models that describe the interactions between a person and the test items. Persons may possess different traits and instruments may be designed to measure more than one trait. In this study we only discuss IRT models that describe one single trait. These models are often referred to as unidimensional IRT (UIRT).
Consider an educational testing situation in which \( n \) individuals answer \( I \) questions or items. For \( j = 1, \ldots, n \) and \( i = 1, \ldots, I \), let \( Y_{ij} \) be random variables associated with the response of individual \( j \) to item \( i \). These responses may be binary (e.g., correct/incorrect) or may be discrete with a number of categories. Let \( \Omega_Y \) denote the set of possible values of the \( Y_{ij} \), assumed to be identical for each item in the test. Let \( \theta_j \) denote the latent trait of ability for individual \( j \), and let \( \eta_i \) denote a set of parameters that will be used to model item (question) characteristics. Different IRT models arise from different sets of possible responses \( \Omega_Y \) and different functional forms assumed to describe the probabilities with which the \( Y_{ij} \) assume those values, namely

\[
P(Y_{ij} = y | \theta_j, \eta_i) = f(y | \theta_j, \eta_i); \ y \in \Omega_Y \quad (2.1)
\]

The item parameters \( \eta_i \) may include three distinct types of parameters: a \textit{discrimination} parameter \( a_i \), a \textit{difficulty} parameter \( b_i \), and a \textit{guessing} parameter \( c_i \). The discrimination parameter \( a_i \) will be related to how rapidly the probability in equation (2.1) changes with changes in ability \( \theta_j \). The difficulty parameter models how difficult the item is. And the guessing parameter provides the probability that an examinee with a very low ability will answer the item correctly.

The common IRT model for items with only two response options is

\[
P(y_{ij} = 1 | \theta_j, a_i, b_i, c_i) = c_i + (1 - c_i) \frac{1}{1 + e^{-a_i(\theta_j - b_i)}} \quad (2.2)
\]

Equation 2.2 indicates that the probability of a response is the function of the items’ discrimination and the difference between the person’s location and item’s location. Graphical output of a model in 2.2 having 14 items with no guessing parameter \( (c_i = 0) \) is displayed in Figure 2.2. These are logistic functions, one for each item (question) showing the probability of a correct response as a function of the respondent’s score on the underlying latent variable ability.
The common IRT models for items with more than two response options will be discussed in the next section where we present an application of IRT. The technical detail is presented in the appendices.
2.2.2 Model Assumptions

Two key assumptions in an IRT model are (1) unidimensionality of the latent traits and (2) local independence. The first assumption implies that the items share a common primary construct and that the model creates a single \( \theta_j \) for each respondent. This means that the items collectively measure a unique underlying latent trait for each examinee and that only one latent trait influences the item responses. Other factors affecting these responses are treated as random error (DeMars, 2010, p. 38). This is a strong assumption and may not be reasonable in many situations as tests or survey instruments may be designed to measure multiple traits. However, as previously stated, we restrict our discussion to this unidimensionality assumption and refer to the IRT models that meet this assumption as unidimensional IRT, or UIRT, models. When the assumption does not hold, estimates of parameters and standard errors may be questionable.

The second assumption, local independence, indicates that if the assumption of unidimensionality holds, then the response of a subject to one item will be independent of his or her response to another item, conditional on the latent trait. In other words, if items are locally independent, they will not be correlated after conditioning on \( \theta_j \) (DeMars, 2010). Letting \( \mathbf{y}_j = (y_{1j}, y_{2j}, \ldots, y_{ij}, \ldots, y_{lj}) \), with \( i = 1 \ldots I \) and \( j = 1 \ldots n \), be the vector of \( I \) observed responses from the \( j^{th} \) subject having an ability \( \theta_j \), the assumption of local independence can be expressed as

\[
P(\mathbf{y}_j | \theta_j, \eta) = P(y_{1j} | \theta_j, \eta_1)P(y_{2j} | \theta_j, \eta_2) \ldots P(y_{lj} | \theta_j, \eta_l)
\]

\[
= \prod_{i=1}^{l} P(y_{ij} | \theta_j, \eta_i)
\]

\[
= \prod_{i=1}^{l} f(y_{ij} | \theta_j, \eta_i)
\]  \( (2.3) \)

where \( P(\mathbf{y}_j | \theta_j) \) is the probability that the vector of observed item scores for a person with trait level \( \theta_j \) has the pattern \( \mathbf{y}_j \), and \( \Pi \) is the symbol for the product of the individual probability \( P(y_{ij} | \theta_j) \) for a person with a trait level \( \theta_j \) obtaining a score of \( y_{ij} \) on item \( i \). Expressing equation (2.3) in term of the \( \theta_j \) we obtain the likelihood function
This can be generalized to the probability, \( P(y|\theta) \), of a complete set of responses from \( n \) persons to \( I \) items on an instrument, where \( \theta = \{\theta_1, \theta_2, \ldots, \theta_n\} \) represents the vector of latent scores for all respondents, as shown in equation 2.5:

\[
P(y|\theta, \eta) = \prod_{j=1}^{n} P(y_j|\theta_j, \eta_j) = \prod_{j=1}^{n} \prod_{i=1}^{I} f(y_{ij}|\theta_j, \eta_i)
\] (2.5)
2.3.1 Data Description

The survey sample was drawn from the National School Boards Association’s database of school boards and superintendents in the United States. The sample was stratified and included all board members and superintendents in 118 urban districts belonging to the National School Boards Association’s Council of Urban Boards of Education (CUBE) as well as the groups from 400 smaller districts with enrollment of 1000 or more. A link to the online survey was sent to 3805 board members and superintendents in the 518 districts. A paper survey was also distributed to those members who did not provide a valid email address. Of those surveyed, 1200 responded, including 900 board members and 120 superintendents from 418 districts. The response rates were 23.6% for board members and 22.5% for superintendents. The survey consisted of 90 questions in total. Aside from some demographic questions, all of the questions related to the issue of student achievement. Respondents only answered the set of questions out of the 90 total survey questions that were relevant to their roles, as a board member, a superintendent, or a board chair. Of those 90 questions, 23 were common for all respondents. At least one response was received from 80.1 percent of the districts surveyed.

Demographic information from the survey indicates the types of individuals who served on school boards. The data suggest that nationally, about 80.7% were white, 12.5% were African-American, and 3.1% were Hispanic. Larger school districts were more likely to include minority board members. Overall, 58% were males; however, the gender distribution was more evenly equal in larger school districts as can be seen in Figure 2.3. Half of all board members have served in their current districts for more than five years and more than a quarter of them were current or former educators themselves. On the whole, board members were better educated than the general population, as more than half reported having earned an advanced degree (Masters, PhD, or EdD). They were moderate to conservative in their politics and professional or business-men and women in their careers. The majority were in the middle-aged group, as 70% were fifty or older and 54% had family annual income of $100,000 or more. Most served on the board for public-service motives.
Figure 2.3  Bar graph of gender within four school enrollment categories.

Question 21 was designed to assess the attitude of board members about factors related to student learning. Participants were to indicate on a 5-point scale (0 = “not at all important”, 1 = “somewhat important”, 2 = “moderately important”, 3 = “very important”, 4 = “extremely important”) how important each of 11 approaches were to improving student learning outcomes. For all these items, higher scores indicate greater levels of importance. Overall, 7% of the participants did not respond to all parts of this question. Item by item, the missing value percentages range from 1.4% to 2.6%, or from 14 to 28 observations per item. Frequencies and proportions of responses to these items are presented in Tables 2.1 and 2.2. For some items the responses tend to concentrate in the upper two or three categories (negative skew), but this is especially true for items C, J, and K. The pattern is opposite for item I (charter school), with 69% of votes in the lowest score.
Question 21: How important do you think each of these approaches is to improving student learning?

<table>
<thead>
<tr>
<th>Item</th>
<th>Not at all important 0</th>
<th>Somewhat important 1</th>
<th>Moderately important 2</th>
<th>Very important 3</th>
<th>Extremely important 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Aggressively recruiting non-traditional teachers</td>
<td>192</td>
<td>225</td>
<td>298</td>
<td>194</td>
<td>85</td>
</tr>
<tr>
<td>B. Increasing school choice within the district</td>
<td>367</td>
<td>177</td>
<td>208</td>
<td>166</td>
<td>74</td>
</tr>
<tr>
<td>C. Professional development</td>
<td>3</td>
<td>23</td>
<td>102</td>
<td>395</td>
<td>483</td>
</tr>
<tr>
<td>D. Reducing class size</td>
<td>62</td>
<td>133</td>
<td>283</td>
<td>298</td>
<td>228</td>
</tr>
<tr>
<td>E. Linking teacher pay to student performance</td>
<td>218</td>
<td>190</td>
<td>218</td>
<td>210</td>
<td>167</td>
</tr>
<tr>
<td>F. Boosting pay for teachers across the board</td>
<td>192</td>
<td>211</td>
<td>240</td>
<td>216</td>
<td>142</td>
</tr>
<tr>
<td>G. Improving the quality of district leadership</td>
<td>61</td>
<td>107</td>
<td>177</td>
<td>342</td>
<td>315</td>
</tr>
<tr>
<td>H. Implementing a year-round school calendar</td>
<td>417</td>
<td>217</td>
<td>183</td>
<td>120</td>
<td>66</td>
</tr>
<tr>
<td>I. Supporting the creation of new charter schools</td>
<td>683</td>
<td>164</td>
<td>79</td>
<td>48</td>
<td>23</td>
</tr>
<tr>
<td>J. Frequent use of assessment data to guide decisions</td>
<td>9</td>
<td>71</td>
<td>136</td>
<td>325</td>
<td>460</td>
</tr>
<tr>
<td>K. Improving the quality of school leadership</td>
<td>36</td>
<td>66</td>
<td>147</td>
<td>332</td>
<td>422</td>
</tr>
</tbody>
</table>

Table 2.1  Frequencies of the responses to the 11 items in question 21. Item I received the most zero ratings, while item C had the largest number of highest ratings.

<table>
<thead>
<tr>
<th>Item</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Non Tradition Teachers</td>
<td>0.19</td>
<td>0.23</td>
<td>0.30</td>
<td>0.20</td>
<td>0.09</td>
<td>0.025</td>
</tr>
<tr>
<td>B. School Choice</td>
<td>0.37</td>
<td>0.18</td>
<td>0.21</td>
<td>0.17</td>
<td>0.07</td>
<td>0.027</td>
</tr>
<tr>
<td>C. Professional development</td>
<td>0.00</td>
<td>0.02</td>
<td>0.10</td>
<td>0.39</td>
<td>0.48</td>
<td>0.014</td>
</tr>
<tr>
<td>D. Reduce class size</td>
<td>0.06</td>
<td>0.13</td>
<td>0.28</td>
<td>0.30</td>
<td>0.23</td>
<td>0.017</td>
</tr>
<tr>
<td>E. Performance pay</td>
<td>0.22</td>
<td>0.19</td>
<td>0.22</td>
<td>0.21</td>
<td>0.17</td>
<td>0.017</td>
</tr>
<tr>
<td>F. Increase pay</td>
<td>0.19</td>
<td>0.21</td>
<td>0.24</td>
<td>0.22</td>
<td>0.14</td>
<td>0.019</td>
</tr>
<tr>
<td>G. District leadership</td>
<td>0.06</td>
<td>0.11</td>
<td>0.18</td>
<td>0.34</td>
<td>0.31</td>
<td>0.018</td>
</tr>
<tr>
<td>H. Year-round schools</td>
<td>0.42</td>
<td>0.22</td>
<td>0.18</td>
<td>0.12</td>
<td>0.07</td>
<td>0.017</td>
</tr>
<tr>
<td>I. Charter schools</td>
<td>0.69</td>
<td>0.16</td>
<td>0.08</td>
<td>0.05</td>
<td>0.02</td>
<td>0.023</td>
</tr>
<tr>
<td>J. Data-driven decision</td>
<td>0.01</td>
<td>0.07</td>
<td>0.14</td>
<td>0.32</td>
<td>0.46</td>
<td>0.019</td>
</tr>
<tr>
<td>K. School leadership</td>
<td>0.04</td>
<td>0.07</td>
<td>0.15</td>
<td>0.33</td>
<td>0.42</td>
<td>0.017</td>
</tr>
</tbody>
</table>

0= Not at all, 1= Somewhat, 2= Moderately, 3= Very, and 4= Extremely important

Table 2.2  Proportions of the responses to the 11 items in question 21. For some items the responses are concentrated in the upper two or three categories; for others the scores are spread across all five response options. The missing rates range from 1.4% to 2.5%.
In the subsequent analyses, these rating categories were changed to a new scale from 1 to 5 for ease of implementing the functions in the package *ltm* [14] in R. Kendall correlation values and significance test results are displayed in Table 2.3. The values in the lower diagonal indicate the $p-$values for the correlation tests. We observe some large $p-$values (> .05) and some small $p-$values (< 0.05). This result suggests that the 11-item scale might contain more than one trait. In the next section, we present the dimensionality assessment for these items.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.259</td>
<td>0.059</td>
<td>0.019</td>
<td>0.237</td>
<td>0.04</td>
<td>0.122</td>
<td>0.154</td>
<td>0.195</td>
<td>0.045</td>
<td>0.130</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>&lt;0.001****</td>
<td>0.033</td>
<td>0.057</td>
<td>0.190</td>
<td>0.077</td>
<td>0.141</td>
<td>0.115</td>
<td>0.289</td>
<td>0.002</td>
<td>0.142</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0.032</td>
<td>0.054</td>
<td>0.101</td>
<td>0.186</td>
<td>0.017</td>
<td>0.141</td>
<td>0.049</td>
<td>-0.006</td>
<td>0.273</td>
<td>0.200</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>0.042</td>
<td>0.031</td>
<td>&lt;0.001****</td>
<td>0.040</td>
<td>0.240</td>
<td>0.103</td>
<td>0.062</td>
<td>-0.029</td>
<td>0.019</td>
<td>0.073</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>0.334</td>
<td>0.120</td>
<td>****</td>
<td>0.08</td>
<td>0.130</td>
<td>0.122</td>
<td>0.212</td>
<td>0.166</td>
<td>0.163</td>
</tr>
<tr>
<td>F</td>
<td>0.115</td>
<td>&lt;0.003</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>0.002</td>
<td>****</td>
<td>0.111</td>
<td>0.073</td>
<td>-0.000</td>
<td>0.166</td>
<td>0.136</td>
</tr>
<tr>
<td>G</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>****</td>
<td>0.182</td>
<td>0.145</td>
<td>0.214</td>
<td>0.677</td>
</tr>
<tr>
<td>H</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>0.074</td>
<td>0.017</td>
<td>&lt;0.001</td>
<td>0.005</td>
<td>&lt;0.001</td>
<td>****</td>
<td>0.152</td>
<td>0.070</td>
<td>0.182</td>
</tr>
<tr>
<td>I</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>0.827</td>
<td>0.287</td>
<td>&lt;0.001</td>
<td>0.599</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>****</td>
<td>0.100</td>
<td>0.147</td>
</tr>
<tr>
<td>J</td>
<td>&lt;0.003</td>
<td>0.928</td>
<td>&lt;0.001</td>
<td>0.170</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>0.001</td>
<td>&lt;0.001</td>
<td>****</td>
<td>0.323</td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>0.007</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>****</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.3 Significance tests of Kendall’s correlations. The upper diagonal part contains correlation coefficient estimates and the lower diagonal part contains corresponding $p$-values. The presence of small and large (> .05) $p$-values indicates that the 11-item scale may have more than one trait.

### 2.3.2 Identifying Latent Traits

To assess the dimensionality of the constructs represented by the data, a scree plot of the eigenvalues of the tetrachoric (polytomous items) correlation matrix was constructed. The plot in Figure 2.4 appears to have two dominant eigen-values indicating that the items may cluster in two dimensions rather than all together. Scale unidimensionality was then assessed by the method of confirmatory factor analysis for ordinal data using the software LISREL version 8.8. The factor loadings displayed in Table 2.4 suggest the creation of two factors (from applying either Varimax or Promax rotation). Using the criterion that items are assigned to the factor on which the item has a higher loading, **Factor 1** comprises 8 items: A *(Non-traditional teachers)*,
B (School choice), E (Performance pay scale), G (District leadership), H (Year-round schools), I (Charter schools), J (Data-driven decision making), and K (School leadership). **Factor 2** consists of 3 items: C (Professional development), D (Small class size), and F (Pay increase). Since the items in the second factor represent a knowledge-based and traditional approach, we labeled this the *Classical* factor. In the same spirit, because the items in the first factor represent more innovative characteristics, we named the factor *Innovative*. Thus the 11-item scale represents two domains, *Classical* and *Innovative*, with three and eight items in each domain, respectively.

Figure 2.4  Scree plot of the eigenvalues. The scree clearly begins to flatten at the third eigenvalue.
Table 2.4 Table of factor loadings generated from the ordinal factor analysis using LISREL 8.8 software. Two dominant factors emerge, with 8 items having high loadings on the first factor (items A, B, E, G, H, I, J, and K) and 3 on the second factor (items C, D, and F).

The analysis that follows separates the data into two dimensions, with the association of latent traits of attitude (or enthusiasm) about Classical and Innovative approaches to improving student learning.

### 2.4 The Graded Response Model (grm)

IRT analyses were conducted for each of the two domains, Classical and Innovative, to obtain an attitude score for each respondent on each domain. The graded response model assumes that the item response is an ordered categorical variable whose values are not separated by equal intervals. This model fits the rating scale of our polytomous items and therefore was selected to model the interaction between the respondents and the multiple-response category items in each domain. The package ltm in R was utilized for this analysis. A brief overview of the model formulation and analysis is described next.
2.4.1 Model Formulation and Analysis

As previously discussed, an UIRT model operates under the assumptions of unidimensionality and local independence. A GRM model derives the probability of a response for a particular item in a test as a function of the latent trait $\theta$ and the item parameters. We are interested in the probability of responding in a specific category. In a GRM, the cumulative probability or the probability of responding in or above a given category is modeled. Then the probability of responding in a specific category is modeled as the difference between two adjacent cumulative probabilities.

For $i = 1, \ldots, I; j = 1, \ldots, n; \text{ and } k = 1, \ldots, K$, let $I$ denote the number of items, $n$ the number of persons, and $K_i$ the number of response categories which we assume is the same for all items. Let $Y_{ijk}$ be response $k$ to item $i$ for person $j$. Let $a_i$ represent the discrimination parameter for item $i$, and $b_{ik}$ be the category boundaries or thresholds for category $k$ of item $i$. There are $K-1$ thresholds, $b_{ik}$s, between the response options. These thresholds are the boundaries between two adjacent cumulative scores; for example, $b_{i3}$ is the threshold between a score of 3 or higher and a score of 2 or lower.

Let $P(Y_{ijk} \geq k|\theta_j, b_{ik}, a_i)$ be the cumulative probability of scoring in or above category $k$ of item $i$ given the item and person parameters. These cumulative probabilities are then modeled as

\[
P(Y_{ijk} \geq 1|\theta_j, b_{i1}, a_i) = 1
\]
\[
P(Y_{ijk} \geq 2|\theta_j, b_{i2}, a_i) = \frac{1}{1 + e^{-a_i(\theta_j - b_{i2})}}
\]
\[
P(Y_{ijk} \geq 3|\theta_j, b_{i3}, a_i) = \frac{1}{1 + e^{-a_i(\theta_j - b_{i3})}}
\]
\[\vdots\]
\[
P(Y_{ijk} \geq K + 1|\theta_j, b_{iK}, a_i) = 0
\]

and they lead to the form of the graded response model as

\[
P(Y_{ijk} = k|\theta_j, b_{ik}, a_i) = P(Y_{ijk} \geq k|\theta_j, b_{ik}, a_i) - P(Y_{ijk} \geq k + 1|\theta_j, b_{ik}, a_i)
\]
\[= \frac{1}{1 + e^{-a_i(\theta_j - b_{ik})}} - \frac{1}{1 + e^{-a_i(\theta_j - b_{ik+1})}}
\]
where \( P(Y_{ijk} = k|\theta_j, b_{ik}, a_i) \) is the probability of responding at a specific category \( k \).

The example plots (not from the board data) of the boundary probabilities, \( P(Y_{ijk} \geq k|\theta_j, b_{ik}, a_i) \), and the probabilities of responding at a specific category in an item, \( P(Y_{ijk} = k|\theta_j, b_{ik}, a_i) \), are displayed in Figure 2.5. These curves are referred to as the item operating characteristic curve (OCC) and the item category characteristics curve (ICC), respectively. Each item has its own OCC and ICC curves, and each curve represents the relationship between the latent trait and the observed ratings. The OCCs specify the cumulative probability of a response in the categories above or below the threshold and the ICCs show the probability of each score category 1, 2, 3, and 4 for a person at a specific \( \theta \) level. The OCCs cross the .5 probability at the points equal to the category difficulties (thresholds) and their slopes are steepest at that point. Although the ICC curves for the lowest and highest categories (1 and 4) cross the .5 probability line (horizontal dotted blue line) at the item thresholds \( b_1 = -3.06 \), and \( b_3 = 1.54 \), the curves for the middle categories (2 and 3) do not necessarily correspond to the item thresholds. The peaks of the curves do not have any obvious connection to the \( b_{ik} \) parameters.
Figure 2.5 OCC and ICC curves for an item with four response categories. The OCC curves (upper) represent the cumulative probability functions crossing the .5 probability at the step difficulty parameters (threshold) $b_1 = -3.06, b_2 = 0.91$, and $b_3 = 1.54$ (see the light blue vertical lines). Each of the five ICC curves (lower) represents the probability for each response category. The two curves for the lowest and highest categories (1 and 4) cross the .5 probability line (horizontal dotted blue line) at the item threshold $b_1 = -3.06$, and $b_3 = 1.54$. However, the curves for the middle categories (2 and 3) do not intersect at the item thresholds.

A summary of item category threshold $b_{ik}$ is presented in Table 2.5. The OCC and ICC of item 4 are displayed above. The cut-points in the last line, 1 vs 2-4, 1-2 vs 3-4, etc coincide with the legend of the OCC figure.

We fitted the IRT grm model to the two domains, Classical and Innovative, using a constrained grm model (model with equal slopes) and unconstrained model (unequal slopes) for
Table 2.5 Examples of item category thresholds for four items. The last row contains the cut-points at each category option.

each domain. A likelihood ratio test was used to compare the fit of the equal or unequal slope models. Results of the analysis and discussion of the model fit are presented in section 2.5. The next section describes some technical aspect of model assessment and parameter estimation.

2.4.2 Model-Data Fit Assessment

We utilized a simulation based approach to assess the model fit in which we constructed a discrepancy measure of goodness of fit using the Kullback-Leibler divergence statistics given by equation 2.8

\[ D_{KL}(P||Q) = \sum_{i=1}^{I} \ln \left( \frac{P(i)}{Q(i)} \right) P(i) \]  

(2.8)

where \( P(i) \) is the distribution of the proportion of responses in each category for item \( i \) from the observed data, and \( Q(i) \) is the distribution of the proportions of responses in each category for item \( i \) from the fitted data. Treating the model as a data-generating mechanism, we simulated \( M \) data sets, computed \( M \) values of \( D_{KL}s \), and obtained the mean \( \bar{D}_{KL} \) as the observed test statistics \( T_{obs} \). The procedure was repeated again \( M \) times, each time utilizing one data set from the \( M \) simulated data sets as actual data. We then obtained \( M \) values of \( \bar{D}_{KL} \). These values form a reference distribution of \( \bar{D}_{KL} \) from which to compare the observed test statistic. The proportion of \( \bar{D}_{KL} \) greater than \( T_{obs} \) is used as the \( p \)-value, which is the probability of obtaining a \( T_{obs} \) of equal value or higher. If the \( p \)-value is less than .05, we would have evidence against the null hypothesis of the fitted model. Otherwise, we would fail to reject \( H_0 \) and conclude that we did not have enough evidence to reject the possibility of \( H_0 \).

To discover how much we can learn about the latent trait from each item, we can use the concept of item information. The information curves for each item are displayed in Figure 2.6.
Each item provides some information about the underlying latent trait and the information function describes how well each item performs as a function of the latent trait. Thus, a high information value is associated with a small standard error of measurement.

Figure 2.6  Left: Item Information Curves (IIC) for four items. Item 3 gives the most information \( IIC_3 \) about the latent trait compared to items 1, 2, and 4. Right: Test or total Information Curve. The curve resembles the shape of the \( IIC_3 \) since item 3 has the largest amount of information.

The term *information* in IRT describes how certain we feel about the estimate of a person’s location \( \theta_j \) and is based on the Fisher Information matrix

\[
I(\theta) = E \left( \left( \frac{\partial \ln L}{\partial \theta} \right)^2 \right) = -E \left[ \frac{\partial^2 \ln L}{\partial \theta^2} \right] \tag{2.9}
\]

where \( L \) is the likelihood function defined in (2.4).

Individually, each dichotomous item information function is defined as

\[
I_i(\theta) = -E \left( \frac{\partial^2 \ln P_i(\theta)}{\partial \theta^2} \right) \tag{2.10}
\]
where \( P_i(\theta) \) is the probability of a correct response to dichotomous item \( i \). For polytomous items, the item information function, \( I_i(\theta) \), is the sum of the category information functions, \( I_{ik}(\theta) \). \( I_{ik}(\theta) \) is based on the Fisher information matrix and defined by Samejima (1969, pp. 36-38) as in equation 2.11

\[
I_{ik}(\theta) = -\frac{\partial^2 \ln P_{ik}(\theta)}{\partial \theta^2} = -\frac{\partial}{\partial \theta} \left[ \frac{P'_{ik}(\theta)}{P_{ik}(\theta)} \right] = \frac{\left[ P'_{ik}(\theta) \right]^2 - P_{ik}(\theta)P''_{ik}(\theta)}{[P_{ik}(\theta)]^2} \tag{2.11}
\]

where \( P_{ik}(\theta) \) is the probability of a response in category \( k \) to item \( i \) defined in equation (2.7), and \( P'_{ik}(\theta) \) and \( P''_{ik}(\theta) \) are the first and second derivatives of \( P_{ik}(\theta) \).

Thus, the item information function is calculated as

\[
I_i(\theta) = \sum_{k=1}^{K} I_{ik}(\theta)P_{ik}(\theta) \tag{2.12}
\]

and the total or test information is the simple sum of the item information

\[
T(\theta) = I(\theta) = \sum_{i=1}^{I} I_i(\theta) \tag{2.13}
\]

### 2.4.3 Parameter Estimations

The graded response model parameters are estimated using the Marginal Maximum Likelihood (MML) or the Joint Maximum Likelihood (JML) estimation technique. In the MML method, the estimation of item parameters and the estimation of person parameters are performed in two separated steps. The item parameters are estimated first, assuming a distribution for \( \theta \). This is followed by the estimation of the person parameters, which technically is a prediction since the \( \theta_j \) are treated as random variables.

**Item Parameter Estimation**

With the likelihood conditional on \( \theta = (\theta_1, \theta_2, \ldots, \theta_n) \) given in expression 2.5 and an assumed normal distributional form \( g(\theta_j|\psi) \) for the independent and identically distributed latent traits, the marginal log likelihood for item parameters may be written as

\[
l(\eta) = \sum_{i=1}^{I} \sum_{j=1}^{n} \log \int f(y_{ij}|\theta_j, \eta_i)g(\theta_j|\psi)d\theta_j \tag{2.14}
\]
Because of the assumption of local independence mentioned previously, maximization of
2.14 reduces to maximization of

\[
l_i(\eta_i) = \sum_{j=1}^{n} \log f(y_{ij}|\theta_j, \eta_i)g(\theta_j|\psi)d\theta_j
\]  

(2.15)
for one item at a time, where \( \eta_i = (a_i, b_i, c_i) \) for \( i = 1, ..., I \) and \( \psi \) is the set of hyper parameters for mean and standard deviation, usually set at 0 and 1 respectively. The integrals in 2.15 are numerically approximated using a Gauss-Hermite quadrature algorithm. After the item parameters are estimated, they are used to update information on the distribution of \( \theta \), and the item parameters are re-estimated. The procedure is repeated until the estimated values stabilize or converge. After the item parameters and the \( \theta \) distribution have been estimated, the \( \theta \) score for each subject can be estimated.

**Person Parameter Estimation**

After the item parameters and the \( \theta \) distribution have been estimated, the \( \theta \) score for each subject can be estimated using Expected a Posteriori or Modal a Posteriori procedures. Each examinee has his or her own \( \theta \) posterior distribution. For examinee \( j \), the posterior density is

\[
P(\theta|\hat{\eta}, \hat{\psi}, y_j)
\]

(2.16)

The Modal a Posteriori, or MAP, procedure estimates \( \theta_j \) by using the mode of equation (2.16) as the maximum value, while the Expected a Posteriori (EAP) procedure uses an estimated expected value. In this analysis we used MAP.

The mode of equation (2.16) can be found by applying the Fisher Scoring Method. The \((t+1)^{th}\) iteration is

\[
\hat{\theta}_{t+1} = \hat{\theta}_t - \left[\frac{\partial l(\theta_j|\eta_i,y_{ij})}{\partial \theta_j} \right] / \left[\frac{\partial^2 l(\theta_j|\eta_i,y_{ij})}{\partial \theta^2_j} \right]_{t}
\]

(2.17)

where \( \frac{\partial l(\theta_j|\eta_i,y_{ij})}{\partial \theta_j} / \frac{\partial^2 l(\theta_j|\eta_i,y_{ij})}{\partial \theta^2_j} \) is the ratio of the first derivative of the loglikelihood function of \( \theta_j \) and the Hessian matrix, which is the matrix of second derivatives of the loglikelihood function.
of $\theta_j$. This is an iterative method with $\hat{\theta}_t$ being updated until convergence is achieved.

There is usually a normal distribution assumption for the $\theta_j$s. The estimates for the $\theta_j$s are more accurate if they span the entire range of item difficulties. The graph of the distribution of a set of estimated $\theta_j$s is shown in Figure 2.7. By default, the mean score was set to zero and the standard deviation of the scores was set to 1.

![Kernel Density Estimation](image)

Figure 2.7  Estimated distribution of latent scores, $\theta_j$s. The plot has a heavier lower tail than would be expected in a normal distribution $\theta_j$.

Analyses are done separately for each domain. We will discuss the Classical domain first followed by the Innovative domain.

### 2.5 Analysis and Results

#### 2.5.1 Analysis of the Classical Domain

The Classical domain consists of three items: C (Professional development), D (Reduced class size), and F (Pay increase). We fitted two graded response models to the data, one constraining the discriminant parameters, $a_i$, to be the same, and the other allowing the $a_i$ to vary. The model is displayed here again for convenience.
In the reduced model with \( a_1 = a_2 = \ldots = a_I = a \) all items are assumed to provide the same amount of information about how positive (or negative) respondents’ attitudes were about the importance of classical or traditional factors for student learning. In the full model with distinct value of \( a_i \) for \( i = 1, \ldots, I \), the items might carry different amount of information about attitudes. The reduced and full models were compared using a likelihood ratio test. The \( p \)-value for this test was 0.241, so the data do not provide enough evidence to reject the reduced or constrained model.

To assess the goodness of fit of the model, we used our simulation-based Kullback-Leibler approach discussed in Section 2.4.2. The large \( p \)-value (.5488) from 1000 simulations suggested that the constrained grm model may be useful and that it would be possible to extract and interpret the results from it. We reviewed the results for the following four components, the \textit{item parameters}, \textit{person parameters}, \textit{information function}, and \textit{reliability of the estimators}.

**Item Parameters**

In evaluating the item parameter estimates we aimed to answer two questions: “\textit{What is the spread of the item category difficulties?}” and “\textit{How discriminating is each item?}” The test results presented above suggest that all three items have similar discriminating power \( a \), but we still need to determine how well and in what segment of the latent trait the items discriminate among the respondents. Table 2.6 shows the estimated item parameters for the constrained model, the item category difficulties \( b_{ik} \), and their discriminating power \( a \). The cut-points that separate the categories are shown in the bottom row. There are four thresholds separating five categories. Everyone had a 100\% chance of choosing \textit{Not at all important} or higher, so there was no threshold for that option.
Table 2.6  Estimated parameters for constrained graded response model for the Classical domain.

For item C, the thresholds span the negative section of the trait. A score of 5 was the most probable for respondents above the zero latent trait level ($b_{C4} = .10$). The rest of the group was more likely to rate it with a score of 4. Other options were unlikely as the other thresholds were more than two standard deviations below the mean zero. This indicates that this item was considered an “easy” item to rate high and was unable to differentiate between low and high trait-level respondents only between the 4-score and 5-score groups. The thresholds of items D and F are more evenly spread out in the range of $\theta$ with item F spread further to the positive side. They were “harder” items, therefore less likely to receive a concentration of high scores. They were also better than item C at separating between the low and high attitude groups.

The item parameters can also be interpreted graphically. Figure 2.8 displays the probability of choosing each category or higher. This probability is a function of the latent attitude. The curves are called the item operating characteristic curves, or OCCs. The thresholds are the intersections at the 0.5 probability lines. For item C, the thresholds span across the negative area while the thresholds of items D and F are well spread across the attitude range. For this constrained model, every curve has the same slope at the 0.5 probability line.
Figure 2.8  Plot of item Operating Characteristics Curves (OCCs), or boundary curves, for items C, D, and F. The curves describe the relationship between the probability $P(\theta)$ of choosing a category option or higher in an item for all respondents. The $b_{ik}$s are the intersection between the curves and the horizontal line where $P(\theta) = .50$.

Figure 2.9 displays the probability for each item’s response category. These curves are called the Item Characteristics Curves or ICCs. The curves for options 4 and 5 are most dominant for item C, while all other options are more nearly equally probable at the different trait levels for items D and F. So how well do these items discriminate between low and high attitude groups? The relatively high discrimination parameter ($a = 1.13$) is reflected in the sharp peaks of the ICC curves and the steep slope of the OCC curves. The probability of switching between two adjacent categories changes rather rapidly at the intersections of the OCCs, indicating a clear distinction between two different attitude groups.
Figure 2.9  Plots of Item Characteristic Curves. As the probability of choosing one option decreases, the probability of selecting the next option increases. For item C, a score of 1 (less than 2) appears to be seldom selected. The curves of the first and last categories have opposite monotonic patterns while the curves of the middle categories have unimodal shapes.

Next we look at the distribution of the person estimates, $\theta_j$, and compare it to the distribution of the item category thresholds.

**Person Parameters**

A *latent (attitude)* score was estimated for each response pattern along with the score distribution using the empirical Bayesian scoring method with a normal prior applied to the likelihood. These scores were set to center at 0 with standard deviation of 1. The curve appears to have a normal bell shape with a heavy left tail. Figure 2.10 displays the graph of the estimated distribution and other related components of the estimated attitude scores. Although the distribution is not exactly normal (from the Q-Q plot and Shapiro-Wilk test result), the scores spread between $-2$ and 2 and concentrate around zero. Item C provides more information on the negative side while the category estimates of items D and F roughly cover the range of $\theta$ and provide more information in the coverage area. This leads to our subsequent discussion of
the test and item information.

Test Information and Item Information

We now turn to the concept of item information to discover how much we can learn about the latent trait from each item. The information function, $I(\theta)$, indicates the precision of the attitude estimates. The plots of the total information function and the individual item information functions are displayed in Figure 2.11. Overall, the scale provided about 50% of the information on precision of the attitude estimates. This low number may be due to the limited length of the scale with only three items. The plot on the right lists the contribution of each item within the $\theta$ range of $(-2, 2)$. Each item gives more information near its thresholds. Item C provides more information on the negative side, while items D and F provide information in the range of $-2$ to $2$. Overall, our estimates are more precise within the above range as the information curves are high.
Reliability

How reliable are the model estimates, $\theta_j$, $a_i$, and $b_{ik}$? As indicated above, due to the limitation of length, the parameter estimates are reliable only in a certain range of $\theta$. The model estimates were more precise where the standard error curve was at its lowest range, which is between $-2$ and $1$ as seen in Figure 2.12. Reliability of $a$ and $b_{ik}$ could be assessed through the standard errors of their estimates.
Figure 2.12  Plot of instrument information and standard error of measurement. The estimates are more precise in the range where the standard error curve is the lowest, which is $-2 < \theta < .5$.

2.5.2 Analysis of the Innovative Domain

Following the same procedure as used in analysis of the Classical domain, we applied the IRT technique to the Innovative domain. This domain consisted of eight items: A (Non-traditional teachers), B (School choice), E (Performance pay scale), G (District leadership), H (Year-round schools), I (Charter schools), J (Data-driven decision making), and K (School leadership). Two grm models were fitted to the data, with constrained and unconstrained slopes. The likelihood ratio test indicated that the unconstrained model is a better option ($p$ - value < 0.001). This is clearly indicated in the slopes of the OCCs and ICCs in Figures 2.13 and 2.14. The curves appear to have two distinct groups with very different slopes. The first one consists of the two items G and K, whose slopes are markedly steep, and the second consists of all the other six items (A, B, E, H, I, and J), whose slopes are fairly flat.
Figure 2.13  ICC curves. The middle categories in Items G and K have a narrow unimodal shape while for the other items the curves are very flat. For items E, B, H, and I the score of 1 seems to be the most likely response for a large group of people, especially for item I (Charter schools). On the contrary, scores of 4 or 5 are the most likely selected options, and 1 is a seldom used score.

Notice the visible black curves for items B, E, H, and I. For these items a score of 1 is the most probable score, particularly for item I (Charter schools). This is the most “difficult” item to endorse, and only those with a very high attitude score, at least in the upper range of $\theta > 2.8$, would acknowledge this item as somewhat important or more. This is in contrast to the pattern of item C, Professional development. Items B, E, and H, are also “difficult” because endorsing these items with a score of 4 or higher (very important to extremely important) would require an attitude larger than 2, essentially in the upper 2% of the respondents. The black curves for these items span dominantly across the attitude range, revealing the fact that for these items the score of 1, or Strongly Disagree, is the most likely score. That is, most respondents did not
support Charter School or Year-round school, School choice, and Merit-based performance pay as approaches to improving learning outcomes.

Figure 2.14 Plot of the OCC curves. Categories in Items G and K have similar and very steep slopes while the other six items have very flat slopes.
To assess the fit of the model, we again utilized the Kullback-Leibler discrepancy index. Results from 500 simulations reject the null-hypothesis ($p$-value = 0.018), indicating that the unconstrained model may have some deficiencies in describing the data.

We conducted a number of exploratory procedures to attempt to determine the manner in which the model does not adequately reflect the patterns in the observed data. In particular, we examined the number of response patterns and the proportion of responses in each category for each question. Visually, they did not appear to be major discrepancies between the actual data and the data simulated from the fitted model. Neither were we able to detect consistent patterns in discrepancies between actual and simulated data sets. This led us to conjecture that the small $p$-value from our goodness of fit procedure primarily reflects the effects of a large amount of data resulting in a test that is highly sensitive to small departure from the proposed model. To investigate this conjecture in more detail we conducted the entire analysis with a data set that resulted from randomly selecting half of each gender-stratified sample from the actual data. This reduced data set was used to estimate the model, examine inferential conclusions, and conduct the goodness of fit procedure. Results were similar to the full analysis in terms of substantive inference, but the goodness of fit test returned a $p$-value of 0.467, resulting in failure to reject the model as an adequate representation of the data. We interpret this outcome as support for our conjecture that the original $p$-value of 0.018 produced from the full data set was the result of a hyper-sensitive procedure as the result of a large sample size.

Table 2.7 shows the values of the estimated item parameters. In general, the category estimates cover the whole spectrum of the attitude scale. The thresholds of items G and H spread to the negative side and are closer together; thus the ICC curves peak near the center of the $b$-parameters. These items can differentiate among three respondent groups of low, average, and high attitudes. For the other items, the categories are very far apart, which might add information about the latent trait at different locations. However, the ICCs were very flat, indicating that the items were unable to reliably distinguish low and high attitude groups.
Table 2.7 Table of item parameter estimates, category thresholds and item discriminants of the Innovative domain.

The distribution of $\theta$ is centered around zero and spreads quite evenly across the range from $-2$ to $2$, as seen in Figure 2.15. Figure 2.16 presents the plots of the test information, standard error, and the item information curves. The estimates of $\theta$ in the left plot appear to be more precise in the range from $-2$ to $1.5$. The plot on the right shows that only leadership items G and K contribute the most to the estimates of $\theta$. The curves of the other items are flat and thus are non-informative about $\theta$. This finding suggests that there are five distinct groups of board members each having a different attitude level on school and leadership leadership issue.
Figure 2.15  Plot of the distribution of the estimated attitude scores. The curve appears to have a normal distribution with a center at zero and standard deviation of 1.

Figure 2.16  Left: plot of test information and standard error of measurement. The estimates of $\theta$ appear to be more precise in the range from $-2$ to 1.5. Right: plot of item information curves. Only items G and K contribute the most to the estimates of $\theta$. The curves of the other items are flat and thus are non-informative about $\theta$. 


2.6 Discussion and Conclusion

This study was part of an effort to address beliefs of school board members and on strategies for improving students’ learning achievement. However, caution is needed in interpreting the meaning of the attitude trait. In this context, a low attitude score does not represent a negative or bad attitude. Likewise, a high attitude score does not perfectly equate to a high or good viewpoint. The scores are just representations of the respondents’ opinions on a certain issue and the diverse range of responses reflects the controversial aspect of the matter being discussed.

Reviewing the first issue in the Classical domain, Professional development, we observed that this item received a high level of support from the respondents. Our model predicted that more than 97.5% of the respondents were in the attitude range to endorse this item with a score of 4 or 5. Professional development (PD) is probably as classic as the education system itself because it has always been one of the primary venues for educators to obtain new ideas, knowledge, and teaching strategies. However, critics of PD, who often have referred to it as primordial and unreliable methods of education reform, may criticize the boards’ overwhelming support for this issue.

The second issue, Reducing class size, has been a topic of debate for many decades as to whether or not it is an effective way to improve learning. Responses on this item spread out more than for item C. It is apparent that less support was given to the class size reduction as a strategy of boosting achievement as compared to PD. Nevertheless, the majority still agreed that Reducing class size was a moderately to extremely important approach that affects student learning outcomes.

For item F, Increasing pay for teachers across the board, the category parameters were evenly spaced out across the attitude trait. However its range shifts further to the right. This item therefore appears to be more difficult to be rated high compared to items C and D for some respondents. Only those in the high range of attitude would endorse this item as very important or higher. It is interesting to note that less agreement is shown in viewing increasing teacher salary as a strategy for boosting student performance.

Results for the Innovative domain reveal three important factors. The first factor relates to
the matter of *Improving district and school leadership* in items G and K. Overwhelmingly, the majority of board members supported this issue. These two items were considered “easy” items because respondents only needed to be above the $-1$ *attitude* range to rate the items with a high score of 4 or 5 (very important or higher). This is an interesting finding as it appears to indicate that boards were strongly advocating for developing and improving systemic leadership for raising student achievement.

The second factor refers to item J *Frequent use of assessment data to guide decision making.* This item proves to be another “easy” or “no brainer” question because it received a uniform rating of *very* to *extremely important* from the vast majority of respondents. This is an intriguing finding as it appeared to support current movement in the educational system, an emphasis on using data to inform practice.

Finally, the issues that received the least support were *merit pay* and *implementing alternative school systems.* Board members were much more skeptical in their attitude regarding whether policy changes such as charter schooling or merit pay related to improving student learning. A substantial number (around 70%) of board members viewed *Charter schools, Increasing school choice for children within districts,* and *Year round calendar* as *not at all important.* This is another rather intriguing finding. Given the support and investment of the Obama administration for charter schooling, only a small percentage of board members though the creation of new charter schools is an *extremely or very important* approach for increasing learning standards. It would be of greater interest to understand why so many board members appeared unsupportive of charter schools and some of the most urgently needed *school choice* reform for children (Hess, 2010).

In conclusion, given the pivotal role of school boards in governing our nation’s public school systems, this research on school boards provides important information on the board priorities and actions that might impact district culture and achievement. The three most popular strategies supported were professional development, frequent use of assessment data, and improving the quality of school and district leadership. These strategies represented the belief in quality training, good data, and strong leadership as factors improving student performance. The three least supported tactics were charter schools, merit pay, and alternative school choice.
These reflected a less confident attitude in some of the current strategies for boosting learning outcomes.

Finally, the objective of school effectiveness research is to investigate the relationship between explanatory and outcome factors (Fox, 2007). In that process, our goal is to increase the reliability in the estimates that may require utilizing IRT in analyzing standardized tests for binary or polytomous data. IRT allows for inclusion of the interaction between the respondents and the test items and therefore providing more precise estimates of a person’s latent trait attitudes. We illustrated some basic features of IRT analysis via the application to data from a national survey of school board members, thus providing a comprehensive analysis on board members’ attitudes. In addition, we showed how IRT may be used for scaling purposes and illustrated a simulation-based approach using the Kullback-Leibler discrepancy to assess the graded response model data fit.

It is important to address some limitations of this study. We could have conducted multidimensional IRT rather than two separate unidimensional IRTs and compared the results from the two processes. Nevertheless current analyses allowed us to demonstrate the procedure twice on two different scales, which provided more examples for new IRT adventurers. Additionally, the models could be made richer by including predictor variables to explain factors affecting the rating scores. Multilevel modeling may also be considered at the school district level, as board members were clustered within districts. Finally, rescaling is also an option to increase the reliability of the estimates.

As a final note, since IRT is gaining popularity, open source software on IRT is becoming more abundant. This wider availability of functional software could prove advantageous if researchers familiarize themselves with IRT and with the techniques elaborated in this study.

**Acknowledgements**

This chapter is based on a study presented at the Joint Statistical Meeting in August 2011. I would like to acknowledge Hilary Lamonte and Mary Delagardelle from the Iowa Association of School Boards, and Rick Hess and Olivia Meeks from the American Enterprise Institute for their permission to use the school board data for this study.
BIBLIOGRAPHY


CHAPTER 3. STUDENT LEARNING EXPERIENCES RELATED TO USING TECHNOLOGY IN A STATISTICS COURSE

Dai-Trang Le, Ana-Paula Correia, Mack Shelley, and Kevin Kasper

A study to be submitted to the Journal of Technology Innovations in Statistics Education

Abstract

This study examines, evaluates, and discusses results from a study on the effectiveness of the teaching approaches that emphasize teaching statistics concepts using technology, data visualization, and data analysis tools. The data included responses from 84 students from a graduate-level statistics service course. Pre- and post- surveys were developed to examine students’ changes in attitude toward statistics after experiencing the teaching method. Data analysis methods included quantitative and qualitative techniques utilizing item response theory and content analysis. Although previous research generally has suggested that students’ attitudes toward statistics have been negative, results from our study indicated that students at this level have more positive attitudes than negative. Moreover, our approach to teaching statistics with an emphasis on visualization and immersion in technology has helped enhance student attitude and engagement, increase their motivation, and elicit greater interaction and participation in learning statistics. Therefore, the results of this study suggest that our approach does enhance the student learning experience.

3.1 Introduction

“Few college students escape taking an introductory course in statistics. But for many of these students, the lessons don’t seem to stick. They remember the pain but not the substance”
Nowadays, most college students (undergraduate and graduate levels) have to take at least an introductory statistics or a quantitative research methods course. Unfortunately, instead of viewing the class as important and useful, students often hold negative attitudes toward it. Many view these courses as non-engaging, boring, irrelevant, and difficult. Some drop out or felt resentful after completing the course. Students often perceive the course as saturated with jargon, demanding, and not readily applicable. Statistics educators cannot help but wonder why introductory statistics courses have earned such a negative image and what implications this may bring to the education of statistics.

As stated by Schau (2012) and Wentzel and Wigfield (2009), research on educational and cognitive theories found that students’ attitudes toward the discipline are important course outcomes, and, in fact are at least as important as knowledge and skills in the discipline. This is particularly true in statistics. As attitudes toward statistics can be a significant determinant of student performance and may influence student learning and adoption of statistical thinking, student achievement, and their willingness to apply what they learn to their professional lives (Gal, Ginsburg, & Schau, 2003), students who left statistics courses with negative feelings are unlikely to retain and apply what they learned. Therefore, understanding these attitudes and their relationship to learning should become more important in developing teaching approaches to enhance students’ learning experience.

Consequently, interest in improving instructional methods to improve students’ attitude and boost achievement has been rising for decades. Understanding these attitudes and their relationship to learning is important in developing teaching approaches to enhance students’ learning experience. Statistics educators have explored a variety of teaching strategies in the last few years, and there has been a shift on how to teach statistics to students of different ages as well as in a variety of different fields. In 1992, the American Statistical Association (ASA) and the Mathematical Association of America (MAA) formed a joint committee to study the teaching of introductory statistics. The main recommendations from that study were to *emphasize statistical thinking, incorporate data and emphasize concepts using less theory and fewer 'recipes',* and *to foster active learning* (Cobb, 1992). In more recent years, a movement to reform the teaching of statistics calls for researchers and teachers to focus on the synergy
among content, pedagogy, and technology (Moore, 1997). Not only should students be active participants assigned with structured activities that focus on statistical concepts and ideas that are not mathematical in nature, but content and pedagogy should be strongly influenced by technology (Moore, 1997).

Influenced by these recommendations many advocate utilizing technology, especially web-based software application and visualization technology to enhance student engagement, increase motivation and interaction, and boost achievement (Chick & Pierce, 2010, Forbes, 2012, Phelps & Caer, 2010). These researchers suggest that using visualization tools in teaching statistics with real-world application increases student attention and interest in the subject. In addition, these strategies influence instructors to teach quantitative reasoning skills through exploring data, while making theory and data analysis come alive (Forbes, 2012).

**Research Problem**

Recently, educators have begun to utilize popular web-based data visualization tools, such as Gapminder. This software has been used in a variety of Technology, Entertainment, Design (TED) talks to illustrate global and national economic, public health and many other issues. The Gapminder website (http://www.gapminder.org/) provides useful information such as lesson plans, videos, and worksheets for teachers and other educators who wish to include Gapminder in their teaching.

However, to date, no study has investigated the impact of Gapminder in teaching statistics and certainly no study has ever been done on the efficacy of combining Gapminder and JMP (a software used for data analysis). Our study seeks to fill this gap. What follows in this section is the listing of our research attributes, purpose, questions, and contributions.

**Unique Research Attributes**

There are eight several notable aspects and special features in our study.

1. Two different surveys for pre and post administration. The pre survey adopted 24 items from the Survey of Attitude Toward Statistics (SATS-36), a widely used instrument to measure six aspects of the students’ attitudes toward statistics (Schau, Stevens, Dauphinee, & Del Vecchio, 1995). We added questions on technology to probe students’ technol-
ogy skills. The post survey contained 29 questions of Likert-type and open-ended formats. Six of the questions on attitude were similar to the pre-survey and the others were new with content allowing for assessing students’ attitude changes within the questions.

2. The course in which the alternative teaching strategy was implemented is listed at a level appropriate to its audience of primarily master’s and doctoral graduate students. Thus, the course content and teaching methods are focused on practical research applications.

3. Course participants, graduate students and a few highly motivated undergraduates, came from a wide variety of disciplines (survey results showed more than 60 majors).

4. The method of delivery of course material in the class offered a blended teaching environment, in which more than half of students were enrolled to take the course online.

5. The focus of the teaching strategy was on using technology to facilitate independent learning and understanding, including a dynamic online learning environment using the MyStatLab learning management system, pre-recorded lectures, and live-recorded labs.

6. The class emphasized a teaching strategy that accented on using real-life data and applications of statistics.

7. Data analysis methods utilized item response theory (IRT) were employed to analyze the response patterns to all survey questions, rather than using simple mean or aggregate score, to assess students’ attitudes. Attitude here is treated as a latent variable to be extracted from the manifest variables and the raw survey responses. This IRT method provides an alternative way of analyzing Likert-type data. It takes into account the characteristics of the respondents and the survey questions. Specifically, IRT considers the interaction between the respondents and the survey questions.

8. Gapminder and JMP were used in combination as the main visualization technology. A review of the literature shows that no study has been conducted on the effectiveness of the conjunctive use of Gapminder and JMP in teaching statistics.
3.1.1 Research Purpose

The purpose of this study was to evaluate the effectiveness of a teaching approach with an emphasis on technology, visualization, and its impact on student learning experiences in statistics. We investigated how utilizing Gapminder in conjunction with JMP, a dynamic graphical emphasis statistical software, and other technologies would impact students’ overall learning experience. Our aim was to provide research-based information to improve the teaching and learning in statistical methods service courses. Specifically, our research aimed to answer the following broad research question and three related sub-questions.

3.1.2 Research Questions

To what extent did the use of the visualization tools, Gapminder and JMP, and related course technology, affect students’ learning experience?

We addressed the specific components of learning experience via these sub questions:

1. Was there an improvement in attitude over a period of six weeks?

2. If such a change occurred, what factors influenced the change?

3. Did any of the strategies we employed, such as the teaching method with a visualization focus, the immersion of technology, the use of Gapminder and JMP, the hands-on technique with real-life data, and the online learning environment MyLabPlus, play a major role?

3.1.3 Research Contributions

The unique attributes of this study contribute to the literature on statistics education in several ways. First, currently very little research has been conducted on the effect of teaching statistics with visualization technology particularly to graduate students. Second, we present an illustration of using IRT statistical methods to examine students’ attitudes and perceptions of the learning experience. IRT is well known and has been widely used in psychology, measurement, and testing, but still is not an popular method utilized by many quantitative-oriented
educational researchers. This study helps shed light on how IRT can be used to analyze survey data, particularly related to evaluating the impact of technology-driven education interventions. Third, this study shares for possible broader applications our innovative use of technology and its implications related to students’ learning experience in statistics.

In this study, we measure students’ attitude based on their expressed responses to survey items addressing their perception on statistics. We adopt this definition on attitude.

“Students’ attitude towards statistics is referred to as their feeling, interest, viewpoint about the relevance and worth of statistics, their belief in their own cognitive ability and in the difficulty of statistics as a subject, and the effort they are willing to put forward to learning.” (Schau, 2005)

3.2 Review of Literature

In this section we examine three relevant areas of statistics education: (1) teaching and learning statistics (2) using data visualization tools in teaching statistics, and (3) students’ attitude toward statistics. We first review the historical development and reform movement in teaching and learning statistics. Next, we examine the practice of using data visualization tools and technology in teaching statistics. Here, a special focus is given to the use of Gapminder and JMP. Finally, we discuss research on student attitudes toward statistics and present a brief overview of the instrument Students’ Attitude Toward Statistics (SATS) because some questions have been adopted from it. The section is organized into three subsections, corresponding to the three topics listed above.

3.2.1 Teaching and Learning Statistics

As our society becomes increasingly data-oriented and information-based, statistics is becoming an increasingly important scientific field of study. Statistics literacy, i.e., the ability to develop statistical thinking and reasoning skills, to understand data and charts, provides an important set of skills for today’s citizens and yields enhanced assets for a competitive workforce. Consequently, statistics training has become an integral part of the curriculum at every level of education. Currently, in post-secondary education, almost every student is required
to take a statistics course, regardless of their major. Similarly, most graduate programs include a requirement in quantitative research methods or statistics. In secondary education, since 2000, the National Council of Teachers of Mathematics (NCTM) has implemented standards related to data analysis and probability in the Pre-K-12 curriculum. Effective 2010, statistics has become a key component of the Common Core State Standards for Mathematics (http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf). In response to the rise in popularity and importance of teaching and learning statistics, the American Statistical Association (ASA) and the NCTM are calling on universities and school administrators across the country to prepare high-quality teachers of mathematics and statistics and to provide opportunities for appropriate professional development. Statistics researchers and educators are responding to the reform movement by focusing on improving the quality of teaching statistics at all levels, including the training of Pre-K-12 teachers. This effort focuses on making statistics an integral part of mathematics education in Pre-K-12 as it is more urgent now than ever. An example of this focus is the establishment of The Guideline for Assessment and Instruction in Statistics Education (GAISE) (http://www.amstat.org/education/gaise/GaiseCollege_Full.pdf) in 2007. GAISE recommended six features for introductory statistics courses:

1. Emphasize statistical thinking and literacy over other outcomes.

2. Use real data where possible.

3. Emphasize conceptual rather than procedural understanding.

4. Take an active learning approach.

5. Analyze data using technology rather than by hand.

6. Focus on supporting student learning with assessments.

The desired result of all introductory statistics courses is to produce statistically educated students (GAISE 2010). This means that students should develop statistical literacy and the ability to think statistically outside the classroom and in their academic and professional discipline, to be statistically literate and a wise consumer of information in a high-tech and modern society.
Technology has been a part of this movement and has influenced the way educators teach and students learn.

### 3.2.2 Data Visualization Technology in Teaching Statistics

The movement to reform the teaching of statistics calls for researchers and teachers to focus on content, pedagogy, technology, and active learning. Data visualization is a component of the technology emphasis. Research has suggested that using visuals in teaching results in a greater degree of learning (Stokes, 2003). One definition of data visualization is a graphical representation of data or concepts and it is a process of representing data as a visual image, (Zeidan, 2012). Another definition considers data visualization as the transformation of abstract data to a visual representation that can be easily understood by users, (Zeidan, 2012). The use of visual representations and interaction techniques should be intended to increase the understanding of abstract information. However, as the use of visuals must be carefully planned to be effective. It should focus on provoking and encouraging thoughtful analysis of the underlying meaning besides the aspects of presentation to excite and entertain learners.

A large body of literature focuses on the impact of visualization in teaching and learning, especially in teaching mathematics. Research suggests that using visual elements in teaching increases learning outcomes (LeGates, 2004). As we are increasingly surrounded with visual images, the emphasis on visual learning strategies and visual technologies continues to gain attention. Various studies of the power of visual learning in secondary mathematics education have been conducted. Visual learning is defined as “absorbing information from illustrations, photos, diagrams, graphs, symbols, icons, and other visual models. It is about making sense of complex information quickly—literally being able to comprehend ideas at a glance” (Murphy, 2003, p.2). Figure 3.1 displays some popular information visualization tools and Figure 3.2 presents the sketch of an information process via visualizing data.
However, visual learning is not only about acquiring but also communicating information. “The concept of visual literacy is defined as the ability to interpret images as well as to generate images for communicating ideas and concepts” (Stokes, 2005). “We are surrounded by increasingly sophisticated visual images” (Howells, 2003). Consequently, many people would say that their preferred learning style is visual. Surveys show that about 65% of people identify themselves as visual learners and the other 35% are kinesthetic and auditory (http://library.rpcc.edu/docs/LearningStylesAssessment-TRIO.pdf). However, based on our personal experience in both teaching and learning domains, very few people learn in only one way but everyone benefits from visual methods. Therefore, visual learning strategies are becoming increasingly powerful tools in teaching and learning, especially in mathematics and statistics. Teachers of statistics face the challenge in teaching quantitative reasoning skills through exploring data, while making theory and data analysis come alive and relevant rather than as just numbers on spreadsheets. Teachers need to utilize more visualizing techniques and software to help students explore data and learn statistical concepts. However, as I stated before, the effective use of any technology in teaching requires thoughtful consideration and planning. Asking question such as “Is there a specific technology that will enhance student learning?” is equivalent to asking whether the chalk board or the text book will enhance learning. A tool’s
learning benefits depend on when, where, how, and why you use it. “Visualization should help make sense of data that may have seemed previously unintelligible” (Stokes, 2005) and make it possible to take advantage of human perceptual system and evaluate whether visualizations do a better job than other methods to assist the understanding of huge datasets.

Figure 3.2 Schema of process of interpretation and understanding of information via data visualization. Source: Graphic made by Dürsteler. J. from “The digital magazine of InfoVis.net”

A review of popular interactive web-based information visualization tools identifies the following software that have been used frequently by many educators and researchers.

1. **Gapminder**: this tool displays statistical data in graphical dynamic ways, allows users to create, and interact with a moving motion chart of multiple dimensional data. However, it does not allow users to import their own data. [www.gapminder.org](http://www.gapminder.org)

2. **Google Motion Chart API**: this software allows users to upload data from any source, from text files to full databases, and to see their data merged and compared in well-
designed visualizations.

3. **Many Eyes**: this software also allows users to create visualizations.

   http://www-958.ibm.com/software/data/cognos/manyeyes/visualizations

4. **Geographic Information System (GIS)**: This software permits users to capture, manage, analyze, and display information geographically.

5. **Google Earth**: This lets users view geographical space from different angles and distances through the use of superimposed satellite imagery. It also allows users to import publicly available datasets and map them geographically.

   http://www.google.com/earth/index.html


   This is an efficient, open-source, interactive, and dynamic applet that can be used for visualizing longitudinal multivariate data.

7. **JMP**: This is a powerful statistical software that can be used to analyze data. It can generate informative graphs quickly and easily. It can also link statistics with graphics, making information accessible in ways a spreadsheet never could.

There are other visualization tools designed and used specifically in the public health and global health areas. They include:

1. **Gapminder**: www.gapminder.org (see description above)

2. **HealthMap**: this tool introduces the challenge of tracking disease outbreaks in real time.

   It provides an updated and interactive map of information on infectious disease outbreaks.

   www.healthmap.org

3. **Worldmapper**: provides global maps with countries reshaped according to the prevalence of a variety of conditions such as malaria, infant deaths, or even Internet users.

   www.worldmapper.org
4. **Food Environment Atlas**: This tool was produced by the Economic Research Service of the US Department of Agriculture. It has interactive maps of county-level data about the social determinants of health, food availability and pricing, physical activity, and selected health measures. [http://www.ers.usda.gov/foodatlas/](http://www.ers.usda.gov/foodatlas/)

In addition to its many advantages, data visualization does have many limitations of which novice users may not be aware. We discuss the shortcomings of data visualization next.

**Limitations of Data Visualization Tools**

It is important to understand and evaluate some limitations of these visualization tools to help users understand and evaluate their decisions when it comes to utilizing data sets from these software options. We provide here a discussion of the three most important limitations.

First and foremost, most of the large data sets from governmental agencies are usually aggregate data (at the national level), which do not show the variability in the data and frequently are missing a great deal of the important regional information. This fact is not at all obvious to users/students; therefore, we recommend that clear and concise explanations be given to learners of statistics.

Second, the value of the graphics depends on the quality of the data. Thus the quality of data must be inspected before creating and utilizing the visual image. It is important to investgate and gain as much information as possible on who collected the data, where they were collected, and how and when the data were collected. We all should be reminded not to accept information at face value.

Third, information visualization tools can be powerful teaching resources, but they will not enhance learning automatically. As data visualization tools become more abundant with the advance of technology, teachers will need to provide students with the skills to interpret and understand the information. Teachers need to be trained in using the software, and learn about its limitations and how to integrate it in teaching. Educators need to help students in identifying the limitations and biases of the data as they construct arguments.

Since Gapminder and JMP are utilized in our study, we review how each software package has been used in teaching.
Teaching with Gapminder and JMP

To utilize the tools we first need to understand what they are, what they can be used for, and their limitations. We provide here a brief overview of each software package, their capabilities, and their limitations.

About Gapminder

Gapminder is a web-based data visualization tool that was created by Hans Rosling, a professor of International Health at Karolinska Institute and director of the Gapminder Foundation. It is a web-service tool that allows users to see all sorts of interactive global statistics ranging from wealth and health of the nations to home, schools, education, human development, and many other topics. Gapminder was launched in 1998 to provide and enhance the public’s understanding of world health. The most important feature of this graphing software is its ability to illustrate changes over time of multiple indicators using moving graphics. The data consumer will benefit from this important information and will access a much richer picture of the data than tables or a series of static graphs can provide. A typical Gapminder display can show a number of countries, their regional location, their population, their gross domestic product per capita, and the life expectancy of their citizens, year by year, from 1800 through 2010. With Gapminder the data are displayed dynamically before your eyes. The software contains a collection of almost 500 different data sets available for anyone to use. Data are updated and added on a regular basis. As a result, many topics of interest can be investigated by selecting a set of indicators and observing the interactions. Some of the most interesting videos created by Gapminder can be viewed from TED talks (www.ted.com) or at http://www.gapminder.org/videos/.

Gapminder as a Teaching Tool

To start, as a teacher one should explore the Gapminder site, http://www.gapminder.org/forteachers/. This site provides teachers with ideas, lesson plans, and resources to incorporate Gapminder in their classrooms. The second place to view its use in action is the Gapminder
course at the NYC iSchool, where 10th and 11th graders were challenged to use a quantitative lens to analyze global developmental data from 1800 until today [http://tinyurl.com/7cd5dn3](http://tinyurl.com/7cd5dn3).

The Gapminder website provides excellent information, many lesson plans for educators to teach using Gapminder, and many informative short 10-20-minute videos to engage students easily. More and more educators around the world have utilized Gapminder in their teaching. However, empirical research on the benefits of this software in teaching (whether it will enhance student learning) has not been conducted or published. Its use in teaching statistics is still limited due to two primary reasons. First, Gapminder is used for creating motion pictures but not for analyzing data. Second, designing activities for effective use of Gapminder in teaching, like any other forms of technology, is time-consuming as it requires thoughtful consideration and planning. Thus, the benefits of Gapminder in learning statistics depends on how, when, where, and why you use it.

**About JMP**

JMP is a flexible general statistical software package that provides visual data analysis that can link statistics with graphics and make information accessible in ways a spreadsheet never could. JMP can be used for data management and database queries. Its visual interface supports exploratory data analysis to gain a deeper understanding of data types, distribution, and general data representation. In addition, JMP can generate informative graphs swiftly and easily. One can also quickly create customized analytical applications that can be shared, allowing everyone to collaborate and interact with the data. Using JMP as an analytic hub, one can easily work with SAS®, Microsoft Excel, and R from within JMP itself. This paper discusses the relevant facts about JMP [http://www.jmp.com/software/jmp10/pdf/jmp10_fact_sheet.pdf](http://www.jmp.com/software/jmp10/pdf/jmp10_fact_sheet.pdf). There are many reliable online resources for learning JMP. SAS incorporation has its own site that provides support for teaching with JMP [http://www.jmp.com/academic/academic-resources.shtml](http://www.jmp.com/academic/academic-resources.shtml).

The above discussion explains the usage and advantages of utilization tools in teaching and learning statistics. Abstract or more difficult-to-understand concepts can be illustrated visually to help enhance students learning experience especially in introductory statistics courses. While technology has become a useful component in teaching statistics students’ attitude and
engagement also need to be reviewed because they may influence the learning process. This next section alerts educators to the importance of assessing student attitudes and beliefs regarding statistics.

### 3.2.3 Students’ Attitudes Toward Statistics

Many researchers have written about students’ apprehension about statistics. Students’ anxiety and belief in their ability to learn statistics can affect their performance in statistics courses (Mills, 2004, Gal and Ginsburg 1994). Thus, addressing students’ attitudes toward statistics and ways to enhance them have continued to be discussed actively by many statistics instructors.

As mentioned in the introduction chapter, previous research has indicated that these introductory courses have rather a poor image in our society and schools. Many students, especially those majoring in social science disciplines, get the impression that these courses are the most difficult and least enjoyable. Field (2000, p. xiv) states that "since time immemorial, social science students have despised statistics." Generally speaking, students entering these disciplines (social sciences) often do not have a strong mathematics background and often dislike anything “mathematical”. They find it hard to apply the statistical methods they learn to advance their research and careers. They lack the exposure to large-scale national or international datasets to investigate social issues. Some questions for statistic educators are: “How can we make our courses interesting, relevant, and even fun for students?” “How can we leverage students interest and steer their enthusiasm into rigorous applications of statistical methods?” “How can we motivate students to actually apply statistical concepts in their practice and thus advance the profession?” And, most importantly “How can we engender in students a positive view of statistics and an appreciation for the potential uses of statistics in future personal and professional areas relevant to each student?” (Gal, Ginsburg, & Schau, 1997, pp. 37-51)

Instructors have explored different ways to improve the teaching and learning of statistics. Many have incorporated technology in their teaching to motivate and enhance student understanding and learning of statistical concepts. Motivating and engaging strategies only work when they enable learners to actively take part in the learning process and when there is a
strong connection between the subject and the student’s interest. Wallman (1993, p.1) has defined statistical literacy as “The ability to understand and critically evaluate statistical results that permeate our daily lives – coupled with the ability to appreciate the contributions that statistical thinking can make in public and private, professional and personal decisions.”

Research has indicated that students’ attitudes and beliefs can impede or assist learning statistics and may affect their willingness to apply what they have learned (Gal, I. & Garfield, J., 1997). Consequently, researchers have given more credence to the crucial role that the affective domain has in both teaching and learning, which has also spurred an increase interest in this important area of inquiry. Figure 3.3 represents the “communication” between the two domains of learning, affective and cognitive. Researchers are increasingly looking at some important questions related to students’ attitudes, values, biases, interests, perceptions, beliefs, etc., and their relationship to teaching and learning.

Moreover, it is generally agreed that there is a connection between mathematics proficiency, previous experience in mathematics, and the ability to learn statistics. Beliefs and attitudes related to math may play a powerful role in affective responses to statistics since students often expect that the study of statistics will include a heavy dose of mathematics, including complex algebra and formulae (Simon & Bruce, 1991). Many students identify statistics with mathematics as we often hear statements expressed by some students in the introductory statistics course, *I am not good at math, and therefore I can’t learn or I fear statistics.*

As expected, the literature indicates that results from most research on attitude in statistics show more positive attitudes among students with better math grades and more math-oriented
education background (DeVaney, 2010, Pimenta, Faria, Pereira, Costa, & Vieira, 2010). However, this pattern was not homogeneous across the dimensions of attitudes toward statistics which include interest, appreciation, and perspective on the difficulty level of the subject. Students may not like statistics but still can appreciate and see the significant contribution statistics can have in their professional lives. Thus demystifying the connection between math ability and learning statistics may help increase positive attitudes toward statistics. Teachers of introductory statistics should model and convey the fact that if theoretical statistics depends heavily on mathematics, the practice of statistics does not, or very little indeed. This is not to downplay the importance of mathematical thinking, but to emphasize the importance of the applications of statistics as applications of statistics span many fields and disciplines. An average person who practices statistics does not need to know advanced mathematics (i.e. concepts beyond college algebra). Statistics thinking and applying should be a tool to advance research in all fields regardless of their connection with mathematics. As one student put it, “I never understand math but statistics makes sense”. One can learn and use statistics without having a strong background in math.

In the post “NSF should understand that Statistics is not Mathematics” on the blog “Simply Statistics”, http://simplystatistics.org/ on January 11, 2013, Rafael Irizarry, a professor in the Department of Biostatistics in the Johns Hopkins Bloomberg School of Public Health, argued that “Statistics is analogous to other disciplines that use mathematics as a fundamental language, like Physics, Engineering, and Computer Science. But like those disciplines, statistics contributes separate and fundamental scientific knowledge. ... Although statisticians rely heavily on theoretical/mathematical thinking, another important distinction from mathematics is that advances in our field are almost exclusively driven by empirical work. Statistics always starts with a specific, concrete real world problem.” This supports our teaching philosophy for introductory or service-level statistics courses, in which a strong emphasis is placed on using concrete real-world applications rather than dwelling on theoretical or mathematical concepts.

In the next topic, we review some of the existing instruments used for assessing students’ attitudes toward statistics.
3.2.4 Existing Attitude Assessment Instruments

A few instruments related to statistics anxiety have been developed. Among these three popular but quite antiquated ones. These three instruments were designed for assessing students' attitudes towards statistics:

1. Survey of Attitudes Toward Statistics –SATS- (Schau, Stevens, Dauphinee, & Del Vecchio, 1995),

2. Attitudes Toward Statistics scale –ATS- (Wise, 1985), and


These instruments complement and improve on the perceived limitations of one another. The biggest drawback, of these instruments is that students are never asked to explain their answers to the Likert-type items. Research on survey methodology (DeVault, http://marketresearch.about.com/), has recommended utilizing a combination of both types of questions in designing a survey. This is because open-ended questions allow respondents to provide more information, including feelings, attitudes, and understanding of the subject. This permits researchers to better access the respondents' true feelings on an issue. Closed-ended questions, because of the simplicity and limited range and lack of detail of the answers, may not offer respondents opportunities to reflect their real feelings. Closed-ended questions also do not allow the respondents to explain if they do not understand the question or do not have an opinion on the issue. Thus more can be learned from responses to Likert-type scales when they also include some open-ended procedure to enable respondents to elaborate on their initial answers. This study has adapted a shorter version of the SATS instrument in the pre- survey and combined selected SATS items with our own custom-made closed-ended and open-ended questions.

3.3 Study Setting and Methods

To understand the effect of the visualization tools in teaching statistics, we incorporated a research project that utilized the instructional visualization software known as Gapminder, in
conjunction with lectures and labs designed to focus on graphical analysis and visualizing data. In this project students were required to use Gapminder and JMP, a data analysis software, to investigate topics related to a global issue in which they were interested. The project was introduced during the first week and ended after the fifth week (January 17 to February 28, 2013). We conducted two surveys, pre and post, on students’ attitudes to evaluate the change, if any, in attitudes over the six-week period. Some of the questions in the pre survey instrument on students’ attitude were adapted from the well-known survey on Students’ Attitude Toward Statistics (SATS).

3.3.1 Study Setting/Description of the Course

The course, “Statistical Methods for Research Workers”, is a service-level statistics course for non-statistics majors, enrolling primarily graduate students and was conducted in a blended learning format. Students registered in the class are expected to have taken or to have experience with a prior undergraduate elementary statistics course. However, based on the survey results, for most students, this was their first course. This lack of exposure may have afforded us the best opportunity to positively impact their attitudes toward statistics because they may not have developed a negative image about statistics. Students could register in an online or face-to-face environment. The course consisted of two structural elements–lectures and labs. All lectures were pre-recorded and posted weekly, while labs were conducted on a weekly basis, broadcasted live, and video-recorded. Students could attend class in person, view the live lab session in real time directly from home, or watch video lectures at their convenience. The course materials were geared to primarily social sciences and education majors, but students from all majors, including advanced undergraduate students, were allowed to enroll. A variety of applied topics were covered, ranging from descriptive statistics to inferential statistics that included regression, analysis of variance, analysis of covariance, and repeated measures within-subjects designs.

3.3.2 Profile of Study Participants

The pre-survey participants were students from the statistics course described above. The group consisted of 84 students, of which 66 (79%) were graduate students and 18 (21%) under-
graduates. There were 67 different academic majors classified in two categories: STEM (48%) and Non-STEM (52%) STEM\(^1\) majors. There are slightly more males (51%) than females (49%). About 44% were enrolled on-campus and 56% were online; 75% enrolled full-time and 25% part-time. Participants were classified in four age-groups, 18-29, 30-39, 40-49, and 50-59, with the majority (54) in the first group, 18 in the second, 8 in the third, and 4 in the oldest age group. For most students, this was a required course, but some registered because the course fit their interest. The majority expected or planned to work hard to earn an “A” grade. More than half (55%) identified themselves as visual learners, 39% tactile focused on learning by doing, and only 6% as auditory.

3.3.3 The Study Instruments

Two online surveys were conducted, at the beginning of the course and six weeks after. The pre-survey was designed to measure students’ initial attitudes toward statistics and their general engagement in college. The post-survey was constructed to assess the change in attitudes towards the subject after six weeks of involvement in the course and being exposed to the visualization-emphasis and technology-immersion teaching method. The pre-survey was formed by adapting questions from two established online surveys (see Appendix C.1). The questions in the post-survey are more specific to students learning experience and involvement in the class after a six-week period. For that reason we formed the questions more directly relevant to students' experience in the course rather than reusing the items from the pre-survey instrument.

There are four major components in the pre-survey: (1) demographic information (12 questions), (2) student attitudes toward statistics (24 questions), (3) students’ level engagement in learning (24 questions), and (4) technology (4 questions). For the attitude items, we adopted 24 out of 36 questions from the original SATS survey. Since we evaluated students' attitudes, and not engagement in this study, only questions on attitudes were included. Although the possible range for each of the original items from SATS is between 1 and 7, the range was structured to be from 1 to 5 for ease of analysis and interpretation while keeping the scale balanced with a neutral option. Specifically, the response options are: 1 = Strongly Disagree, 2 = Disagree, 3 =

\(^1\)STEM majors are Science, Technology, Engineering, and Math majors.
Neither Disagree nor Agree, 4 = Agree, and 5 = Strongly Agree. We preserve the six attitude scales of the SATS listed above: Affect, Cognitive Competence, Value, Difficulty, Interest, and Effort. Table 3.1 displays the description and the numbers of items adopted from the original SATS survey for each attitude subscale.

Table 3.1 Descriptions of the six subscales of the attitude scale and the number of items adopted from the original SATS survey in each subscale.

<table>
<thead>
<tr>
<th>Attitude Scale</th>
<th>Measure</th>
<th>Number of items adopted from SATS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affect</td>
<td>Positive or negative feeling concerning statistics</td>
<td>4 out of 6</td>
</tr>
<tr>
<td>Cognitive Competence</td>
<td>Attitudes about own intellectual knowledge and skills when applied to statistics</td>
<td>4 out of 6</td>
</tr>
<tr>
<td>Value</td>
<td>Attitudes about the usefulness relevance, and worth of statistics in personal and professional life</td>
<td>6 out of 9</td>
</tr>
<tr>
<td>Difficulty</td>
<td>Attitudes about the difficulty of statistics as a subject</td>
<td>4 out of 7</td>
</tr>
<tr>
<td>Interest</td>
<td>Level of individual interest in statistics</td>
<td>3 out of 4</td>
</tr>
<tr>
<td>Effort</td>
<td>Amount of work the student expends to learn statistics.</td>
<td>3 out of 4</td>
</tr>
</tbody>
</table>

The post-survey also contained four main components, as in the pre-survey. However, significant changes were made to three of the four components (attitude, engagement, and technology) to assess attitude change over time. In the following section the components of the pre- and post-surveys are presented together in the tabular form and the change of questions are discussed in detail. Tables 3.2 to 3.5 contain the contents of the demographic, attitude, and technology components.

Specifically, the first group of questions is on the demographic component, for which 9 out of the 12 original SATS items (items 1 to 8 and 12) remain the same. The other three items (9, 10, and 11) were modified to evaluate the change over the six-week period of the course. Table 3.2 lists all demographic questions in pre- and post-surveys. For example, question 10 was changed from “What challenges do you anticipate in this course?” in the pre-survey to “What challenge did you face at this juncture of the course? Has your anticipation changed over the course of 6
weeks?” In question 11, “What grade do you expect to receive in this course?” in the pre-survey was changed to “Did your grade so far match your expectation? Is it higher or lower?” in the post-survey. The modified items were anticipated to provide information on students’ progress in the course.

Table 3.2 Demographic questions in the pre- and post-surveys

<table>
<thead>
<tr>
<th>Question</th>
<th>Content</th>
<th>Survey</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>Are you a Graduate or Undergraduate?</td>
<td>Pre-Post</td>
</tr>
<tr>
<td>Q2</td>
<td>What degree are you seeking?</td>
<td>Pre-Post</td>
</tr>
<tr>
<td>Q3</td>
<td>What is your major or academic discipline? Second major (if applicable)?</td>
<td>Pre-Post</td>
</tr>
<tr>
<td>Q4</td>
<td>What is your gender?</td>
<td>Pre-Post</td>
</tr>
<tr>
<td>Q5</td>
<td>Which best describes you? (Ethnicity)</td>
<td>Pre-Post</td>
</tr>
<tr>
<td>Q6</td>
<td>What is your age group?</td>
<td>Pre-Post</td>
</tr>
<tr>
<td>Q7</td>
<td>Are you registered as an on-campus or online student?</td>
<td>Pre-Post</td>
</tr>
<tr>
<td>Q8</td>
<td>Are you enrolled as a full-time or part-time student?</td>
<td>Pre-Post</td>
</tr>
<tr>
<td>Q9</td>
<td>Why are you taking this course?</td>
<td>Pre</td>
</tr>
<tr>
<td>Q10</td>
<td>Have you ever taken any statistics courses before?</td>
<td>Post</td>
</tr>
<tr>
<td>Q11</td>
<td>What challenges do you anticipate in this course?</td>
<td>Pre</td>
</tr>
<tr>
<td></td>
<td>What challenges did you face at this juncture of the course?</td>
<td>Post</td>
</tr>
<tr>
<td></td>
<td>Have your expectations changed over the course of the first six weeks?</td>
<td></td>
</tr>
<tr>
<td>Q12</td>
<td>What grade do you expect to receive in this course?</td>
<td>Pre</td>
</tr>
<tr>
<td></td>
<td>Did your grade so far match your expectation? Is it higher or lower than</td>
<td>Post</td>
</tr>
<tr>
<td></td>
<td>you expected it?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>What best describes your learning style?</td>
<td>Pre-Post</td>
</tr>
</tbody>
</table>

The second group of items consists of questions on attitude that were modified to reflect student experience and attitude change after several weeks of being taught under the new teaching method. Table 3.3 contains the pre- and post-survey questions on attitude. Due to having a different focus, the six original constructs in the pre-survey in Table 3.1 are no longer retained. Instead, the post-survey consisted of questions tailored to measure specific improvement in attitude, feeling, and perspective about the subject at this juncture of the course. For instance, in Q17, the students were asked “Is your attitude towards statistics more positive after six weeks in the course?” Responses to this question would inform us whether any expected change really occurred. This group consists of questions on a 5-point Likert scale
and three open-ended or multiple-choice questions where students can list items or resources that are helpful and stimulate them to learn statistics. It was hoped that the new format and content of these questions provide detailed information on variables affecting student attitude change.

The third group of items consisted of questions on technology that were geared toward measuring student self-reported proficiency with the technology used in class, Gapminder and JMP in particular. The pre-survey questions were aimed to gauge students’ familiarity with Gapminder, statistical software, and their choice of favorite software. Table 3.5 list the technology questions from the pre-survey and Table 3.6 list the technology questions from the post-survey.
Table 3.3 Questions on *attitude* in the pre-survey.

<table>
<thead>
<tr>
<th>Content of sub-questions (Type of Subscale)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. I can learn statistics. (Cognitive Component1)</td>
</tr>
<tr>
<td>2. I understand statistical equations. (Cognitive Component2)</td>
</tr>
<tr>
<td>3. I find it difficult to understand statistical concepts. (Cognitive Component3)</td>
</tr>
<tr>
<td>4. I have trouble understanding statistics because of how I think. (Cognitive Component4)</td>
</tr>
<tr>
<td>5. I like statistics. (Affect1)</td>
</tr>
<tr>
<td>6. I feel insecure when I have to do statistics problems. (Affect2)</td>
</tr>
<tr>
<td>7. I am afraid of statistics. (Affect3)</td>
</tr>
<tr>
<td>8. I enjoy taking statistics courses. (Affect4)</td>
</tr>
<tr>
<td>9. I use statistics in my everyday life. (Value1)</td>
</tr>
<tr>
<td>10. I will not need statistics in my profession. (Value2)</td>
</tr>
<tr>
<td>11. Statistics is irrelevant in my life. (Value3)</td>
</tr>
<tr>
<td>12. Statistical skills will make me more employable. (Value4)</td>
</tr>
<tr>
<td>24. Statistical thinking is not applicable in my life outside my job. (Value5)</td>
</tr>
<tr>
<td>17. Statistics should be a required part of my professional training. (Value6)</td>
</tr>
<tr>
<td>13. I plan to work hard in my statistics course. (Effort1)</td>
</tr>
<tr>
<td>20. I plan to complete all of my statistics assignments. (Effort2)</td>
</tr>
<tr>
<td>16. I plan to spend extra time on this course. (Effort3)</td>
</tr>
<tr>
<td>19. I am interested in being able to communicate statistical information to others. (Interest1)</td>
</tr>
<tr>
<td>21. I am interested in using statistics to understand the world. (Interest2)</td>
</tr>
<tr>
<td>23. I am interested in learning statistics. (Interest3)</td>
</tr>
<tr>
<td>14. Most people have to learn a new way of thinking to do statistics. (Difficulty1)</td>
</tr>
<tr>
<td>15. Learning Statistics requires a great deal of skills and discipline. (Difficulty2)</td>
</tr>
<tr>
<td>18. Statistics is a subject quickly learned by most people. (Difficulty3)</td>
</tr>
<tr>
<td>22. Statistics is a complicated subject. (Difficulty4)</td>
</tr>
</tbody>
</table>
Table 3.4  Questions on *attitude* in the post-survey.

<table>
<thead>
<tr>
<th>Post-Survey: Content of new or modified questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q13. <strong>Attitude toward statistics</strong></td>
</tr>
<tr>
<td>1. I now see the value of learning statistics.</td>
</tr>
<tr>
<td>2. My attitude about statistics become more positive as the course progresses.</td>
</tr>
<tr>
<td>3. My anxiety about statistics has reduced so far.</td>
</tr>
<tr>
<td>Q14. <strong>Your feeling about statistics at this juncture of the course</strong></td>
</tr>
<tr>
<td>1. After experiencing the course work, I plan to spend more time studying for this course.</td>
</tr>
<tr>
<td>2. The course has improved my understanding of statistical concepts.</td>
</tr>
<tr>
<td>3. I feel that if I work hard I can do well in this course. This course helps preparing me for my research involving data analysis.</td>
</tr>
<tr>
<td>4. I like statistics more than before.</td>
</tr>
<tr>
<td>5. I like statistics teaching methods that emphasizes data visualization.</td>
</tr>
<tr>
<td>Q15. List two or three specific items about this course that most stimulated your study of statistics.</td>
</tr>
<tr>
<td>Q16. List two or three specific items about this course that most helped your understanding of statistics.</td>
</tr>
<tr>
<td>Q17. Is your attitude towards statistics more positive after six weeks in the course?</td>
</tr>
</tbody>
</table>
Table 3.5  Questions on technology in the pre- survey.

<table>
<thead>
<tr>
<th>Pre-survey: Part IV: Questions on Visualization Tools (open-ended)</th>
</tr>
</thead>
</table>

1. Have you ever heard of Professor Hans Rosling and his data visualization tool, Gapminder? If yes, please explain in which context/situation you encountered Gapminder?

2. What is your favorite educational gadget or software?

3. In terms of using statistical data analysis software such as JMP, SAS, SPSS, or R are you a beginner? Intermediate user? Or advanced user?

4. Given a choice of a point-click menu-driven statistical software or a program statement and command interface software, what would you choose?
Table 3.6 Questions on technology in the post survey. These questions are specific to the technologies used in the class.

<table>
<thead>
<tr>
<th>Post-survey Part IV: Questions on Technology/Software Experience (rating + open-ended)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q24. List two or three specific technology items we use in this course that are most helpful to you in learning statistics.</td>
</tr>
<tr>
<td>Q25. Did you like the idea of using web-based technology such as Gapminder in learning statistics? Please elaborate.</td>
</tr>
<tr>
<td>Q26. How proficient do you think you are with the following software? (i.e. being able to do what you want to do with them): So so, Good, Very Good</td>
</tr>
<tr>
<td>Gapminder, JMP, MyLabPlus, Jing, Other.</td>
</tr>
<tr>
<td>Q27. Do you think using JMP rather than SAS or any programming-based statistical software help you in learning statistics in a better manner?</td>
</tr>
<tr>
<td>Q28. How proficient do you think you are with JMP so far? Are you able to use JMP to do your homework and create graphics such as histogram, scatter plot, charts, etc?</td>
</tr>
<tr>
<td>Q29. On the scale from 1 to 5, from least (1) to most favorite (5), rate the following components?</td>
</tr>
<tr>
<td>Gapminder, JMP, MyLabPlus, Jing, Blackboard, Gapminder Wiki, Lecture Videos, Lab Videos. Other, please specify.</td>
</tr>
</tbody>
</table>

3.3.4 Data Sources and Procedures

In compliance with the human subjects in research protection protocols, an exemption was obtained from the Institutional Review Board before implementing the survey. The primary source of data for this study is the pre- and post-surveys conducted before and after the completion of the class project focused on data visualization and technology. We offered students a minimal amount of extra credit (1% of course grade for each survey) for participating in the surveys when the response rate reached 90%. Participation was completely anonymous, but
students’ IP addresses were collected in the post-survey to avoid the possibility of duplicate surveys from the same individuals. Future research will explore the option of tailoring a unique survey link for each IP address, thereby ensuring the possibility of a single survey entry per respondent.

The response rate for pre-survey reached 93% (84 out of 90 students enrolled in the course), while the rate for the post-survey was 92% (82 out of 89). Although 19 records of duplicate email addresses were found, all respondents were retained in the dataset. The rationale for doing this was twofold: multiple students may have used the same public computers that recorded the same IP addresses, and because it was not possible to monitor the IP address in the pre-survey there was no practical way to track IP addresses between pre- and post-surveys.

In addition to the two surveys, data were collected from students’ weekly submissions, opinion papers, videos, data files, and final research papers for the Gapminder project. However, this study uses only data from the two surveys. Future research will make use of the additional data.

3.3.5 Study Variable Measures

Participants were asked to provide their demographic information such as gender, age group, ethnicity, major of study, degree sought, enrollment format (online or face-to-face, and full-time versus part-time), learning style, grade expectation, and anticipation of challenges in the course. To assess students’ attitudes towards statistics, latent attitude variable were constructed for each survey. Since many questions were different between the pre- and post-surveys, the number of latent constructs and the meaning of each construct may be different in the two surveys. The analysis also utilizes information from the open-ended questions.

3.3.6 Analysis Methods

To assess and evaluate the change in students’ attitudes towards statistics we constructed a latent variable for attitudes from each survey from the questions related to attitude. The statistical methods of analyses consist of descriptive statistics, item analysis, and Item Response
Theory. Qualitative analysis was employed to analyze response data from open-ended and multiple-choice questions.

3.4 Data Analysis

Quantitative analysis was conducted on numerical responses to Likert-type response questions data and qualitative analysis was conducted on text responses to open-ended survey questions to evaluate the change in students’ attitudes about statistics.

3.4.1 Quantitative Data Analysis

Data Summary

The pre- and post-surveys on attitudes and engagement in statistics were administered to the students in Statistics 401D-XW at Iowa State University in Spring 2013. This is a statistical methods course taken largely by graduate students with majors from 67 disciplines, primarily in the social and behavioral sciences and education. Out of 92 enrolled students, 84 responded to the pre-survey and 80 responded to the post-survey. The surveys were online, voluntary, and completely anonymous; as such, no identification of individuals was possible. The surveys can be viewed in the Appendices C.1 and C.2.

In this analysis, we extracted data from student responses to the section on attitude from both surveys. The questions asked respondents to indicate on a 5-point scale (1=Strongly Disagree, 2=Disagree, 3=Neutral, 4=Agree, and 5=Strongly Agree) their level of agreement. For all items, higher scores indicate greater agreement with the content. To facilitate comparison and evaluation of attitude change, the questions were separated into two sets. The first set, labeled as perfectly-matched questions, consists of six items that were included in both surveys. The second labeled as Question Set 2, consists of nine questions asked only in the post-survey, all related to the student change or shift in attitude over the course of six weeks in the class. Tables 3.7 and 3.8 display the two sets of questions.
Table 3.7 Pre- and post-survey perfectly-matched question set. These questions are later on grouped into two primary constructs: Interest-Value-Affect and Difficulty.

<table>
<thead>
<tr>
<th>Perfectly-matched Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. I can learn Statistics. (Affect1)</td>
</tr>
<tr>
<td>2. Statistics should be a required part of my professional training. (Value6)</td>
</tr>
<tr>
<td>3. I am interested in using statistics to understand the world. (Interest1)</td>
</tr>
<tr>
<td>4. I am interested in learning statistics. (Interest2)</td>
</tr>
<tr>
<td>5. Statistics is a complicated subject. (Difficulty1)</td>
</tr>
<tr>
<td>6. Learning Statistics requires a great deal of skills and discipline. (Difficulty2)</td>
</tr>
</tbody>
</table>

Table 3.8 Nine post-survey items in Set 2. These items contain elements of attitude change, shift, or progress related to statistics. They are categorized into five attitude components: Affect, Interest, Value, Cognitive Competence, and Effort. The label of each item is based on the subscale the item belongs to and its order in the subscale.

<table>
<thead>
<tr>
<th>Post Survey Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. I now see the value of learning statistics. (Value)</td>
</tr>
<tr>
<td>2. My attitude about statistics become more positive as the course progresses. (Affect)</td>
</tr>
<tr>
<td>3. My anxiety about statistics has reduced so far. (Cognitive Competence)</td>
</tr>
<tr>
<td>4. I am enthusiastic about using statistics. (Interest)</td>
</tr>
<tr>
<td>5. I like statistics more than before. (Affect)</td>
</tr>
<tr>
<td>6. After experiencing the course work, I plan to spend more time studying for this course. (Effort)</td>
</tr>
<tr>
<td>7. The course has improved my understanding of statistical concepts. (Cognitive Competence)</td>
</tr>
<tr>
<td>8. I feel that if I work hard I can do well in this course. (Effort)</td>
</tr>
<tr>
<td>9. This course helps preparing me for my research involving data analysis. (Value) (Interest)</td>
</tr>
</tbody>
</table>
The next two sections describe the analysis and results for each set of questions.

3.4.1.1 Analysis of Question Set 1 \textit{(perfectly-matched)}

The first set of questions consists of six items, each with five response options as stated previously. These items were intended to measure each student’s \textit{attitude} about statistics at the beginning of the course and how \textit{attitude evolved} over the first six weeks of experiencing the course. Since \textit{attitude} was not a variable directly observed, it was treated as a latent variable manifested from a pattern of responses to multiple items. Item Response Theory (IRT) was utilized to estimate this variable because IRT takes into account the interaction between the items and the respondents by using the response pattern of each person. Rather than using a simple average, which ignores individual responses and only calculates group responses, IRT extracts a latent score for each individual and uses it as the latent attitude score. Before performing the IRT analysis it is important to examine the data for the validity of the underlying IRT assumptions of dimensionality and local independence of the items.

\textbf{Dimensionality} \hspace{1em} The first assumption of IRT relates to the dimensionality of the items in an instrument. This assumption implies that the construct being measured is in fact unidimensional, that is the covariance among the items can be explained by a single underlying dimension. Techniques such as Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) can be used in conjunction to evaluate the assumption. Alternatively, comparing unidimensional and multidimensional IRT models can also be helpful for assessing the dimensionality of the data.

Figure 3.4 displays the EFA and CFA’s graphical output which includes the eigenvalues, scree plots, and factor loadings. The scree plot reveals two relatively large eigenvalues (2.61 and 1.24). This indicates that more than one latent variables may exist. The CFA factor loadings appear to fall into distinct groups with items \textit{difficulty1} and \textit{difficulty2} forming their own group, while the other items falling into a second group. However, the 20 simulated scree plots (the lines bunched together) seem to suggest one dominant dimension. The simulation was performed by simulating random normal data with the same number of subjects and items.
as in the original data set. An indeterminant situation in deciding the number of dimensions is commonly encountered in performing item analysis for dimensionality. One can conduct more sophisticated tests to ensure the dimensionality but it is not possible here due to the short instrument and the small number of participants. IRT was utilized to help identify the number of factors present in the scale.

Figure 3.4 Display of scree plot, eigenvalues, and confirmatory factor analysis results. The lines bunched together are 20 screeplots obtained from simulated random normal data. The items’ abbreviations are Affect1 = I can learn Statistics; Value6 = Statistics should be a required part of my professional training; Interest2 = I am interested in using statistics to understand the world. Interest3 = I am interested in learning statistics; Difficulty1 = Statistics is a complicated subject; and Difficulty2 = Learning Statistics requires a great deal of skill and discipline.

To guide the item selection, we examined the item information functions from a unidimensional graded response model IRT model and the factor loadings from a two-dimensional IRT model. Figure 3.5 displays the plot of the information curves from the UIRT model (left) and the factor loadings (right) from a two-dimensional IRT model. The information plots of the items difficulty1 and difficulty2 lie flattened at the bottom, indicating that they do not contribute any information about the estimate of the latent variable and should be excluded.
from the list. The other four items have higher and distinct information curves and appear to be better contributors. The factor loadings from the two-dimensional IRT model indicate the same two group distinction. Consequently, we proceeded to form two subscales, the first one containing four items (affect1, value6, interest3, and interest2) and the second one containing the two items difficulty1 and difficulty2. These scales were labeled as pre attitude-interest and pre attitude-about-difficulty respectively.

Figure 3.5  Left: Information curves obtained from a graded response unidimensional IRT model. Right: Factor loadings from a graded response two-dimensional IRT model. Both plots show evidence of two factors, with the last two items (difficulty1 and difficulty2) forming their own groups.

To facilitate a comparison of the information obtained from the pre- and post-surveys, we constructed two similar subscales for the post-survey and labeled them as post-attitude-interest and post-attitude-about-difficulty. As confirmation, item analysis was conducted and the results (not included here) were similar to those for the pre-survey, with two groups formed. In the discussion that follows each scale is evaluated separately.
**IRT analysis of the pre- and post-attitude-interest Scale**  The IRT calibration of the four items in the pre-attitude-interest scale was performed under the reduced and full graded response models (common slope and unique slopes for all items respectively). The likelihood ratio test rejected the null hypothesis of the reduced model ($p$-value = 0.001) indicating that the model includes four unique slope parameters. Results for the goodness of fit of the model using the Kullback-Leibler discrepancy procedure confirm the fit of the unique slopes model.

Table 3.9 shows the estimated item category parameters, item location and discrimination (or slope), for the pre-survey. The *slope* represents the rate at which the response probability changes as *attitude* increases. The *thresholds*, each indicates the points at which a respondent with a specific *attitude* score has a 50:50 chance of choosing the designated option or higher on an item. The item location is the mean of the thresholds and indicates the overall difficulty of the item. All four items are located below zero, with item 2 being farthest and was thus considered the easiest to endorse while item 1 was the hardest. For all items, everyone has a 100% chance of choosing *Strongly Disagree* or higher, so there is no threshold for that option. For item 1, the probability of choosing at least the *Disagree* category is 50% for respondents with $\text{attitude} = -2.127$; the probability of choosing at least the *Neutral* category is 50% for those with $\text{attitude} = -0.754$; the probability of choosing at least the *Agree* category is 50% for those with $\text{attitude} = 0.296$; and the probability of choosing the *Strongly Agree* category is 50% for those with $\text{attitude} = 2.409$. The *attitude* scores were assumed to have a normal distribution with zero mean and standard deviation of 1. The threshold can be interpreted relative to this distribution. For this survey all four items were considered “easy” for respondents to choose *agree* or *strongly agree*. This was especially true for items 3 and 4 since respondents only need an *attitude* higher than zero to *strongly agree*. This result indicates that at the beginning of the course students were very interested in learning and using statistics. The majority believed they could learn statistics and expected it as a required part of their professional training.
Table 3.9 Estimates of item parameters (category thresholds, item locations, and unique discriminations or slopes) for pre-survey. D=Disagree, N=Neutral, A=Agree, SA=Strongly Agree.

<table>
<thead>
<tr>
<th>Pre-survey</th>
<th>Threshold for</th>
<th>Location</th>
<th>Slope</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D</td>
<td>N</td>
<td>A</td>
</tr>
<tr>
<td>1. I can learn statistics</td>
<td>-2.13</td>
<td>-0.75</td>
<td>0.30</td>
</tr>
<tr>
<td>2. Statistics should be a required part</td>
<td>-3.69</td>
<td>-2.21</td>
<td>-0.81</td>
</tr>
<tr>
<td>3. I am interested in using statistics</td>
<td>-2.23</td>
<td>-1.32</td>
<td>-0.62</td>
</tr>
<tr>
<td>4. I am interested in learning statistics</td>
<td>-2.01</td>
<td>-1.46</td>
<td>-0.36</td>
</tr>
</tbody>
</table>

Similar analysis was conducted for the question set in the post-survey. The reduced or constrained graded response model with equal slopes was a better fit for this data set. Parameter estimate results for the post-survey are displayed in Table 3.10. Item 1 became the easiest to endorse, followed by items 2, 4, and 3, in that order. Similar to the pre-survey, all the items can be considered “easy” to endorse with a score of agree to strongly agree. However, it appeared “easier” for students to agree or strongly agree with items 1 and 2. This indicates that students seemed to be more confident about their ability to learn statistics after experiencing the course. However, their interest levels remained the same as before.

The item parameters can also be interpreted graphically as displayed in Figures 3.6 and 3.7. They are the plots of the probability of choosing a category option or higher for each item. The curves in Figure 3.6 for items 3 and 4 in the pre-survey have steeper slopes because the discriminations are higher while all the items have the same slopes in the post-survey. The category thresholds are not well spread across the *attitude* range for all four items as there are more thresholds below zero. Thus, this set of questions can measure lower *attitude* levels particularly well.
Figure 3.6  Plot of the OCC curves, the probability of choosing each category option or higher for all four items in pre-survey. The curves for items 3 and 4 have steeper slopes.
Figure 3.7  Plot of the probability of choosing each category option or higher for all four items in the post-survey. All the curves have an equal slope.

One can also view the response probability from the Item Category Curves displayed in Figures 3.8 and 3.9. Within each category, the curves represent the probability of an item being selected. The pre-survey curves showed more variation across the five categories than those in the post-survey. Those in the lower range of the attitude location \((\theta \leq -2)\) tend to disagree with all the items. The higher the attitude score, the more likely that respondents agreed with the last two items, *I am interested in learning statistics* and *I am interested in using statistics.*
Figure 3.8  Plot of category characteristic curves for the pre-survey. The probabilities of endorsing each category vary for each item.

The distributions of scores are more consistent for the post-survey. Figure 3.9 shows a uniform pattern in each category of responses. The curves appear to clump together, and form distinct segments on the $\theta$ scale. Proceeding from left to right, we observed that category 1 (strongly disagree), was selected by a very small group of respondents, those in the range of $\theta < -2$. Category 2 (disagree) yet receives more endorsements by those in the range $-2 < \theta < -1$. It is interesting to observe that the black curve was well below the other curves in this range. This indicates that there was some affection for statistics even though the attitude scores on interest or value level were still low. The sentiment for those in the range $-1 < \theta < 0$ tend to be neutral (category 3) on all issues. Finally, those with higher attitude level ($\theta > 0$) tend to show strong support (agree to strongly agree) for these items.

It was essential to be able to separate the groups based on their pattern of responses in the post-survey. After six weeks of experiencing statistics, students could be categorized into three distinct groups based on their attitude scores, disagree to strongly disagree, neutral, and agree to strongly agree. These groups corresponded with three attitude divisions: low ($\theta < -1$), neutral ($-1 \leq \theta < 0$), and high ($\theta \geq 0$).
Figure 3.9  Plot of category characteristic curves for the post-survey. The curves are tight and close together representing the items being uniformly rated.

The patterns described above can also be seen in the item characteristic curves in Figure 3.10. The plots in the post surveys are more alike as they appear to have the same pattern of spreading out. In general, respondents in the attitude range larger than zero tended to endorse all four items with a score of 4 or higher (agree to strongly agree).
Figure 3.10  Item characteristics curves for perfectly-matched items in the pre and post surveys. Overall categories 4 and 5 items dominate the whole spectrum.
The distribution of the attitude scores was estimated (using the Gauss-Hermite quadrature rule) along with the item parameters and placed on the same metric. The distribution was assumed to have a standard normal configuration. The histogram and density of the estimated distributions for pre- and post-attitude-interest scores along with the table of summary statistics are available in Figure 3.11. Both distributions appear to be normal, with scores concentrated around zero.

Returning to the question of interest, “Was there a change or shift in the attitude about statistics?, we examined the estimates for the latent scores attitudes. The two means (not displayed) are nearly the same (−0.01). The distributions of the post-scores seem to shift to the right, making the scores more positive. The interquartile ranges for the pre- and post-distribution are 1.37 and .96 respectively, indicating the scores were more concentrated in the post- than in the pre-survey. Overall, students’ attitudes have shifted toward being more positive for those in the first and second quartiles. The proportion of positive latent scores is higher in the post- than in the pre-survey as seen in the attitude panel of Figure 3.11. However, the observed change in latent attitudes on this value-interest scale appeared not to be substantial. The next set of questions will provide us more evidence of changes (if any) because the questions were tailored to evoke evaluation of individual changes over the six-week period.
Additionally, the precision of the estimates can be addressed via the item information function. Figure 3.12 displays the information curves for the scales. Each item gives more information near its threshold. In the pre-scale, more information can be gained from the last two items as their curves are much higher, showing different peaks and valleys. However, in the post-survey, all four items contributed approximately equally to the estimation of the latent scores as the curves are very much at the same level.
Figure 3.12 Item information curves for pre- and post-surveys.

The model estimates that the values of $\theta$ are more accurately in the range where they have more information. The standard errors of the attitude scores can be approximated by the square root of the inverse of the information function. The instrument information curve and standard error of the measurement are displayed in Figure 3.13. Overall, there is more information within the range of $-2$ to $1.5$. This is because the instrument information function is the sum of the item information functions; thus estimates of attitude scores for the survey are more accurate within that range.
Figure 3.13   Instrument information curves versus the standard error of measurement curves.

The instrument information function $I(\theta)$ is the sum of the individual item information functions $I(\theta_i)$. The standard error is: $SE(\theta) = 1/\sqrt{I(\theta)}$.

In summary, for this scale the model appears to estimate the scores with high precision. We observed a shift in attitudes toward “higher ground” and more positive for the middle groups of respondents. However, the change may not be very substantial.

**IRT analysis of the pre- and post-attitude-about-difficulty scale**   What does this scale tell us? Keep in mind that this is the scale related to the attitude about the difficulty level of statistics. Descriptive statistics are displayed in Table 3.4.1.1. Separating the responses into three categories of Disagree, Neutral, and Agree, there appear to be no significant changes in each group. This indicates that in the pre-survey more respondents agreed to strongly agreed that “Statistics is a complicated subject.” and “It requires a great deal of skills and discipline.” This sentiment did not appear to be different based on data from the post-survey. As there are only two items in this scale, IRT analysis may not be accurate and therefore was not employed.
Proportions for each level of response

<table>
<thead>
<tr>
<th>Statistics is a complicated subject</th>
<th>D</th>
<th>N</th>
<th>A</th>
<th>D</th>
<th>N</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.09</td>
<td>.31</td>
<td>.59</td>
<td>.16</td>
<td>.23</td>
<td>.43</td>
</tr>
<tr>
<td>It requires a great deal of skills and discipline</td>
<td>.07</td>
<td>.26</td>
<td>.67</td>
<td>.05</td>
<td>.23</td>
<td>.55</td>
</tr>
</tbody>
</table>

Table 3.11 Table of proportions of the categories for two items on difficulty level. The values stay approximately the same from pre- to post-survey.

### 3.4.1.2 IRT Analysis of Question Set 2

The nine post-survey questions comprising question set 2 all contain elements of change or shift in attitude, and thus can be used to assess the attitude change over the course of six weeks. These questions form the six constructs identified in the SATS instrument by Schau (2003): Affect, Cognitive Competence, Value, Difficulty, Interest, and Effort. Table 3.12 displays the question content and item response frequencies in proportions.

<table>
<thead>
<tr>
<th>Items</th>
<th>1=SD</th>
<th>2=D</th>
<th>3=N</th>
<th>4=A</th>
<th>5=SA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. I now see the value of learning statistics</td>
<td>0.03</td>
<td>0.04</td>
<td>0.16</td>
<td>0.54</td>
<td>0.24</td>
</tr>
<tr>
<td>2. My attitude about statistics became more positive</td>
<td>0.05</td>
<td>0.12</td>
<td>0.23</td>
<td>0.46</td>
<td>0.14</td>
</tr>
<tr>
<td>3. My anxiety about statistics has reduced so far</td>
<td>0.06</td>
<td>0.15</td>
<td>0.42</td>
<td>0.33</td>
<td>0.04</td>
</tr>
<tr>
<td>4. I am enthusiastic about using statistics</td>
<td>0.06</td>
<td>0.06</td>
<td>0.40</td>
<td>0.38</td>
<td>0.10</td>
</tr>
<tr>
<td>5. I plan to spend more time studying for this course</td>
<td>0.03</td>
<td>0.09</td>
<td>0.39</td>
<td>0.39</td>
<td>0.11</td>
</tr>
<tr>
<td>6. The course has improved my understanding</td>
<td>0.01</td>
<td>0.07</td>
<td>0.14</td>
<td>0.62</td>
<td>0.15</td>
</tr>
<tr>
<td>7. I feel that if I work hard I can do well in this course</td>
<td>0.01</td>
<td>0.05</td>
<td>0.10</td>
<td>0.50</td>
<td>0.34</td>
</tr>
<tr>
<td>8. This course helps prepare me for research involving data analysis</td>
<td>0.04</td>
<td>0.07</td>
<td>0.11</td>
<td>0.51</td>
<td>0.26</td>
</tr>
<tr>
<td>9. I like statistics more than before</td>
<td>0.09</td>
<td>0.15</td>
<td>0.33</td>
<td>0.33</td>
<td>0.11</td>
</tr>
</tbody>
</table>

Table 3.12 Proportions for each level of response. The sample includes 9 items, and 80 sample units with no missing values

Addressing the research questions mentioned before, we are interested in addressing two
main issues. The first issue is “Was there an improvement in attitude over a period of six weeks?” The second issue is: If such a change occurred, what factors influenced the change? Did any of the strategies we employed such as the teaching method with a visualization focus, the immersion of technology, the use of Gapminder and JMP, the hands-on technique with real-life data, and the online learning environment MyLabPlus, play a major role?”

Analysis of results obtained through application of Item Response Theory is employed to assess the latent construct attitude via the respondents’ pattern of responses. First, an overview of the data is obtained by looking at some descriptive statistics.

**Descriptive Statistics**

The items in this dataset contained responses in every category. As can be seen in Table 3.12, few respondents in this sample endorsed the response category associated with strongly disagree for any item. Over 75% of the respondents agreed or strongly agreed with items 1, 6, 7, and 8. They asserted that they now saw the value of learning statistics (item 1) as the course has improved their understanding of statistics concepts (item 6) and helped prepare them for research involving data analysis (item 8). Most importantly, they strongly felt competent about their ability to do well if they worked hard (item 7). Over half of respondents stated that their attitudes about statistics were more positive than before (item 2) and they were enthusiastic about using statistics (item 4). Perhaps experiencing the course work and perhaps the positive feeling motivated them to want to spend more time studying for the course (item 5). While about 37% reported to have reduced their anxiety (item 3), the majority (43%) stayed neutral on this issue and roughly 20% did not agree. This phenomenon was not unique to this group because in general the anxiety level is high in a statistics course. The important point to recognize is the positive feeling they had related to their ability to learn, the willingness to engage in course activities, the interest in learning, the recognition of the value and usefulness of the subject, and the desire to work hard and apply the knowledge learned.

Additional descriptive statistics are provided in Table 3.13, specifically, they are the item-total correlation values, Cronbach’s alpha values without item, \( \alpha_{\text{without}} \), (which is the index of item reliability when being withdrawn from the instrument set), means, median scores, and standard deviations. As can be seen, except for item 5, all other items have satisfactory prop-
erties with item correlation greater than .60 and \( \alpha_{\text{without}} \geq .89 \). The overall \( \alpha \) of the set is 0.91, which denotes a high level of reliability and that the instrument is composed of appropriate items. Furthermore, the mean scores are fairly high (out of 5), all are above the neutral score of 3, with maximum of 4.1. The medians are closely equivalent to the means, indicating that the data do not have a ceiling or floor effect, i.e., no concentration of extreme values appeared. However, the high standard deviations reveal that the scores are not evenly distributed throughout the five categories. This means that some categories have higher frequencies or the responses are more frequent in some categories.

<table>
<thead>
<tr>
<th>Item</th>
<th>( \alpha )</th>
<th>mean</th>
<th>median</th>
<th>sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total w/o</td>
<td>0.68</td>
<td>0.90</td>
<td>3.92</td>
<td>4.00</td>
</tr>
<tr>
<td>1. I now see the value of learning statistics</td>
<td>0.83</td>
<td>0.89</td>
<td>3.51</td>
<td>4.00</td>
</tr>
<tr>
<td>2. My attitude about statistics became more positive</td>
<td>0.66</td>
<td>0.90</td>
<td>3.12</td>
<td>3.00</td>
</tr>
<tr>
<td>3. My anxiety about statistics has reduced so far</td>
<td>0.70</td>
<td>0.90</td>
<td>3.39</td>
<td>3.00</td>
</tr>
<tr>
<td>4. I am enthusiastic about using statistics</td>
<td>0.45</td>
<td>0.92</td>
<td>3.48</td>
<td>3.50</td>
</tr>
<tr>
<td>5. I plan to spend more time studying for this course</td>
<td>0.77</td>
<td>0.89</td>
<td>3.83</td>
<td>4.00</td>
</tr>
<tr>
<td>6. The course has improved my understanding of statistics</td>
<td>0.61</td>
<td>0.90</td>
<td>4.10</td>
<td>4.00</td>
</tr>
<tr>
<td>7. I feel that if I work hard I can do well in this course</td>
<td>0.69</td>
<td>0.90</td>
<td>3.89</td>
<td>4.00</td>
</tr>
<tr>
<td>8. This course helps prepare me for research</td>
<td>0.82</td>
<td>0.89</td>
<td>3.23</td>
<td>3.00</td>
</tr>
</tbody>
</table>

Table 3.13  Descriptive statistics: item-total correlations, Cronbach’s alpha values, mean, median, and standard deviation scores.

The descriptive statistics above, although useful in providing a quick summary of the response data, are aggregate information and do not provide detail about the individual response patterns. Assessing the response patterns can enhance the ability to identify or classify sub-
jects based on their attitudes and gain more information about the intervention. Therefore, we elected to utilize Item Response Theory to examine the response patterns and the probability of each response score.

For parametric IRT analysis, we first had to check the assumptions of unidimensionality and local independence. The acceptable item-total correlation scores in Table 3.13 suggests that a common factor adequately represents these items. The scree plot, the extracted eigenvalues, and the magnitude of the factor loadings displayed in Figure 3.14 were also strongly suggestive of a single factor. More rigorous tests of dimensionality were not practical due to the short instrument length and small sample. Thus, the 9-item set was considered sufficiently unidimensional for IRT analysis. We are now ready to proceed with the analysis.

![Scree Plot](image)

**Figure 3.14** Scree plot, eigenvalues, and factor loadings suggest that this set of items is sufficiently unidimensional.

**IRT analysis for question set 2**
Estimation of the items was performed under the reduced and full graded response models (common slope and unique slopes for all items, respectively). The likelihood ratio test rejected the null hypothesis of a reduced model \((p-value < 0.01)\), indicating that the model including all nine unique slope parameters is a better fit. Results for model goodness of fit from the Kullback-Leibler discrepancy procedure indicated that there was no evidence against the hypothesis of the fit of the full model, or that the model does in fact describe the data well \((p-value = .4213)\).

Table 3.14 shows the estimates of the item category parameters.

<table>
<thead>
<tr>
<th>Item Description</th>
<th>D</th>
<th>N</th>
<th>A</th>
<th>S.A</th>
<th>Slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. I now see the value of learning statistics</td>
<td>-3.03</td>
<td>-1.88</td>
<td>-0.78</td>
<td>1.00</td>
<td>2.24</td>
</tr>
<tr>
<td>2. My attitude about statistics become more positive</td>
<td>-1.80</td>
<td>-0.88</td>
<td>-0.06</td>
<td>1.20</td>
<td>4.06</td>
</tr>
<tr>
<td>3. My anxiety about statistics has reduced so far</td>
<td>-2.07</td>
<td>-0.87</td>
<td>0.62</td>
<td>2.58</td>
<td>1.88</td>
</tr>
<tr>
<td>4. I am enthusiastic about using statistics</td>
<td>-1.90</td>
<td>-1.34</td>
<td>0.22</td>
<td>1.59</td>
<td>2.52</td>
</tr>
<tr>
<td>5. I plan to spend more time studying for this course</td>
<td>-3.98</td>
<td>-2.30</td>
<td>0.12</td>
<td>2.38</td>
<td>1.04</td>
</tr>
<tr>
<td>6. The course has improved my understanding of statistics</td>
<td>-2.81</td>
<td>-1.52</td>
<td>-0.76</td>
<td>1.35</td>
<td>2.57</td>
</tr>
<tr>
<td>7. I feel that if I work hard I can do well in this course</td>
<td>-3.19</td>
<td>-2.08</td>
<td>-1.23</td>
<td>0.73</td>
<td>1.88</td>
</tr>
<tr>
<td>8. This course helps preparing me for my research</td>
<td>-2.50</td>
<td>-1.49</td>
<td>-0.84</td>
<td>0.92</td>
<td>1.90</td>
</tr>
<tr>
<td>9. I like statistics more than before</td>
<td>-1.49</td>
<td>-0.58</td>
<td>0.38</td>
<td>1.39</td>
<td>3.90</td>
</tr>
</tbody>
</table>

Table 3.14 Table of distribution of the category threshold levels.

The results suggest that despite strongly expressing their feeling in both pre- and post-surveys that “Statistics is a complicated subject” \((\approx 60\% \text{ of responses})\) and “It requires a great deal of skills and discipline” \((\approx 70\% \text{ of responses})\), most respondents felt more confident about their ability to learn statistics. This was revealed in their overwhelming \((84\%)\) agreement with the statement “I feel that if I work hard I can do well in this course”. Translating to the attitude
scores, 84\% of support is approximately equivalent with a location of $\theta > -1.2$. This rise in confidence hopefully would equate to a reduction in fear and anxiety about learning statistics. For more than half of the students, these questions were considered very easy to be endorsed with *agree* or *strongly agree* as all the item locations are below zero. As seen in Table 3.15, listed in order of decreasing easiness are items 7, 1, 8, 5, 6, 2, 4, 9, and 3. According to Pinrich and Schunk (2002), students who believe they can do well expend more effort and tend to be more satisfied. The order of the easiest four items (7, 1, 8, and 5) supports this finding. The items convey that students believe that “if they worked hard they can do well in the course” (item 7), they “now see the value of learning statistics” (item 1), and appreciate that “the course helps prepare them for doing research” (item 8). Consequently, they “plan to spend more time studying for this course”.

<table>
<thead>
<tr>
<th>Item</th>
<th>7</th>
<th>1</th>
<th>8</th>
<th>5</th>
<th>6</th>
<th>2</th>
<th>4</th>
<th>9</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td>−1.44</td>
<td>−1.17</td>
<td>−0.98</td>
<td>−0.94</td>
<td>−0.93</td>
<td>−0.38</td>
<td>−0.36</td>
<td>−0.07</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Table 3.15  Item locations = item easiness = item difficulty. The items are listed in order of decreasing easiness. The smaller the value, the easier the item to be endorsed with a score of *agree* or *strongly agree*.

As can be seen in Figure 3.15, the items that provided the most information on the estimates of attitude $\theta$ are items 2 and 9, as they have the largest values of discriminating powers. However, except for item 5, the other six items contribute more information at the range where items 2 and 9 are lowest.
Figure 3.15  Plot of information curves for items in question set 2.

The latent attitude scores span an area from $-2.85 \leq \theta \leq 2.41$ (Figure 3.16), covering the entire positive side of the items’ locations.

Figure 3.16  Distribution and summary of attitude scores for question set 2.
Three-quarters of the $\theta$ scores (above the first quartile $Q_1 = -0.367$) are above all nine of the items’ neutral locations, indicating that for these respondents the scoring options agree or strongly agree are the most probable. This explains that there was a positive change in attitude across almost all components of the attitude construct. Affection for the subject was high (items 9 and 2) – My attitude about statistics became more positive, and I like statistics more than before. Value and appreciation increased (item 1) ($Q_1 = -0.367$ is larger than $b_{Agree,item1} = -0.779$ and $b_{Agree,item8} = -0.84$) – I now see the value of learning statistics. Attitudes about respondents’ own cognitive competence (items 6, 8, and 7) in statistics increased ($Q_1$ is larger than $b_{Agree,item6} = -0.76$ and $b_{Agree,item7} = -1.23$) – The course has improved my understanding of statistical concepts; helped prepare me for research; and I felt I could do well in this course. Individual interest level and effort (items 4 and 5) increased, as half of the students stand above the Agree location of item 4 ($Q_2 = 0.246 > b_{Agree,item3} (0.617)$ – I am enthusiastic about using statistics and I plan to spend more time studying for the course. Consistent with feelings about the difficulty level of statistics, more than 70% of responses to item 3 (My anxiety about statistics was reduced so far) were below the Agree location, ($b_{Agree,item3} = 0.617 < Q_3 = 0.68$).

Returning to the first research question, we concluded that there was a substantial change in attitude over a period of six weeks. The results of the detailed analysis above provide evidence to this claim. They suggest that the students’ attitudes were shifted in a positive direction, i.e., being more positive, feeling more competent about self-ability, and increasing in interest, appreciation, and values for the course. These effects are associated with increased effort invested in learning the course. An observed growth in the willingness to work hard also stemmed from the unwavering recognition of the difficulty level and the demanding requirements of course work. As we have noticed, students’ anxiety level remained high (item 3) and for most students the biggest challenge at the six-week juncture of the course remained Time (as shown in the analysis of question 10 in the previous section). Their biggest concern was having time to study and being able to balance time to devote to reading, viewing lectures, and practicing the course material.

This is evidence of growth and engagement in course work for two reasons. As we have observed, there was no single mention of a lack of resources or challenging material. The abundance
of resources such as lecture videos, online textbook, instant feedback, computer-adaptive-test practice exercise, and self-study plans, provided from the web-based tool MyLabsPlus became catalysts for independent learning and performance. Thus the most critical component to make learning happen remained time. This was the first evidence of growth and engagement. Second was the students’ level of confidence in their own ability to learn the subject. Literature has shown that learners in general are willing to work harder when they know they can succeed and that positive experience motivates learning and performance.

While the numerical data analysis provides us with critical information to observe the positive change in students’ attitudes about statistics, to answer the second research question we would still need to obtain more information from the open-ended questions. The text responses will supplement the answers on “What factors influence the change? Did any of the strategies we employed such as web-based learning tool MyLabsPlus, visualization focus using Gapminder and JMP, real-life data, immersion in technology, play a major role?” Specifically, what factors did they feel were most helpful in learning statistics, what resources did they consider most effective at scaffolding their understanding as well as stimulating their learning and desire for applying what they learned. These questions are answered in the next section using results from the qualitative analysis of some open-ended questions.

3.4.2 Qualitative Data Analysis of Open-Ended Questions

Qualitative analysis methods were applied by examining the contents of the open-ended questions. Seven open-ended and multiple-choice questions were designed to assess students’ attitudinal improvement. This procedure allowed us to elaborate and supplement the numerical responses, putting a human voice and feelings to the “dry” numbers of responses.

The analysis of open-ended questions was conducted by looking at the common themes that emerged from the text responses, compared and contrasted between pre- and post-survey responses (if the questions were included in both surveys), and interpreted the meanings of the themes.
Question 10

Pre-survey: What challenges do you anticipate in this course?

Post-survey: What challenges did you face at this juncture of the course. Have your expectation change over the course of the first six week?

This question was included in both surveys to assess student anticipations of the course challenges. The content was modified in the post survey intended to assess student evaluation of the existing challenges they were facing. Table 3.16 presents the question contents and a summary of the responses for both pre- and post-surveys.

<table>
<thead>
<tr>
<th>Question 10</th>
<th>Common Themes</th>
</tr>
</thead>
</table>
| Pre: What challenges do you anticipate in this course? | -Time  
-Online course  
-Learning to use data analysis software  
-Level of difficult  
-Statistics  
-Math skill |
| Post: What challenges did you face at this juncture of the course? Have your expectation change over the course of the first six week? | -Time, keeping up with the course  
-Variety of technologies used  
-Course organization  
-Gapminder project  
-Group work  
-Difficulty level of exam  
-Demanding of assignments  
-Application of material learned  
-No change in expectation |

Table 3.16  Content and common themes of question 10 in pre- and post-survey.

Students’ comments were very upfront related to this question. The summary of the responses are displayed in Table 3.16. The common themes that emerged from the responses in
the pre-survey were Time, Taking the course as an online student, Worrying about self ability to learn statistics due to low math skill, and Apprehension about learning to use data analysis software.

Compared to the pre-survey, some challenges remained the same in the post survey. Time was the barrier many quoted to be their biggest challenge. Sufficient time to learn the material, proper use of software, and engaging in group work were listed by many as factors that impeded learning. Some felt that time spent on applying what they were learning was in short supply and that the course lacked origination, connection, and conceptual linking. However, many also stated that their expectation remained unchanged at this juncture in the course. This could be considered both negative and positive, as it includes both optimistic and pessimistic expectations. On the definitive positive side, the concerns about taking the course online were on many students’ minds at the beginning of the course, but that did not surface in the post-survey. This could be considered a positive point, as students appeared to enjoy having the freedom of learning independently, reviewing lectures anytime, within the comfort of their environment.

Question 11

Pre-survey: What grade do you expect to get in this course?

Post-survey: Did your grade so far match your expectation? Is it higher or lower than your expectation?

Students’ enthusiasm and level of optimism appeared high. Figure 3.17 displayed the comparison of score expectation between on-campus and on-line students. For the pre-survey, the majority of students expected to receive an “A”, more so for on-campus students (86%) than online students (57%). The responses to the post-survey suggest that the majority of students were at this point satisfied with their performance (grade). Of the 80 responses in the post-survey, only 14% received a lower grade than expected, 23% received a higher grade, and the remaining 63% received their expected grade (matched). We recognized that though grade was not the absolute and sole indication of performance, receiving a good grade boosts self-efficacy and the high expectation for a good grade was a sign of willingness to work hard to earn it.
Thus, the post-survey results painted an encouraging picture of the learning outcomes thus far.

<table>
<thead>
<tr>
<th>Pre: <em>What grade do you expect to receive in this course?</em></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ON-CAMPUS STUDENTS</strong> <strong>VS</strong> <strong>DISTANCE STUDENTS</strong></td>
</tr>
<tr>
<td><strong>GRADES</strong></td>
</tr>
<tr>
<td><strong>A</strong></td>
</tr>
<tr>
<td><strong>B</strong></td>
</tr>
<tr>
<td><strong>C</strong></td>
</tr>
<tr>
<td><strong>Other</strong></td>
</tr>
</tbody>
</table>

More on-campus (yellow) expected an A grade than on-line students (red). Over 90% expected a B or better.

Table 3.17 Contents and summary of question 11 in the pre- and post-survey.

**Question 15**

*List two or three specific items about this course that most stimulated your study of statistics.*

The responses clearly indicated a varied list of stimuli; the three most frequently mentioned items were: *MyStatLab*, Gapminder, and JMP. Among the other items, but not listed in order of importance, were: live labs, exams, homework, tech assignments, notes, visualization, and textbook examples. Students were positive towards MyStatLab for various reasons: the web-based tool learning environment contains an *online textbook*, assignments that can provide *instant feedback*, the dynamic of the *built-in help features* included in each assignment with which students could get immediate help when doing homework, the *study plan* that can be tailored to meet individual needs, and the dynamic nature of the multimedia library that includes various media formats such as videos, animation applets, powerpoint presentations, summaries, etc. Students’ level of enthusiasm about this web-based learning environment spoke volumes about the quality of the software.
Question 16

*List two or three specific items about this course that most helped your understanding of statistics.*

The responses show that the stimuli listed in question 15 served a dual role as items that most helped students to understand statistics. Among the repeats are *MyStatLab*, visual learning, Gapminder, and JMP, plus a new group that includes instant homework, feedback, own discretion in learning, the textbook, hands-on assignments, and accessibility of materials and relevant information. Most students identified *JMP* and *lab lectures* as the most helpful items in helping them understand statistics. The word cloud obtained from the text responses is displayed in Figure 3.17. JMP, lab lectures and videos, homework, MyLabsPlus, and Gapminder, and a few others appear dominantly among other words in the figure. This is because of their being mentioned frequently in the responses. The outcomes were very positive, and suggest that the majority of students were pleased with the variety of class activities and technologies used and felt that they were very useful and contributed to learning statistics.

![Figure 3.17](image_url)

*Figure 3.17* The word cloud obtained from responses to question 16. The cloud gives greater prominence to words that appear more frequently in the text. JMP, lab lectures, homework, Gapminder, and MyLabsPlus are a few dominant stimulus items.
Question 17

Is your attitude towards statistics more positive after six weeks in the course?

Approximately 50% of students reported that their attitudes were more positive after six weeks in the course, 32% responded Remained unchanged, and 19% responded No. The elaborated text responses showed that the reasons for the change cover a broad scope. Examples include: “I like the pace”, “It isn’t as difficult as I thought”, “I believe statistics was important, now I am more certain.” We attributed the attitudinal improvement to two significant reasons.

The first reason for students’ increase in positive attitude was that as they gained understanding of statistics or the class, it became easier to understand the concepts taught. This phenomenon motivated them to want to learn more. A few other representative comments were: “Very much. I like to do statistics homework, believe it or not. Reading online textbook gives me more freedom to read anytime anywhere.”; “I’m actually enjoying statistics for the first time!”; “I can get to understand statistics and work on it” ; “I was looking forward to the course, but now I want to learn it even more.”; “It has been a lot of review from my AP Stats course that I took in high school, but now I see a lot more of the meaning behind the WHY of statistics.”; “So far, this class has done a good job of looking at practical applications of statistics.”

The second reason for the positive attitudinal change was attributed to the teaching team and the excellent resources provided in the course. Those who had taken statistics before mentioned that it was easier to understand and learn the material in depth the second time. Knowing the “whys” this second time made it possible to apply statistics and work with real-life applications.

However, a significant number of respondents scored in the neutral zone, attitude unchanged. The primary reasons people were less positive are their not liking math or numbers, feeling lost and overwhelmed with course work, not liking the Gapminder project and group work, and unsure of how to apply what they learned. As indicated in the literature, students who fear math typically perform worse in statistics courses (Adams & Holcomb, 1986; Zeiden, 1991). For those whose attitudes stayed neutral i.e., liking statistics neither more nor less, the primary comments were that they had always been aware of the value and importance of statistics: “I
always know statistics is important and useful and still do feel the same way now”, “I find it useful, as I knew I would. But numbers bore me; they always have.”; “I do not like Statistics that much, I have a Statophobia.”; “I understand it is important but it is hard to get excited about.”

On the other side of the coin, negative comments include: “Stats is just plain not exciting”, “It is becoming more difficult, I don’t feel prepared to perform the required mathematical analysis.” The reasons for some of the negative comments seemed to concentrate on the difficulty encountered and not lack of understanding. The comments on Gapminder were few but very definitive and on the negative side, mostly due to group work and the use of different technologies. Listed below are some other comments from the negative-attitude group.

“This level class of statistics is overkill for the kind of statistics I will actually use in my research, as my research is entirely qualitative.”; “I feel like topics are not discussed in enough depth before moving on.”; “I don’t like math and stats. Basically, numbers make me nervous. However, I see where knowing these facts and software can help me in my scholarly pursuits. So I’m just trying to survive the course.”; “Honestly the information contained in the homework has not helped me personally on Exam 1. There is a disconnect somewhere that I don’t understand. My homework grade is NOT a direct reflection of my exam grade or vice versa.”

**Question 18**

*Which of the items below do you find useful in helping you learn statistics? Mark all that apply.*

Using MyLabPlus and JMP are standouts in the responses to this question. These two items consistently were perceived as valuable, usable, appropriate, and have been ranked high in responses to previous questions. The other items frequently selected are collaborating with instructors and teaching assistants, and working in groups, completing assignments and projects, working with real-world issues and data. Many expressed satisfactory feelings about the textbook, such as: “The book is amazing!”; “The textbook! It is very well written and helpful in gaining background knowledge for the course content.”
Being able to collaborate with teaching assistants and classmates was also a significant factor in helping them learn. A student expressed her/his gratitude for the teaching team as “I sincerely complement all the instructors of this course.” In general, based on what was observed the changes in attitudes toward in statistics were influenced by the degree to which students experienced being able to learn and understand the course materials.

Question 19

*Approximately how much time a day have you spent studying for this course?*

More than 60% of respondents indicated spending 1-2 hours, and about 1/4 spending three or more hours, every day to study for the class. Compared to the pre-survey this was a substantial increase in the amount of time invested in studying each day. This change in time engaging in studying could be attributed to many reasons such as the demanding assignments for the course, or that the course activities had required more independent study, or the increase in interest in learning statistics.

Question 21

*Does being exposed and having used Gapminder increase your appreciation for statistics? Please explain.*

In this question the surprisingly positive comments from respondents related to the software Gapminder included *useful, I liked it, wished I could have spent more time using it, I liked the display of data, different ways of visualizing data, provided me with real world applications,* and more. Nearly 65% of the respondents commented they were satisfied with learning and using Gapminder because Gapminder helped them recognize the value of learning and using statistics, and that the visualization and animated graphs made numbers come alive and stimulate their desire to use data and learn statistics.

The negative side, however, was mainly on the timing of the course, its organization, and the requirement of the Gapminder activity. Many suggested structuring this activity later in the course where they would be able to perform more advanced data analysis and therefore
create more meaningful applications. Also the early introduction of the project with the requirements of learning about and use of different technologies made the course overwhelming at the beginning, time consuming, and therefore less enjoyable. Some of the negative comments included: too much time spent to learn how to use technology, overwhelming, did not aid my learning much, problem with group work. These responses were consistent with answers to other questions.

As course designers, we could learn from these results to redesign or restructure the project at different time in the course. However, Gapminder is meant to offer versatility and the potential to impact the teaching and learning of statistics.

In summary, the analysis of open-ended questions provided more information about students’ attitude improvement, which supplements the results in the numerical analyses of both question sets 1 and 2, and enriches our understanding on this issue. We now recognize that students’ biggest challenge was finding time to study for the course. This result did not surface in the statistical analysis although students expressed that their level of anxiety remained high we were not able to understand the deep reason behind this concern. Moreover, we now have a better understanding of the “why’s” students did or did not like the course, and the “what’s” factors best stimulating and helpful for their learning.

The combination of results from the qualitative and quantitative analyses enable us to gain great insights into our teaching practice. We were able to realize that the variety of technologies used at that point appeared to be an obstacle for some to keep up with the demands of the course. However, overall the majority was satisfied with their progress as their grade met or went beyond their expectation. The three most important software technologies that stimulated the study of statistics were MyStatLab, Gapminder, and JMP. JMP, lab lectures, and homework assignments were the top three items selected as having helped students understand statistics. About half reported having more positive attitudes due to being able to understand statistical concepts, which caused an increase in motivation. Having a good teaching team and abundant resources for independent learning also received high ratings. The negative comments were due largely to the difficulty level of the course activities and assignments. Many expressed satisfaction with the web-based online environment, MyLabsPlus JMP, the textbook, and good
instructions. Most students spent at least one to two hours every day studying for this course. Finally, a large majority responded positively to the impact of Gapminder on helping increase their appreciation for statistics even though some felt overwhelmed by the activity at the time.

3.5 Discussion and Conclusion

The aim of this study was to evaluate the effects of the use of various technologies such as Gapminder, JMP, and MyStatLab in improving students’ learning experience in a graduate-level statistics service course. We designed, developed, and implemented a small research-based activity using Gapminder software. In conjunction, we employed a teaching method that focused on using the web-based learning environment MyStatLab to enhance learning, employing data analysis software JMP for visualization emphasis, and making statistics connected by using real-life and relevant data sets. The results from both quantitative and qualitative analyses demonstrate that there was meaningful positive change in attitude towards statistics. Level of engagement was increased due to the immersion of technology and teaching methods emphasizing in technology and visualization. This teaching method, even though assessed over a short period, influenced students’ attitudes in a positive way, suggesting that long-term use of this method may be worthwhile.

3.5.1 Students’ Change in Attitude

Previous research findings suggest that many students in introductory statistics courses exhibited negative attitudes, anxiety, or fear toward the subject (Peterson, 1991; Gal, Ginsburg, & Schau, 1997; Schau, 2002; Schau, 2004 & Sorge ). Recent research has revealed some optimistic results that today’s students of statistics have a more positive attitude toward statistics (Forbes, 2012; Mills, 2004). Findings from our study are in agreement with recent findings that the majority of students were not afraid of statistics and viewed the subjects as important and useful for advancing their study and research. However, the composition of our students (graduate students and highly-motivated undergraduates) and the fact that the course is a graduate service class may cause the results to be different from what is reported from other research concentrated on undergraduate students.
To what can we attribute this shift in attitude? How has it been discussed in the literature? More information and insights on the issues were obtained from other non-numerical questions in the post-survey. Students’ responses to the open-ended questions suggest that as students got more involved with class work, they discovered for themselves that they were in fact capable of learning and performing well in the course. Consequently, they were stimulated to study and willing to put more effort into handling challenging material and felt satisfied with their accomplishment. For many of them, the appreciation and cognitive level for the subject increased dramatically over a period of six weeks.

It appeared that experiencing the course helped students realize that prior belief in their math anxiety did not hinder their learning of statistics. For the most part students have come to realize that they can learn statistics and there is no barrier to entry at the level of the course. This is in contrast with results from earlier research. Sorge and Schau (2002) pointed out that students’ attitudes are expected to be more positive the more math they have had, or the more STEM-oriented their education majors are. However, this pattern was usually not homogeneous across the six dimensions of attitudes toward statistics. That is, despite perceiving statistics as difficult due to math phobia or anxiety, students can still recognize the value and the usefulness, and show appreciation, for the field of statistics.

In their 1994 paper, Gal and Ginsburg stated that “a student’s responses to items assessing usefulness-of-statistics issues might have little to do with feelings or attitudes towards statistics as a subject; instead, they may only reflect the student’s vocational development or knowledge about requirements or content of certain professions” (p. 5). For our study, because the participants were mainly graduate students or highly motivated undergraduate students, the findings indicated that positive attitudes were high in all dimensions for all students, more so for those who majored in STEM fields.

Schau (2003) identified six components in measuring attitude toward statistics using her popular instrument, SATS (see Appendix C.3). They are Affect, Cognitive Competence, Value, Interest, Difficulty, and Effort. In a recent paper, Schau and Emmioglu (2012) found that, in general, on average students entered undergraduate introductory statistics courses with a neutral attitude on Difficulty level and Affection constructs; a positive attitude on the Cognition
Competence, Value, and Interest; and a very positive attitude on the Effort. Their attitudes after completing the course either stayed about the same (Affect, Cognitive Competence, Difficulty) or decreased for Value, Interest, and Effort components.

Findings from the quantitative analysis of the perfectly-matched questions showed that even though their perception of difficulty level remained the same from the pre- to post-survey, students were more positive about their ability to learn statistics. In the post-survey people's attitude scores fell into three distinct categories: low, neutral, and high attitudes. These categories coincide with three groups of raw response scores: (1) strongly disagree or agree, (2) neutral, and (3) agree or strongly agree. Results displayed in Table 3.18 indicate that the proportion of respondents with low attitude scores stayed the same while the proportion of respondents decreased for the neutral group and increased for the high attitude group. Overall, more students felt more positive about statistics; they especially agreed with the statement “I am interested in using statistics to understand the world.”

<table>
<thead>
<tr>
<th>Attitude</th>
<th>low ($\theta &lt; -1$)</th>
<th>neutral $(-1 &lt; \theta &lt; 0)$</th>
<th>high ($\theta &gt; 0$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-survey</td>
<td>12%</td>
<td>38%</td>
<td>50%</td>
</tr>
<tr>
<td>Post-survey</td>
<td>13%</td>
<td>29%</td>
<td>58%</td>
</tr>
</tbody>
</table>

Table 3.18  Summaries of attitude scores for pre- and post-surveys. Scores were divided into three groups: low, neutral, and high.

Results from the analysis of the second set of survey items, on the other hand, suggest that overall students’ attitude about statistics were more positive. Many reported that they liked statistics more than before although their anxiety level did not decline as much. Their positive attitude about the value of statistics increased. They were more confident about their ability to learn and willing to expend the time. This may be due to the increase in level of understanding. They felt that the course had helped prepare them for data analysis.

For these students, these survey items were rather easy to score high because all the item locations are negative. For items 1, 2, 6, 7, and 8, more than half of the students would endorse a score Agree or higher. This result indicates that students’ attitudes about the course have improved. Table 3.19 lists the items in order of increasing difficulty levels. The items’ location values marked the point where 50% or more respondents scored at or above the Neutral category.
of all these items.

<table>
<thead>
<tr>
<th>Item</th>
<th>Location</th>
<th>at least 50% scored above level</th>
</tr>
</thead>
<tbody>
<tr>
<td>7. I feel that if I work hard I can do well in this course.</td>
<td>-1.44150</td>
<td>Neutral</td>
</tr>
<tr>
<td>1. I now see the value of learning statistics</td>
<td>-1.16975</td>
<td>Neutral</td>
</tr>
<tr>
<td>8. This course helps prepare for my research</td>
<td>-0.97900</td>
<td>Agree</td>
</tr>
<tr>
<td>5. I plan to spend more time to study for this course</td>
<td>-0.94375</td>
<td>Neutral</td>
</tr>
<tr>
<td>6. The course has improved my understanding of statistics</td>
<td>-0.93475</td>
<td>Neutral</td>
</tr>
<tr>
<td>2. My attitude about statistics has become more positive</td>
<td>-0.38375</td>
<td>Neutral</td>
</tr>
<tr>
<td>4. I am enthusiastic about using statistics</td>
<td>-0.35750</td>
<td>Neutral</td>
</tr>
<tr>
<td>9. I like statistics more than before</td>
<td>-0.07625</td>
<td>Neutral</td>
</tr>
<tr>
<td>3. My anxiety about statistics has reduced so far</td>
<td>-0.06200</td>
<td>Neutral</td>
</tr>
</tbody>
</table>

Table 3.19 Item locations listed in order of easiness to difficulty. Locations are the averages of item categories. All item locations are above the neutral level.

Results from the qualitative analysis not only complement the quantitative results but also strengthen the data-driven conclusions. Schau and Emmioglu (2012, p. 1) stated that in statistics education, as in other academic disciplines “Students will use statistics in life and work if they believe it is useful. They will use statistics only if they believe that they can do statistics. They will choose to engage in statistical tasks and courses if they find it interesting. If they believe that statistics is too difficult..., it is likely that they will not even try to understand and use statistics”. We did find these patterns in our study, with graduate students as the majority of respondents with students being enthusiastic about using statistics, and we found that their understanding levels increased. As a result many reported feeling prepared for doing data analysis-related research. As discussed in a recent broadcast at UCLA by the two leading figures in Statistics education, Robert Gould of UCLA and Chris Franklin of the University of Georgia, students in social science majors are in fact the group that tend to perform best in introductory statistics classes even though they perceive themselves as less capable
than those in STEM majors at the beginning of the course. The video can be viewed from http://tinyurl.com/lp5ap5l.

The reason for the observed change in attitude is nothing magical, as many students have commented that they discovered for the first time as a result of their recent and positive experience in the course that they were able to understand statistics. This can be attributed to a few dominant factors. The first (and most common) factor was related to students’ math phobia and thus statistics anxiety. Literature has shown that students’ attitudes toward statistics often reflect their attitudes towards mathematics or the belief in their own ability to do math (Gal & Ginsburg, 1994). Students realized that the course does not require a lot of mathematics, instead, it focuses on visual or graphical analyses, interpretation, and real-life applications. The experience has therefore influenced the change in attitude to a stronger affection for, appreciation of, and interest in statistics. Nowadays statistics is applied in almost every field. Service-level statistics courses should emphasize concrete real-world applications rather than dwelling on theoretical or mathematical concepts that may drive students away from learning and applying statistics.

The second factor that can be attributed to the observed positive attitude change was the availability of teaching materials (lecture and lab videos, lesson plans, practice homework, online textbook, open-discussion forum, etc.). The abundance of resources and the technologies therefore open doors for learning and discovering.

The third factor that may not be as visible but has been discussed in this article so far, is the alternative type of assessment that was provided by the web-based tool that we used. MyLabPlus, provided students with opportunities to receive immediate feedback, and to have three attempts on any question plus additional practice to master concepts. This context may also lead to an increase in understanding of the material being taught. The more they understood, the more they enjoyed the class and looked forward to learning, which resulted in their liking and willingness to invest in learning the subject, as reflected in students’ answers to question 17. Consequently, as students became more satisfied with the course they seemed to be more willing to engage, learn, and perform.

Research has shown that students’ attitudes toward statistics are a significant determinant
of their performance and that a negative attitude may influence their willingness to apply their statistical knowledge (Schau, 2003). Results from this study suggest that teaching statistics concepts with an emphasis on technology and data visualization tools such as Gapminder and JMP and the availability of technological resources enhanced student attitude and motivation in learning.

3.5.2 Limitations

Despite having many positive findings, this study has three inherent limitations. The first limitation has to do with the critique of not having a control group or possibility to replicate the reported findings. Had it been designed with a control group the results could be generalized to similar study settings. However, as it was the one and only statistics course taught in this setting at the time of the study, creating a control group was not possible. We could not ethically section off students into groups to administer “treatment” and “control” to teach with and without technology and visualization focus. However, we compensated for that by designing the post-survey to contain questions that can assess students’ attitude changes. Most other studies (SATS, ATS) utilized the same survey for the pre- and post-period. Consequently, we believed that the results from this study could be used for course design and planning purposes as long as they are in the same setting (graduate-level service course, and online).

The second limitation was that students’ beliefs about the usefulness of statistics might change over time as a result of experience, participation, increasing difficulty, and, the level of abstraction level of new concepts. Therefore, students’ viewpoints on some of the survey issues may fluctuate depending on their own circumstances (being ahead or behind on course material) and classroom events (assignments, exams, topics covered). Consequently, interpretation of attitude changes needs to take into account the class circumstantial objectives.

Third, our study did not include end-of-semester data, the findings may not reflect a full picture of the influence of the teaching approaches. It would be of interest to survey students’ attitudes at the end of the course to have a full picture of the influence of the teaching tools and techniques.

Last, replication of the exact study to other courses may present some challenges unless
similar technologies can be utilized and instructors are familiar with the tools and able to devote the time to planning.

3.5.3 Conclusion

Although previous research indicated that negative affective responses to statistics are common among students enrolled in introductory statistics courses overall, results from recent research suggested that students in graduate statistics service courses today have more positive attitudes toward statistics than negative attitudes. The findings of our study suggest that in general students were not only positive but also had higher appreciation for what statistics has to offer. As a result, they were willing to invest extra hours and effort in learning statistics.

One of the goals of our teaching method was to foster an active learning environment, to help establish and maintain a positive feeling toward statistics via learning experiences. The six-week exposure seems to have a positive impact on students’ attitudes toward learning statistics. They were more engaged in the course and more confident about their ability to learn. Utilizing a combination of Likert-type scales and open-ended questions in both of our surveys, we were able to assess more completely a picture of the change in attitude. Students at this level are more likely than not to view statistics positively despite perceiving it to have a strong mathematics emphasis. Although results of our study may not be relevant to other graduate service courses, they do provide hopeful pictures and evidence of the importance of utilizing technology in teaching statistics. Moreover, they shed light on the importance of using data visualization tools in teaching.


CHAPTER 4. BRINGING DATA TO LIFE INTO AN INTRODUCTORY STATISTICS COURSE WITH GAPMINDER

Dai-Trang Le

This study was published in Teaching Statistics in June 2013.

Abstract

Gapminder is a free and easy to use software for visualizing real world data in multiple dimensions. The simple format of the Cartesian coordinate system is used in a dynamic and interactive way to convey a great deal of information. This tool can be readily used to arouse students' natural curiosity regarding world events, and to increase the motivation to understand statistics.

4.1 Introduction

Instructors in introductory statistics classes often struggle to stimulate students’ interest in the subject. Various software is used to teach students how to analyze and interpret the data with which they are working. However, focusing on numerical and theoretical techniques and methods to analyze data alone is not sufficient. To be successful in teaching statistics in the 21st century, instructors should implement more engaging instructional methods and employ multiple forms of discourse such as verbal, numeric, audio and visual to communicate effectively with their audience, to engage and capture their attention. It is particularly true when the audience consists of young learners “who have come of age in a technological world where visual information is the norm” (Murphy 2009 p1). Displaying data in ways that students can both enjoy and understand will enhance their learning experience in statistics.
With Gapminder, a free software for visualizing and animating the fact-based real world data in multiple dimensions, instructors can bring rather boring and mundane factual data to life into their classrooms. The interactive interface of Gapminder makes data exploration and visual analysis fun, dynamic, informative, and very intuitive. The software can be used the software to illustrate the relationship among four different variables of interest using the data collected over time. Note the following example in Figure 4.1 illustrates a bubble chart on Breast cancer and Income per person. Each dot represents a country, the size and color of the dot represent the country’s population size and continent.

![Figure 4.1](image)

**Figure 4.1** This bubble chart of Breast cancer versus Income per person portrays the change over time of four variables, breast cancer, income per person, population size, and geographic region.

In this study, I detail some important aspects of how Gapminder can be used to “seize” student attention and ignite their curiosity in statistics at the introductory level. Students in advanced statistics courses can also benefit from playing with the software. I also include examples and illustrations to show how it can be used in other subjects such as geography, history, social studies, etc in grade levels 6-12, to engage students and connect them with current world wide events. My goal is to advocate the use of visual teaching techniques to increase students’ learning and create equal opportunities for all types of learners and perhaps shape their point of view about the complex images of the world. Finally, I present a case study of Gapminder with my students and their video clips of the stories they created using the software.
4.2 Brief History of Gapminder and its co-founder Hans Rosling

As stated in the introduction, this is a free and user-friendly software for visualizing the real-life world data. It was developed by the Gapminder Foundation in Sweden in 2005 and acquired by Google Inc. in 2006. Its co-founder, Hans Rosling, is a Swedish medical doctor, a professor of public health at Karolinska Institute, and a co-founder of Doctors without Borders. He is also a statistician and an inspirational public speaker of TED (Technology Entertainment and Design) talks. In many of his talks, Rosling uses Gapminder to demonstrate the relationships among some key variables of interest. He created animated and interactive charts to illustrate the data on global health and the World Social and Economic Development. His presentations not only capture the audience interests but also resonate and stimulate their actions. Figure 4.2 displays the image of Rosling and the chart in his famous TED talk, “200 years that changed the world”.

Figure 4.2 Image of the five-minute video 200 years that changed the world. Rosling shows how all the countries of the world have developed since 1809. Link to video: [Video link](#).

One of Rosling’s most famous videos is The Joy of Stats, a documentary video broadcast by BBC in 2010. In this video (Figure 4.3) he demonstrates how interesting and fascinating statistics can be when explained in a dynamic and visualizing format.
Benefits of Using Gapminder

As you may already have experienced, data in spreadsheets are meaningless to most people. With Gapminder one can create simple and intuitive motion charts. The interactive nature of the graphs allows users’ curiosity to lead them to more discoveries about the data. The graphs are very dynamic and yet easy to understand and create. In about five minutes of exploring the software, a novice can generate an eye-catching and powerful chart showing the relationship among multiple variables of interest. Using the software’s tracking capability, users can play the data forward, or retract them to view progression over time.

Data Documentation and Data Quality

Most of the data sets in Gapminder are compiled by the United Nations data bank. The documentation page, Data in Gapminder World describes the methods, sources and data used to produce its various datasets. Thus, one can assess the quality of the data and understand the sources of variation or biases. One can also find the most updated information for each data set using their data blog. When reliable data were not available for some countries or territories, rough estimates or simple guesses were used. Therefore the Gapminder team “discourages the use of their data sets for statistical analysis and advises those who require more exact data to
investigate the available data more carefully.” (Total Population Documentation)

**Gapminder versus Google API**

Given its many advantages, Gapminder has its limitations. It is not a software for data analysis such as SPSS, JMP, Excel, etc. It is used primarily to explore trends and relationships among variables supplied internally. In addition, one can not use their own data because uploading external data is not an option. With Google Visualization API, users can access internal data as well as create or import their own data to create bubble charts similar to Gapminder (as long as the data meet the API’s standard Data Table format). Figure 4.4 displays the chart created with Google Motion Chart. A quick guide to the Google Motion Chart can be found here: [http://www.gapminder.org/upload-data/motion-chart/](http://www.gapminder.org/upload-data/motion-chart/).

![Figure 4.4 Chart created using Google Bubble Chart.](image)

**Creating a Screencast of a Gapminder Motion Chart**

One can use Jing, a free screen capture software, together with Gapminder graph to create a screencast that captures the motion chart while narrating the story unveiled in the data. Using Jing, I created a video (displayed in Figure 4.2) on the relationship of the Number of Internet
Users versus the Level of Income per person in countries across the world. Notice there are five different variables plotted on this chart. Two variables, *Number of Internet Users* and *GDP per person* on the vertical and horizontal axes are used to create the bubble scatter graph. The size of the bubbles represents the population size, and the colors represent the geographic regions.

![Figure 4.5 Plot of Number of Internet Users versus Income Per Person](image)

*Figure 4.5*  Plot of Number of Internet Users versus Income Per Person. Link to video: Global Internet Usage versus Income Per Person.

**Visual Literacy, Statistical Literacy and Equity in Learning**

**Visual Literacy**

According to Stuart Murphy, an expert on visual learning, “Research has shown that visual learning theory is especially appropriate to the attainment of mathematic skills” (Murphy, 2009). This is due to the fact “Everywhere you look information is being transmitted visually” (Murphy, 2009). I believe that visual learning can make a profound difference in all other subjects too. This software facilitates visual learning while allowing for a clever and creative way of exploring data in multiple dimensions. In an introductory statistics class, the goals are often to acquaint students with methods of data explorations in order to identify trends and detect relationships among the study variables. Gapminder can be a teaching tool to empower instructors in designing dynamic presentations of real life data, energizing students’ natural
curiosity and creating lasting impressions. For visual learners, a reference picture or a motion chart can carry a lot of weight when it comes time for remembering or understanding of abstract concepts. Consequently, instructors can teach students how to read and interpret the visual information contained in the data, stimulate their interest, and expand their learning of world events.

Carmel Diezmann stated that “Teaching visual literacy is not only about effective mathematics (thus statistics) instruction. Providing students with the opportunity to develop visual literacy is a matter of social justice.” (Diezmann 1995) Consequently, schools need to educate all students in visual literacy to enable them participate equitably in society as our culture is becoming more and more visual. Visual literacy is important for all groups of learners, the visuals and non-visuals. Visual literacy is especially important for those English language learners who rely on the visual language to understand abstract concepts as they have not yet acquired enough English vocabulary. Research has shown visual teaching strategies are some of the most effective teaching ways to reach all learners regardless of their preferred learning style. Visuals allow students to see and understand mathematical as well as statistical relationship and retain their knowledge.

**Statistical Literacy**

Introducing Gapminder in statistics and all other subjects for 6-12 grades promotes proficiency in statistical literacy and statistical reasoning. Through exploration of using real data and visual information from the graphs, student’s awareness of data will be raised. This could increase the desire to use and understand statistics, and encourage the practice of scientific methods such as identifying questions, collecting evidence (data), discovering and applying tools to interpret the data and make sense of statistical information. With equal access to data about the world, all students can practice seeing patterns, investigating relationships among different variables, making interpretations based on sets of data, graphical representations, and statistical summaries. In his 1996 paper “The connected family, bridging the digital generation gap” Seymour Papert stated that “For information to become knowledge it must be assessed, interpreted, and put into a context.” He believes that the ability to interpret statistics is un-
derived in current educational curricula. Gapminder and other software for visualizing data can provide unique opportunities to increase the interpretation and understanding of statistics in a context.

In summary, teachers of all subject areas need to bring these visualizing tools into the hands of this young generation of internet users to increase their knowledge in statistical literacy, statistical reasoning and statistical thinking. It is my ultimate goal to teach students to become statistical literate citizens.

Usage Examples and Resources for Teachers

The Gapminder website provides rich resources that can be applied across the curriculum including math, science, social studies, language arts, and even physical education. The data page http://www.gapminder.org/data/ contains close to 500 data sets to explore and incorporate into lesson plans. One can start with Gapminder World to learn about the global trends from Wealth & Health of Nations. This interactive slide show Human Development Trends 2005 (active link) presents many topics related to globalization and developing countries.

Teachers at all levels can utilize the actual data to expand their students’ learning about the world and promote real inquiry and problem solving thinking. For instance, Gapminder for Geography was created to help geography teachers teach about world development. They can investigate issues such as: “USA or China, who emits the most CO2?” or "Is child mortality falling?" (Figure 4.6)

History and social studies teachers can use Global development with Gapminder to discuss topics such as global health, the effects of HIV, population growth and carbon dioxide emissions, or issues related to world socio-economic development.

English teachers can use a similar version of Gapminder called Dollar Street, an educational software. Dollar Street (as seen in Figure 4.7) displays the world as a street, the street number represents the daily income per person in a family. Teachers can ask students to write a paper or engage them in discussion about the world.
An example of using real world data can be seen from the Gapminder 2010 World Map in Figure 4.8. The chart compares all countries and territories in the world by their wealth and health. Students can explore the world from Gapminder World by selecting the “Open graph menu” box at the top of the left sidebar to see sample graphs. They can click on the vertical or horizontal axis titles to create a graph for any data they want. The data are loaded slowly, but the wait is worth the results.

Students can also follow the Football (soccer) World Cup in Gapminder. Teachers can challenge students using updated data to answer questions such as: Is your favorite team still in the tournament? Is there a relationship between population size and football results? Does a large population increase the chance of finding 11 good players? Are rich countries better at football (soccer)? Students can explore the correlation between income and the ranking made by the International Football Association, FIFA. The match-up between the richest country in the Football World Cup, USA and the poorest country, Ghana, was also one of the most watched events in the 2010 World Cup. Below are the two interesting graphs (Figures 4.9 and 4.10) that highlight two special moments during the competition.

a) Mighty Brazil versus Tiny Cub North Korea
b) Rich versus Poor, Wealthy USA versus Impoverished Ghana

In conclusion, Gapminder makes using data to connect to real life events more meaningful and exciting. It enables displaying time series data with easily understandable moving graphics and establishes an intuitive method for understanding of relationships and patterns. It can be used at all levels and almost all subjects to turn data into meaningful knowledge and empowers students through visual thinking and learning. Instructors of statistics, can utilize Gapminder to increase level of proficiency in statistical literacy and promote the use and understanding of the world developmental statistics.

Case study of Using Gapminder in an Introductory Statistics Course

In one of my statistics courses at Iowa State University, my students were asked to design a project using Gapminder. The goals of the project were to get students become acquainted with the software and explore the data sets that interest them while incorporating concepts of correlation, simple linear regression and interpreting scatter plots. Specifically, students could generate an animated graph, interpret and explain their graphs by writing a paper to “tell the story of their graph”. They could choose to present their story orally or upload it in the
class wikispace. This encourages them to share their work and learn from others. Students also needed to use Jing to create a video in which they could do a narrative on their graph while capturing the change over time in the data.

The enthusiasm in the project was high. Students were excited about the aspects of learning new technologies to investigate statistics while being able to relate that to their own experience. A few students created interesting video stories and the majority chose to upload an electronic copy. Their stories investigated a whole range of topics that addressed some world wide issues from current events to on-going situations. Many students looked at the relationship between issues related to women such as education, poverty, age at first marriage, fertility rate and the level of child mortality in the world. Some interesting and less serious topics include “Does Money Make Life Happier?” or “Cell phone usage over time.” Here are a few interesting videos that I was granted permission to share: (1) Does population predict CO2 emission?, (2) "Effects of Flood on Nations’ Economy", (3) Urbanization and its Implication to Agriculture and Food Supply.
Figure 4.9 Number of football players versus FIFA ranking. The highlight of the 2010 World Cup, Brazil versus North Korea. A battling performance of the tiny cub North Koreans against the mighty Brazilians, the lowest ranked team versus the highest ranked team.

Student comments on the activity:

After being given feedback, students were asked to voluntarily fill out a survey. Here are a few examples of students answers on the four questions of the survey:

Question 1. What did you like about this activity?

- I thought it was a fascinating and revealing and worthwhile endeavor.
- I really enjoyed this activity as I was able to work with real hands-on data of my own interest. I liked that we could look at a plethora of variables and had to find two issues/situations (variables) that tended to correspond/correlate with each other.”
- I like Gapminder. It’s very informative and simple to read. It is amazing to see the changes of the world in various aspects in our lives in a tremendous long term. I was surprised about that.

Question 2. What are your suggestions to improve the activity?
Figure 4.10 Graph of countries played in the 2010 World Cup and their income level. More than half of all teams in the World Cup are high or upper middle-income countries (red and green). Among the low-income countries (dark blue), only two manage to qualify: Ghana and North Korea.

- I felt everything was easy to access and was very straight-forward with guidelines on what to do!
- You might show more about how to make full use of Gapminder, so that we could develop more findings through Gapminder.
- I especially like the user-friendly interface. Without extra time and effort, even a novice of image and video capturing software can easily use and follow the instructions. I will definitely use this Jing for providing class assignment instructions or introducing new language learning software programs to my students. It will be also very effective if I use Jing for online course providing content knowledge depending on the course.

**Question 3. Do you promote the use of Gapminder as a tool for viewing and exploring data in a statistic class?**

- Gapminder was a very interesting and valuable tool to learn about issues that face our world and would be a good resource for any statistics class, maybe even other realms of education.
• Yes. The gaining and understanding of relationship among issues, their origins, causes and implications were awesome.

• Yes, because it expands knowledge and increases understanding statistics exponentially.

**Question 4. In which other settings (other classes, school levels, etc) do you think Gapminder can be used to stimulate and expand learning?**

• I already told a friend who is a middle school teacher about Gapminder and he possibly might use it for an upcoming assignment! Gapminder is an easy tool that really helps someone understand correlation and its components.

• I believe that combining Gapminder with teaching would make the knowledge more impressive and persuasive to students. It should be used in Sociology.”

### 4.3 Conclusion

Successfully connecting students in the 21st century to the real world requires teaching strategies that are in line with their daily involvement in technology. Gapminder provides statistics instructors the capacity to do just that. By using real-world data, students are connected with statistics about the World Social and Economic Development. Introducing this easily understandable moving graphics in the classroom allows for interactive analysis and aids comparison among different countries and states, facilitates understanding of relationships and patterns, and stimulates interest in statistics. Finally, one of my goals is to inform teachers of statistics of the potential of Gapminder in engaging and improving student understanding of statistics. However, its use is not limited to statistics, I believe educators in all disciplines can also take advantage of this technology to improve instructional methods in their classroom. I hope that more educators will explore Gapminder in their teaching and spread the word about this innovative tool for presenting data and promoting learning.
References


2. Gapminder. [http://www.gapminder.org](http://www.gapminder.org)

3. Google API. [Link: Motion chart](https://chart.googleapis.com/chart)


CHAPTER 5. SYNTHESIS AND CONCLUSION

5.1 Synthesis and Conclusion

The objective of learning effectiveness research is to investigate the relationship between explanatory and outcome factors (Fox, 2007). In that process, our goal is to increase the reliability in the estimates of students’ proficiency or an underlying latent trait that may require utilizing IRT in analyzing standardized tests for binary or polytomous data. Item response theory (IRT) allows for inclusion of interactions between the respondents and the test items that they are taking. Consequently, simultaneously investigating the properties of the items and the characteristics of respondents is important to obtain more precise estimates.

Although IRT has been used predominantly by behavioral scientists, its use in social and educational research is still sporadic at best (Osteen, 2010). Regardless of the reasons underlying the absence of IRT-based analyses in the social science and education literature, these fields will benefit from researchers becoming more familiar with IRT methods. This dissertation attempts in part to provide a general introduction to IRT, to call attention to the IRT framework, and to help familiarize potential new users with this methodology.

The first study provides a brief overview of IRT and includes an introduction that explains how to use it for assessing latent factor scores. The study illustrates some basic theoretical features of IRT via an application to data from a national survey of school board members. In addition, we also show how IRT may be used for scaling purposes as one can identify items with low or high information to drop or keep in a scale. Educators now have additional examples of the use of IRT for analyzing similar data sets.

The second and third studies present the use of data visualization technology in teaching statistics. Findings suggest that technology can help foster active learning, increase understand-
ing of concepts, boost self-confidence, and promote independent learning. Although previous research indicated that negative affective responses to statistics are common among students who are enrolled in statistics courses, the findings of our study indicate that in general students were not only positive toward the content of the course but also had higher appreciation for what statistics has to offer more broadly. Over the course of a six-week period we observed an increase in student engagement and participation. Students became more interested and seemed more eager to tackle their online homework. Focusing on teaching data analysis through data visualization first encourages exploration and discovery.

Using Gapminder as a tool in visualizing data during the first few weeks of the course provides students the opportunity for engaging in interesting and meaningful investigations using real-life data. The Gapminder activities provides the students opportunity to study statistics in an informal environment and appears to excite them about the subject. Using JMP as data analysis software allows students to explore data analysis on their own. Results from the study revealed an optimistic picture and suggest that students in this class were more positive about statistics than expected based on results from previous research in the field.

While technology has become a useful component of statistics courses, choosing the appropriate software has become a challenge for instructors. Nevertheless, appropriate application of data visualization tools can boost students' understanding and engagement in working with data. This, in turn, will influence students' attitude, appreciation, and willingness to apply what they have learned.

Although the introduction to IRT in this dissertation is brief, these applications can help to convey the potential that IRT has to offer to education measurement. We also hope to encourage readers to explore this methodology further. A more detailed discussion of IRT and helpful resources can be found in several of the publications referenced in these two studies for readers who wish to learn more about applying IRT methods.

5.2 Limitations and Suggested Resolutions

Despite its many benefits this research contains some limitations as indicated in each individual study. We present a summary of the limitations of the three studies while elaborating
on possible resolutions.

**Study 1: Analyzing Attitudes Toward Student Learning: An Application of Unidimensional Item Response Theory**

As mentioned earlier, in this study, we could have conducted multidimensional IRT rather than two separate unidimensional IRTs and compared the results from the two processes. Nevertheless, current analyses allowed us to demonstrate the procedure twice on two different scales, which provided more examples for new IRT adventurers. Additionally, the models could be made richer by including predictor variables to explain factors affecting the rating scores. Multilevel modeling may also be considered at the school district level, as board members were clustered in districts. Finally, rescaling is also an option to increase the reliability of the estimates.

**Study 2: Student Learning Experiences Related to Using Technology in a Statistics Course**

A few areas in the study of visualization technology in the second study could be improved. The first area is the sample size. To remedy this, future research could include other sections of the same statistics course. If this were possible, researchers could also expand the study to include control and treatment groups so that comparisons could be made. This would also require coordination and planning among the participant instructors.

We recommend that attention be given to creating very similar environments, to obtain a meaningful comparison. In our study, we were not able to achieve this because the composition of students in this class and the hybrid environment of online and on-campus were unique and not easy to replicate in other courses. Another factor to be considered is that the other equivalent sections did not use the same software or textbook, and did not have identical content focus.

The second area that could be improved is the survey instruments. After their first use, instruments should be reevaluated for improvements. Moreover, our current pre- and post-surveys do not contain items that allow us to identify individual students. In trying to maintain an anonymous environment and not asking for any form of identification, we lost the opportunity to gain individual comparison, in spite of having both pre and post surveys.
The third area in which to gain additional information from the research is to redesign the Gapminder activities in order to provide students more time to collaborate with their team members. The current activities were packed into a three-week period, which did not allow students time to thoroughly investigate their topics of interest.

**Study 3: Bringing Data to Life into an Introductory Statistics Course with Gapminder**

In this published article, editorial suggestions were to provide samples of activities as additional resources for educators who would like to implement Gapminder in their classroom. Accordingly, we included a website providing information on the Gapminder project and an activity worksheet are provided in Appendix D.

**5.3 Implications for Future Research**

The results provided in these studies could be extended to new studies.

**Study 1:** Based on the possible resolutions provided above, future research can be conducted to evaluate the recommended procedures. We suggest that future studies could include open-ended questions in to gain a fuller understanding of the broad attitudes and perspectives to responses on surveyed issues. In addition, researchers could conduct a review or study to compare performance of public and charter schools. This would provide charter schools’ opponents and supporters more compelling information to make informed decisions. On the technical side, simulation study could be conducted to evaluate the performance of our proposed simulation-approach test statistics \((Kullback-Leibler)\) discrepancy). This would make it possible to utilize the test as a standard tool for performing test on model goodness-of-fit.

**Study 2:** Researchers could design a study on a larger scale and expand it to become a randomized study. They could develop and fine-tune survey instruments to obtain further information on the effect of visualization technologies, specifically Gapminder and JMP

**Study 3:** Data should be collected on the use Gapminder in teaching at a larger scale to evaluate the effect of the software on student learning about data.
BIBLIOGRAPHY


APPENDIX A. UNIDIMENSIONAL IRT MODELS

The most common UIRT models are models for items with two response categories and models for items with multiple and ordinal response categories. We present each type of model in the following sections.

A.1 Models for Items with Two Response Categories

There are three primary dichotomous IRT models, the one-parameter logistic (1PL), two-parameter (2PL), and three-parameter (3PL) models. These models were developed to describe the relationship between the items' characteristics and the respondent's trait. The name reflects the number of item parameters contained in a model. Each type depicts the probability $P(\theta_j)$ of a key response (usually a correct response to a test item) for respondent $j$ with a latent trait level $\theta_j$. In these model, the assumptions underlying IRT and parameter estimations presented before are applied.

A.1.1 One-Parameter Logistic Model (Rasch model)

This is the simplest and most widely used IRT model proposed by Rasch in 1960 and named after him. The model characterizes each item in a single parameter and so the name one-parameter model. This parameter is the item's location on the latent scale. In a testing situation, the model represents the probability of a correct response for an individual $j$ with an ability level $\theta_j$ and an item difficulty parameter $b_i$. It is expressed in the language of probability or the chance of a correct answer as in equation A.1:
\[ P(y_{ij} = 1 | \theta_j, b_i) = \frac{e^{(\theta_j - b_i)}}{1 + e^{(\theta_j - b_i)}} = \frac{1}{1 + e^{-(\theta_j - b_i)}} \]  

(A.1)

and can be shown graphically as in Figure A.1. For simplicity, \( P(y_{ij} = 1 | \theta_j, b_i) \) is sometimes referred to as \( P(\theta_j) \). The graph displays the probabilities of a correct response as a function of a person’s latent trait for three items with different levels of difficulty. This is a monotonically increasing function because the probability \( P(\theta_j) \) increases when \( \theta_j \) gets larger. These curves, often called the item characteristic curves (ICC), are steepest at the points where they cross the .5 probability line. The slope at that point is referred to as the item’s discriminant parameters, \( a_i \). In this model \( a_i = 1 \). The ICC plot can be used to assess the difficulty levels of the items. The curves indicate easy items if they are far to the left and hard items if they are far to the right. The question included in the plot should be asked when viewing the Rasch’s ICC curves.

**What is the spread of the item difficulties?**

![Figure A.1 Plot of Item Characteristic Curves (ICC) for three items in a Rasch model with items’ difficulty levels \( b_i = -1, 0, 1 \). The items spread apart, representing varying level of difficulty.](image)

To verify that the curves are steepest at \( P(y_{ij} = 1 | \theta_j, b_i) = .5 \), we first take the first partial derivative of equation A.1. This gives the slope of the ICCs (equation A.2):

\[
\text{slope}_{ICC} = \frac{\partial P(y_{ij} = 1 | \theta_j, b_i)}{\partial \theta} = P(1 - P) = PQ
\]  

(A.2)

where \( P = P(y_{ij} = 1 | \theta_j, b_i) \) and \( Q = 1 - P \). This expression produces a zero-slope at \( P = 0 \) and \( P = 1 \). Substituting \( P = .5 \) to equation A.2 yields an ICC slope of .25. Moreover, when \( P = .5 \), equation A.1 gives an interesting result, \( \theta_j = b_i \). This means that when an examinee’s
trait level is the same as the item parameter, s/he has a .5 probability of answering that item correctly.

Then, to show the slope is steepest at $\theta_j = b_i$, take the first derivative of the slope (equation A.2) and maximize the resulting equation by setting it equal to zero and solve for $P$.

$$\frac{\partial P(1 - P)}{\partial \theta} = P(1 - P)(1 - 2P)$$ (A.3)

The slope is maximized at $P = 0$, .5, or 1. Out of these possible solutions, only .5 yields a finite solution for $\theta_j$ in equation A.1. Hence the slope of the function is steepest when the value of $\theta_j$ corresponds to a .5 probability of a correct or key response.

Putting it all together, the difficult parameter $b_i$ is the location on the $\theta$-scale corresponding to an equal probability of a correct and incorrect response. It is also the place where the item is most discriminating. A larger value of $b_i$ parameter indicates a more difficult item because it corresponds to a larger requirement of $\theta$ for increasing the probability of a correct response. Thus items that are further to the right are more difficult.

As we can see, the Rasch model is simple and somewhat intuitive in its form. However, it has a key limitation due to the assumption of equal discriminating power for all items. This assumption is not practical in many testing situations because test developers would want to include different discriminating items in their test. For this type of model, items differ only in their difficult levels and the estimate of $\theta_j$ depends only on the number of correct responses. The 2PL model improves on that limitation by permitting items to have varying slopes. The 2PL model is discussed next.

### A.1.2 Two-Parameter Logistic Model (2PL model)

The two-parameter model, proposed by Birnbaum (1968), includes a discriminant parameter. The model is expressed as in equation A.4

$$P_2(y_{ij} = 1|\theta_j, a_i, b_i) = \frac{1}{1 + e^{-a_i(\theta_j - b_i)}}$$ (A.4)

where $a_i$ is a parameter related to the maximum slope of the ICC. One can apply the first partial derivative to model A.4 to find $a_i$ just the same for the Rasch model. This gives (equation A.5)
\[
\text{slope}_{ICC} = \frac{\partial P_2(y_{ij} = 1|\theta_j, a_i, b_i)}{\partial \theta} = a_i P_2(1 - P_2)
\]  
(A.5)

where \( P_2 = P_2(y_{ij} = 1|\theta_j, a_i, b_i) \). The subscript 2 is added to the probability to distinguish between the models and will be more useful in the explanation of the 3PL model. The partial derivative of the slope with respect to \( \theta \) is expressed as (see equation A.6)

\[
\frac{\partial a_i P_2(1 - P_2)}{\partial \theta} = a_i^2 P_2(1 - P_2)(1 - 2P_2)
\]  
(A.6)

Similar to the Rasch model, the only finite solution for \( \theta_j \) is \( P_2 = .5 \), which occurs at \( \theta_j = b_i \). This is also the point at which the slope reaches its maximum. If the slope of the items is fairly flat, the items are less capable of differentiating between the respondents with low and high values of the latent trait. When the slope is steep, the probability of a correct response to an item increases rapidly as \( \theta \) increases. Hence, the higher the discrimination parameter \( a_i \), the steeper the slope, and the better the item differentiates between high and low ability levels. Moreover, items with larger \( a_i \) provide more information about the latent trait and thus would be weighted more heavily when estimating \( \theta \). Baker and Kim (2004) provide a thorough discussion of person parameter estimation for different IRT models.

Figure A.2 displays an example of three ICCs with varying discriminating power \( a_i \) and equal difficulty level \( b_i \). The questions included in the plot will guide your exploration and understanding of the plot. Figure A.3 showcases an example of the ICCs with varying \( a_i \) and \( b_i \). The curves do not need to intersect at the .5 level of probability. Notice that unlike the ICC curves in a Rasch model, the ICC curves in a 2PL model cross one another, thus items may differ not only in their difficulty level but also in their discriminating power. Consequently, the estimate of \( \theta \) is dependent not only on the number of correct responses but also on the items being responded to correctly.

Another common version of the two-parameter IRT model for binary data is the probit model. This version is referred to as the normal ogive model in which the ICC is based on a cumulative normal distribution (Fox, 2010) (see equation A.7).
What is the spread of the item difficulties?
How discriminating is each item?

Figure A.2 Item characteristic curves of the 2PL model corresponding to three different discriminant levels $a_i$ and an equal level of difficulty $b_i = 0$. The items do not spread out due to having the same $b_i$, however, they vary in differentiating between low and high performers.

\[
P_2(y_{ij} = 1|\theta_j, a_i, b_i) = \Phi(a_i (\theta_j - b_i)) = \int_{-\infty}^{z=a_i(\theta_j-b_i)} \phi(z)dz \tag{A.7}
\]

where $\Phi(.)$ and $\phi(.)$ are the cumulative normal distribution function and the normal density function\(^1\) respectively. The probit and logit item parameters differ by a constant scaling factor of 1.7 (ex: $a_{i,\text{logit}} = 1.7 \cdot a_{i,\text{probit}}$) so one can use either model in a two-parameter situation.

Extending the 2PL model to include a third parameter, we obtain the 3PL model, which will be discussed next.

A.1.3 Three-Parameter Logistic Model (3PL model)

This model extends the 2PL model to allow for a guessing parameter. This by far is the most useful model for dichotomous data as it applies the IRT logic in a manner that is closest to a real-life testing situation in which all respondents have a non-zero probability of guessing.

---

\(^1\)Random variable $Z$ is normally distributed with mean $\mu$ and variance $\sigma^2$ with probability density function:

\[
\phi(z; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2} (z - \mu)^2\right).
\]
What is the spread of the item difficulties?
How discriminating is each item?

![Plot of ICC curves with varying discriminating power, \( a_i \), and varying difficulty level \( b_i \). Item 2 is more difficult but item 1 has higher discriminating power.](image)

A test item correctly, especially for low-ability examinees. The mathematical form of the 3PL model is (equation A.8)

\[
P_3(y_{ij} = 1|\theta_j, a_i, b_i, c_i) = c_i + (1 - c_i) \frac{1}{1 + e^{-a_i(\theta_j - b_i)}}
\]  

(A.8)

where \( c_i \) represents the guessing probability. For the ICCs in model A.8, the \( c_i \) is referred to as the non-zero lower asymptote. Figure A.4 depicts two items, one of which has a guessing probability and the other does not. Notice that in this figure, for low-ability respondents the effect of guessing is high as the two ICC curves differ by the guessing value at the lower end of the ability scale. However, the guessing effect is attenuates as the ability level increases.

Applying the same procedure to find location for the maximum slope, we obtain the first (equation A.9) and second derivatives (equation A.10) for equation A.8:

\[
slope_{ICC} = \frac{\partial P_3(y_{ij} = 1|\theta_j, a_i, b_i, c_i)}{\partial \theta} = (1 - c_i)a_iP_2(1 - P_2)
\]  

(A.9)

and

\[
\frac{\partial(1 - c_i)a_iP_2(1 - P_2)}{\partial \theta} = (1 - c_i)a_i^2P_2(1 - P_2)(1 - 2P_2)
\]  

(A.10)
Figure A.4 ICC of an item described by the 3PL model including the guessing parameter \( c_i > 0 \) (blue). The ICC of an item in a 2PL model (red) with \( c_i = 0 \). The items have different difficulty levels since the \( b_i \)'s are different values. The two slopes do not intersect and they are not parallel to each other at \( b_i \), thereby indicating that the slopes may not be very different. For low-ability respondents the effect of guessing is high for the blue item. However, this effect diminishes as the ability level increases. The two curves are almost identical at the higher end of \( \theta \).

Notice that here \( P_2 = P_2(y_{ij} = 1|\theta_j, a_i, b_i) \) as in the 2PL model because \( c_i \) is not involved in the denominator. Thus the maximum slope is achieved at \( P_2 = .5 \), which is equivalent to \( P_3 = c_i + (1 - c_i)*.5 \). This is the point halfway from \( c_i \) to 1 (not shown). Similar to the 1PL and 2PL models, the point at which \( \theta = b_i \) is the location of maximum slope of the ICCs. Substituting \( P_2 = .5 \) into equation A.9 we obtain the maximum value of \( a_i \) at \( (1 - c_i)a_i/4 \).

Because not all items are dichotomous, there are IRT models designed for polytomous data. In the next section we will discuss the model for polytomous items.

A.2 Models for Items with More Than Two Response Categories

Polytomous data come from responses to items with more than two response categories. These include multiple-choice items, open-ended mathematics questions, Likert-type, ordinal items, rating-scale responses, and graded responses to test or survey questions. The type of IRT model used to describe the interaction between respondents and test items is dependent on the nature of the data that have been collected. Several models have been used for polytomous data, such as the partial credit model, the generalized partial credit model, the nominal response
model, and the graded response model. The characteristics of these models are described in van der Linden and Hambleton (1997). We focus on the graded response model (grm) proposed by Samejima (1969) because it will be illustrated in the application.

A.2.1 The graded response model (grm)

The grm model is appropriate for items whose response categories are ordered. It describes the probability of scoring or selecting a score equal to \( k \) or higher. The response options include rating scale or Likert type categories such as strongly disagree, disagree, neutral, agree, and strongly agree. For simplicity, we assume that all items have the same \( K \) number of unique categories.

In the graded response model, a test or a survey item is supposed to have more than two categories that are dependent on one another and the successful accomplishment of one step requires the successful accomplishment of the previous steps (Reckase, 2008). The probability of accomplishing \( k \) or more steps is called the cumulative probability is assumed to increase monotonically with an increase of the hypothetical construct underlying the test, \( \theta \). This probability is typically represented by a normal ogive or a logistic model.

Let \( P(Y_{ijk} = k|\theta_j, b_{ik}, a_i) \) be the probability of receiving a score \( k \). Let \( P(Y_{ijk} \geq k|\theta_j, b_{ik}, a_i) \) be the cumulative probability for \( k \) or more steps (the cumulative probability of scoring in or above category \( k \)). Then \( P(Y_{ijk} = k|\theta_j, b_{ik}, a_i) \) is difference between of two cumulative probabilities, for \( k \) or more steps and for \( k + 1 \) or more steps

\[
P(Y_{ijk} = k|\theta_j, b_{ik}, a_i) = P(Y_{ijk} \geq k|\theta_j, b_{ik}, a_i) - P(Y_{ijk} \geq k + 1|\theta_j, b_{ik}, a_i) \tag{A.11}
\]

where \( k = 1, \ldots, K \) is the \( k \)th steps in \( K \) categories, \( P(Y_{ijk} \geq k|\theta_j, b_{ik}, a_i) \) is the cumulative probability of scoring in or above category \( k \) of item \( i \) given \( \theta_j \) and the item parameters; \( a_i \) as before is the item slope; \( b_{ik} \) is the category boundary or threshold for category \( k \) of item \( i \).

The normal ogive form of the grm model is given by

\[
P(Y_{ijk} = k|\theta_j, b_{ik}, a_i) = \frac{1}{\sqrt{2\pi}} \int_{a_i(\theta_j - b_{ik})}^{a_i(\theta_j - b_{ik+1})} e^{-\frac{t^2}{2}} dt \tag{A.12}
\]

The logistic form of the model is expressed as:
\[ P(Y_{ijk} = k|\theta_j, b_{ik}, a_i) = P(Y_{ijk} \geq k|\theta_j, b_{ik}, a_i) - P(Y_{ijk} \geq k + 1|\theta_j, b_{ik}, a_i) \]
\[ = \frac{1}{1 + e^{-a_i(\theta_i - b_{i,k})}} - \frac{1}{1 + e^{-a_i(\theta_i - b_{i,k+1})}} \]  
(A.13)

The two forms are equivalent when the discriminations of the logistic model are multiplied by a constant of 1.7. In general the logistic model is more popular as it is simpler looking and easier to understand than the normal ogive for many people.

As previously discussed, an IRT model derives the probability of a response for a particular item in a survey or test as a function of the latent trait \( \theta \) and the item parameters. We are interested in the probability of responding in a specific category. In the graded response model, the cumulative probabilities are modeled directly. This is the probability of responding \textit{in or above} a given category. Then the probability of responding \textit{in} a specific category is modeled as the difference between two adjacent cumulative probabilities. Let \( K \) note the number of response categories of item \( i \). For simplicity, we assume that all items have the same \( K \) number of unique categories. Then there are \( K-1 \) thresholds between the response options. The cumulative probabilities have the mathematical representation as in equation A.14

\[ P(Y_{ijk} \geq k + 1|\theta_j, b_{ik}, a_i) = 0 \]

and these cumulative probabilities lead to the \textit{graded response model}, or the probability of a response \( Y_{ijk} = k \) to be

\[ P(Y_{ijk} = k|\theta_j, b_{ik}, a_i) = P(Y_{ijk} \geq k|\theta_j, b_{ik}, a_i) - P(Y_{ijk} \geq k + 1|\theta_j, b_{ik}, a_i) \]
\[ = \frac{1}{1 + e^{-a_i(\theta_i - b_{i,k})}} - \frac{1}{1 + e^{-a_i(\theta_i - b_{i,k+1})}} \]  
(A.15)

where \( k = 1, \ldots, K \), \( P(Y_{ijk} \geq k|\theta_j, b_{ik}, a_i) \) is the cumulative probability of scoring in or above category \( k \) of item \( i \) given \( \theta_j \) and the item parameters. \( a_i \) as before is the item slope; \( b_{ik} \) is
the category boundary or threshold for category $k$ of item $i$. The cumulative functions for the middle categories look very much like the 2PL model, except for multiple $b_{ik}$ parameters.

Thus, equation A.15 is the form of the grm model. The plots of the boundary probabilities, $P(Y_{ijk} \geq k | \theta_j, b_{ik}, a_i)$, and the probabilities of responding at a specific category in an item, $P(Y_{ijk} = k | \theta_j, b_{ik}, a_i)$, are displayed in Figure A.5. They are referred to as the item operating characteristic function (OCC) and the item category characteristics functions (ICC), respectively. The OCC curves are the same as the two parameter logistic model for the dichotomous items. The top curves in the (OCC) specify the probability of a response in the categories above or below the threshold. The bottom curves (ICC) show the probability of each score categories 1, 2, 3, and 4 for a person at a specific $\theta$ level. The OCCs cross the .5 probability at the point equal to the step difficulty (threshold) and their slopes are steepest at that point. Although the two ICC curves (bottom) for the lowest and highest categories (1 and 4) cross the .5 probability line (horizontal dotted blue line) at the item threshold $b_1 = -3.06$, and $b_3 = 1.54$, the curves for the middle categories (2 and 3) do not necessarily correspond to the item thresholds. The peaks of the curves do not have any obvious connection to the $b_{ik}$ parameters. We can identify which categories are less likely to be chosen from the ICC curves.

For polytomous items, the questions about the $b_{ik}$—parameters and the $a_i$—parameters should be: “What is the spread of the category difficulties?” and “How discriminating is each item?” If the $b_{ik}$—parameters of an item are spread out, the item can measure across a wider range of $\theta$. If the locations are close together or span a narrow area, this item may not differentiate well among respondents across the area. Also low discriminating items have very flat ICC curves. An example of item with low $a_i$ is shown in Figure A.6. The $a_i$—parameter as before indicates how steep the slope is, or how rapidly the response probability changes as attitude increases. The $b_{ik}$—parameters are the category thresholds where respondents at that attitude location have a 50% chance of choosing a designated option or higher.

The expected score on a test or item in a grm model, similar to the dichotomous items, is the sum of the products of the probability of an item score and the item score and expressed in
Figure A.5  OCC and ICC curves for an item with four response categories. The OCC curves (top) represent the cumulative probability functions crossing the .5 probability at the step difficulty parameters (threshold) $b_{ik} = -3.06, -0.91, $ and $1.54$ (see the light blue vertical lines). Each of the five ICC curves (bottom) represents the probability for each response category. The two curves for the lowest and highest categories (1 and 4) cross the .5 probability line (horizontal dotted blue line) at the item threshold $b_1 = -3.06$, and $b_3 = 1.54$. However, the curves for the middle categories (2 and 3) do not intersect at the item thresholds.

equation A.16.

$$E_i(\text{Item Score}) = E(y_{ijk} = k|\theta_j, b_{ik}, a_i) = \sum_{k=1}^{K} k P(y_{ijk} = k|\theta_j, b_{ik}, a_i) \quad (A.16)$$

To recap, this section provides an overview of the characteristics of some common IRT models for dichotomous and polytomous data. Along with the item parameters, there are several statistics that describe the function of items and tests that are unique to IRT. The next section presents a summary of other descriptive statistics for items and test or instruments.
What is the spread of the category difficulties?
How discriminating is the item?

![Plot of Operating Characteristics Curves of a polytomous item. The slopes of the curves are fairly flat except for the last curve, indicating a low discriminating power.](image)

Figure A.6  Plot of Operating Characteristics Curves of a polytomous item. The slopes of the curves are fairly flat except for the last curve, indicating a low discriminating power.

A.3 Other Descriptive Statistics for Items and Test/Instrument

Item and test information are very important concepts in IRT. They form the building blocks for more advanced applications such as computer adaptive testing (CAT), or test linking and equating. We present the brief discussion of the expected function, the item and test information functions, and connect them to the standard error of measurement.

A.3.1 Expected Function - The Test Characteristic Curve

The test characteristic curve (TCC) shows the number-correct score (for the dichotomous model) as a function of $\theta$. This is actually the expected mean score conditional on $\theta$. For example, on a test of three items, an examinee with an ability $\theta_j$ would be predicted to have a certain number-correct score ranging from zero to 3. An example is shown in Figure A.7. The ICCs of all the items can be summed together to form the test characteristic curve (TCC) or test response function expressed in equation A.17:

$$E_i(\text{Item Score}) = E(y_{ijk} = k|\theta) = \sum_{k=1}^{K}kP(y_{ijk} = k|\theta_j)$$  \hspace{1cm} (A.17)

where $y_{ijk}$ is the score on category $k$ of item $i$ for person with latent trait $\theta_j$. As $\theta$ gets larger, the expected score is expected to increase as well.
The test characteristic curve is scaled from the number of items to the maximum score on the instrument (see equation A.18).

\[
E(\text{Test Score}) = \sum_{i=1}^{I} \sum_{k=1}^{K} kP(y_{ijk} = k|\theta_j)
\]  
(A.18)

**A.3.2 Item Information and Test Information Functions**

In IRT, we are interested in the sampling variance of the person’s underlying latent trait. To obtain an idea of how well an item and the entire instrument can estimate person trait locations, we examine the item information and test information. The information of an instrument depends on the items used as well as the ability of the subjects. Easy items can tell us very little about subjects in the upper end (ex: \(\theta_{\text{Hawking}}\) versus \(\theta_{\text{Einstein}}\)) but they provide information on people at the lower end (ex: \(\theta_{\text{Johnny}} < \theta_{\text{Mary}} < \theta_{\text{Bob}}\)). Similarly, harder items tell us very little about the lower end, but provide information on the upper end. The left plot in Figure A.8 shows the information provided by each item \(I_i(\theta)\), and the plot on the right displays the information provided by the instrument \(I(\theta)\), as a function of \(\theta\). Let’s answer the questions included in the graph.

First, the test information function plot shows “How much information does the test provide over the ability/trait range?” As can be seen, the instrument provides its maximum information
for estimating $\theta$ around $\theta = -2$ to 1.5 or 2, with a dip in the neighborhood of zero. As we move away from this range, the instrument provides less information for estimating $\theta$. This knowledge about how an instrument will behave in estimating person location permits the design of an instrument with specific estimation properties (Ayala 2009, p32).

Next, we use the plot of item information curves to examine “How each item contributes to the test information?” As we can see, item 3 (green) provides more information for estimating $\theta$ in the range of $-2$ to 2 compared to the other items (1, 2, and 4). Item 3 provides its maximum information in the range between $-2$ and 1.5 as in the test information function. Outside of that range the item will not yield precise estimates and the corresponding standard error will be large. Moreover, items 1, 2, and 4 provide very little information for estimating $\theta$. One can consider removing them from the instrument or just keep item 2 if necessary. This knowledge about how each item will behave in estimating person location permits the refining or rescaling of the instrument (Ayala, 2009, p. 32).

The term “information” is a statistical indicator of the quality of the estimate of a parameter (Rekarse, 2009). The term information is used to describe how certain we feel about the estimate of a person’s location $\theta_j$. To describe our uncertainty about an estimate, we can use the concept of a standard error. Fisher (1925) considered the variance of the estimate as a measure of “intrinsic accuracy”. The formula for the information function can be expressed in different ways and is based on the Fisher information matrix as in equation A.19:

$$I(\theta) = E \left[ \left( \frac{\partial \ln L}{\partial \theta} \right)^2 \right] = -E \left[ \frac{\partial^2 \ln L}{\partial \theta^2} \right] \quad \text{(A.19)}$$

Baker and Kim (2004, p. 76) presented the formula for the item information function developed by Samejima (1969) for a dichotomous item (see equation A.22)

$$I_i(\theta) = -E \left( \frac{\partial^2 \ln P_i(\theta)}{\partial \theta^2} \right) = \frac{\left[ P_i'(\theta) \right]^2}{P_i(\theta)Q_i(\theta)} \quad \text{(A.20)}$$

where $P_i(\theta)$ is the probability of a correct response, $Q_i(\theta)$ is the probability of an incorrect response, and $P_i'(\theta) = \partial P_i(\theta)/\partial \theta$. Let $I_{ik}(\theta)$ denote the amount of information associated with a particular item response category. Then the amount of information contributed by a response
How much information does the test provide over the ability/trait range?

How does each item contribute to the test information?

Figure A.8  Left: Item information curves (IIC) for four items. Item 3 gives the most information ($IIC_3$) about the latent trait compared to items 1, 2, and 3. Right: Test information curve. The curve resembles the shape of the $IIC_3$ since item 3 has the largest amount of information.

category to the item information is the product of the category information and the probability of the responses, $I_{ik}(\theta)P_i(\theta)$. Thus, the total information yielded by the dichotomous item is the weighted sum of the information from each response category (see equation A.21)

$$I_i(\theta) = I_{i1}(\theta)P_1(\theta) + I_{i2}(\theta)P_2(\theta)$$

Generalizing equation A.21 to a $K$-category response item, we have (see equation A.22)

$$I_i(\theta) = \sum_{k=1}^{K} I_{ik}(\theta)P_{ik}(\theta)$$

which shows that the item information is comprised of the information provided by each category $I_{ik}(\theta)$. $I_{ik}(\theta)$ is the information function of an item response category $k$, and defined by Samejima (1969, 1972) as in equation A.23
\[ I_{ik}(\theta) = -\frac{\partial^2 \ln P_{ik}(\theta)}{\partial \theta^2} = -\frac{\partial}{\partial \theta} \left[ \frac{P'_{ik}(\theta)}{P_{ik}(\theta)} \right] = \frac{\left[ P'_{ik}(\theta) \right]^2 - P_{ik}(\theta)P''_{ik}(\theta)}{[P_{ik}(\theta)]^2} \]  
\text{(A.23)}

where \( P_{ik}(\theta) \) is the probability of a response in category \( k \) to item \( i \) defined in equation (A.15), \( P'_{ik}(\theta) \) and \( P''_{ik}(\theta) \) are first and second derivatives of \( P_{ik}(\theta) \).

Finally, the full test information function is the simple sum of the individual item information functions (IIF), \( I_i(\theta) \). The sum of the individual items’s information is called the test information function of an instrument. The left plot shows that the item information curve \( I_3(\theta) \) is highest in the range \((-3, 3)\), and therefore contains the largest amount of information compared to the other items. The test information curve has the shape of the most dominant curve, \( I_3(\theta) \), and can be represented as \( T(\theta) \) or \( I(\theta) \) as

\[ T(\theta) = I(\theta) = \sum_{i=1}^{I} I_i(\theta) \]  
\text{(A.24)}

### A.3.3 Standard Errors of Measurement

IRT has a much larger emphasis on the error of measurement for estimates of each subject’s latent trait rather than the global index of reliability. In IRT reliability is treated as a function of \( \theta \). The \( I_i(\theta) \)s are used to calculate the standard error of measurement by taking the reciprocal of the square root of the test information function as shown equation A.25:

\[ SE(\theta) = \frac{1}{\sqrt{T(\theta)}} \]  
\text{(A.25)}

Figure A.9 shows the relationship between test information and standard error of measurement. It illustrates that where there is more information the standard error is smaller, and consequently, the estimates of attitude are more precise in the range between \(-2\) and \(1.5\).
Figure A.9  Relationship between test information and standard error of the latent trait attitude. The standard error is lowest in the range $-2$ to $1.5$ where the information values are highest.
BIBLIOGRAPHY


APPENDIX B. PARAMETER ESTIMATION FOR IRT MODELS

This section provides a brief explanation of how the parameters in an IRT model are estimated. A thorough discussion on the techniques of parameter estimation can be found in Baker and Kim (2004). Two types of estimation need to be performed: estimating the item parameters, and predicting the individual latent score $\theta$.

Estimation of the item parameters is often done by using the marginal maximum likelihood (MML), or joint maximum likelihood (JML) estimation procedure. In marginal maximum likelihood, the process begins with the assumption of the distribution for $\theta$, usually as a standard normal distribution. After the item parameters are estimated, the person parameters can be estimated through maximum likelihood, or using a Bayesian-like approach via the mode or the mean of the posterior distribution, Modal A Posteriori (MAP) or Expected A Posteriori (EAP). We discuss the estimation of models with dichotomous and polytomous items separately.

B.1 Parameter Estimations for Dichotomous Items

Consider an education testing situation in which $n$ individuals answer $I$ questions or items. For $j = 1, \ldots, n$ and $i = 1, \ldots, I$, let $Y_{ij}$ be random variables associated with the binary response of individual $j$ to item $i$. Let $\Omega_Y$ denote the set of possible values of the $Y_{ij}$ for person $j$, with $\Omega_Y$ are assumed to be identical for each item in the test. Let $\theta_j$ denote the latent trait of ability for individual $j$, and let $\eta_i$ denote a set of parameters that will be used to model item (question) characteristics. Different IRT models arise from different sets of possible responses $\Omega_Y$ and different functional forms assumed to describe the probabilities with which the $Y_{ij}$ assume those values, namely
\[ P(Y_{ij} = y_{ij}|\theta_j, \eta_i) = f(y_{ij}|\theta_j, \eta_i); \ y \in \Omega_Y \]  

(B.1)

Letting \( y_j = (y_{1j}, y_{2j}, \ldots, y_{ij}, \ldots, y_{Ij}) \), with \( i = 1 \ldots I \) and \( j = 1 \ldots n \), be the vector of \( I \) observed binary responses from the \( j^{th} \) subject having an ability \( \theta_j \), the likelihood equation for person \( j \) can be expressed as

\[ L(\theta_j, \eta|y_{ij}) = P(y_{ij}|\theta, \eta) = \prod_{i=1}^{I} P_i^{y_{ij}} [1 - P_i]^{1-y_{ij}} \]  

(B.2)

where \( P_i \) is short for \( P(y_{ij}|\theta_j, \eta) \), \( \eta \) is the vector of item parameters as before. The likelihood function for all persons is

\[ L(\theta, \eta|y_{ij}) = \prod_{j=1}^{n} \prod_{i=1}^{I} P_i^{y_{ij}} [1 - P_i]^{1-y_{ij}} \]  

(B.3)

and the full log-likelihood for \( n \) persons is

\[ l(\theta, \eta|y_{ij}) = \sum_{j=1}^{n} \sum_{i=1}^{I} y_{ij} \log(P_i) + (1 - y_{ij}) \log(1 - P_i) \]  

(B.4)

The method of MML utilizes the marginal distribution of the full log-likelihood of the item parameters obtained by integrating out \( \theta \) in equation (B.4)

**Estimating the item parameters**

The log-likelihood of the *marginal distribution* of the item parameters (shown in equation B.2) given the form of the \( \theta \) distribution \( g(\theta_j) \) for the independent and identically distributed latent traits, can be written as

\[ l(\eta) = \sum_{j=1}^{n} \sum_{i=1}^{I} \log \int [y_{ij} \log(P_i) + (1 - y_{ij}) \log(1 - P_i)] g(\theta_j) d\theta_j \]  

(B.5)

The integral can be approximated using the Gauss-Hermit quadrature rule. The goal is to find the values of the components in \( \eta \) that maximize the integrated likelihood with respect to \( \theta_j \). Due to the local independence assumption, we can work with one item at a time. However, within each item, the parameters are not independent so the maximums must be found simultaneously. We can use prior distributions for the item parameters and apply the Bayesian
approach to estimate them. Assuming some prior distributions for the item parameters, one can multiply the priors by the likelihood and approximate the product by a numerical integration to obtain the posterior distribution. This procedure is called Bayesian Marginal Maximum Likelihood, or MML, with priors. After the item parameters have been estimated once, the distribution for $\theta$ can be updated. Then, repeat the process of estimating item parameters and updating the latent trait distribution until the components of $\eta$ converge. The most updated distribution of $\theta$ is considered as the posterior distribution, which can be used in the next step to estimate the individual $\theta_j$ scores.

The item parameters $\eta_i$ may include three distinct types of parameters: a discrimination parameter $a_i$, a difficulty parameter $b_i$, and a guessing parameter $c_i$. The common IRT model for dichotomous items is

$$P(y_{ij} = 1|\theta_j, a_i, b_i, c_i) = c_i + (1 - c_i) \frac{1}{1 + e^{-a_i(\theta_j - b_i)}} \quad \text{(B.6)}$$

Estimating the individual latent scores $\theta_j$s.

After the item parameters and the $\theta$ distribution have been estimated, the $\theta$ score for each subject can be estimated using ML, Expected a Posteriori (EAP), or Modal a Posteriori (MAP) procedures. Each examinee $j$ has his or her own $\theta$ posterior distribution, $g(\theta|\eta, y_j)$ which can be used to estimate $\theta_j$. The MAP procedure estimates $\theta_j$ by using the mode of the distribution as the maximum value, and thus is referred to as Modal A Posteriori or MAP. On the other hand EAP procedure estimates $\theta_j$ by using the mean of the distribution as the expected value, and thus is called Expected A Posteriori or EAP.

Modal a Posteriori

In MAP, the mode is found by applying Fisher Scoring Method as in equation

$$\hat{\theta}_{t+1} = \hat{\theta}_t - \left[ \frac{\partial^2 l(\theta_j|\eta, y_j)}{\partial \theta_j^2} \right]^{-1} \left[ \frac{\partial l(\theta_j|\eta, y_j)}{\partial \theta_j} \right] \quad \text{(B.7)}$$
where $\frac{\partial l(\theta_j|\eta,y_{ij})}{\partial \theta_j}$ is the ratio of the first derivative of the log-likelihood function of $\theta_j$ and the Hessian matrix, which is the matrix of second derivatives of the log-likelihood function of $\theta_j$. This is an iterative method with $\hat{\theta}_t$ being updated until convergence is achieved.

**Expected a Posteriori**

The EAP method uses the Gauss-Hermit quadrature rule to approximate the mean of the distribution,

$$\hat{\theta} = E[g(\theta|\eta,y_j)]$$  \hspace{1cm} (B.8)

EAP is a non-iterative method and therefore is more stable than MAP. Under the quadrature approximation approach, $g(\theta|\eta,y_j)$ can be approximated by finding the area under the curve of the function via a discrete distribution such as a histogram.

**B.2 Parameter Estimations for Polytomous Items**

**Item’s Parameter Estimation in a grm**

With the likelihood conditional on $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$ given in expression 2.5 and an assumed normal distributional form $g(\theta_j|\psi)$, for the independent and identically distributed latent traits, the marginal log likelihood for item parameters may be written as

$$l(\eta) = \sum_{i=1}^{I} \sum_{j=1}^{n} \log \int f(y_{ij}|\theta_j,\eta_i)g(\theta_j|\psi)d\theta_j$$  \hspace{1cm} (B.9)

Because of the assumption of local independence mentioned previously, maximization of B.9 reduces to maximization of

$$l_i(\eta_i) = \sum_{j=1}^{n} \log \int f(y_{ij}|\theta_j,\eta_k)g(\theta_j|\psi)d\theta_j$$  \hspace{1cm} (B.10)

for one item at a time, where $\eta_i = (a_i, b_{ik})$ the set of item parameters for $i = 1, \ldots, I$ and $k = 1, \ldots, m$ and $\psi$ is the set of hyper-parameters for mean and standard deviation, usually set
at 0 and 1 respectively. The integrals in B.10 can be numerically approximated using a Gauss-Hermite quadrature algorithm. After the item parameters are estimated, they are used to update information on the distribution of $\theta$, then the item parameters are re-estimated. Repeat the procedure until the estimated values stabilize or converge. After the item parameters and the $\theta$ distribution have been estimated, the $\theta$ score for each subject can be estimated.

**Person’s Parameter Estimation in a grm**

After the item parameters are estimated using the Marginal Maximum Likelihood, person parameters can be calculated via the MLE, MAP, or EAP procedure.

**The MLE method**

For $i = 1, \ldots, I; j = 1, \ldots, n$; and $k = 1, \ldots, K$, let $I$ denote the number of items, $n$ the number of persons, and $K_i$ the number of response categories of item $i$. For simplicity, we assume that all items have the same $K$ number of unique categories. Let $Y_{ijk}$ be the response $k$ to item $i$ for person with latent trait $\theta_j$. Let $a_i$ represent the discrimination parameter for item $i$, and $b_{ik}$ be the category boundaries or thresholds for category $k$ of item $i$. There are $K-1$ thresholds, $b_{ik}$s, between the response options. These thresholds are the boundaries between two adjacent cumulative scores, for example, $b_{3}$ is the threshold between a score of 3 or higher and a score of 2 or lower.

Let $Y_j = (y_{1j}, y_{2j}, \ldots, y_{Ij})$ represent the response pattern to $I$ items for person $j$ where

\[
\begin{aligned}
y_{ij} &= 1, \text{ if person } j \text{ selected response } k \text{ for item } i \\
y_{ij} &= 0, \text{ otherwise.}
\end{aligned}
\]

The likelihood function for a given examinee of ability $\theta_j$ is the likelihood of a particular item response pattern (Baker & Kim, 2004)

\[
P(Y_j|\theta_j, b_{ik}, a_i) = \prod_{i=1}^{I} K_i \prod_{k=1}^{K_i} P_{ik}^{y_{ik}}
\]

(B.11)

where $P_{ik} = \left(P_{i,k} - P_{i,k+1}\right)$ and $P_{i,k} = P(Y_{ijk} \geq k|\theta_j, b_{ik}, a_i)$ and
\[ P(Y_{ijk} \geq 1|\theta_j, b_{i1}, a_i) = 1 \]
\[ P(Y_{ijk} \geq 2|\theta_j, b_{i2}, a_i) = \frac{1}{1 + e^{-a_i(\theta_j - b_{i2})}} \]
\[ P(Y_{ijk} \geq 3|\theta_j, b_{i3}, a_i) = \frac{1}{1 + e^{-a_i(\theta_j - b_{i3})}} \]
\[ \vdots \]
\[ P(Y_{ijk} \geq K + 1|\theta_j, b_{ik}, a_i) = 0 \]

and the log-likelihood is

\[ \log P(Y_j|\theta_j, b_{ik}, a_i) = \sum_{i=1}^{I} \sum_{k=1}^{K_i} y_{ik} \log P_{ik} \] (B.13)

The first derivative of (B.13) is

\[ \frac{\partial \log L}{\partial \theta_j} = \sum_{i=1}^{I} \sum_{k=1}^{K_i} y_{ik} \frac{\partial P_{ik}}{\partial \theta_j} \] (B.14)

where \( P_{ik} \) is defined as above. The second derivative matrix is more involved and not listed here. Interested readers can find it on page 219 in Baker and Kim (2004).

The Fisher scoring method is employed to approximate the MLE

\[ \hat{\theta}_{t+1} = \hat{\theta}_t - \left[ \frac{\partial(\hat{\theta}_t|\eta, y_{ij})}{\partial \theta_j} \right] \frac{\partial^2(\hat{\theta}_t|\eta, y_{ij})}{\partial \theta_j^2} \right]_t \] (B.15)

The Empirical Bayesian Methods

The estimates for the \( \theta_j \)s are more accurate if they span the range of the entire item difficulties. These factor scores are not a by-product of the item parameter estimation process but a separate step. An alternative to MML is joint maximum likelihood (JML), in which the individual factor scores \( \theta_j \)s are estimated along with the item parameters. With JML the metric is standardized so that the observed standard deviation of the estimates is one (DeMars, 2010).

After estimating or calibrating the item parameters and the \( \theta \) distribution and perhaps the subjects’ \( \theta_j \)s, we can use them to answer some general questions that IRT is designed to address. DeMars (2010) suggests the list of questions that are useful for understanding IRT results as well as for evaluating an instrument, or test. They include: “What is the spread of item
difficulties or category difficulties? How discriminating is each item? What is the distribution of abilities/traits in this group of respondents? How does the ability distribution compare to the item difficulty distribution? How much information does the test provide over the ability/trait range? How does each item contribute to the test information? For a given population or sample distribution, how reliable are the ability/trait estimates?” Moreover, questions on estimation methods, procedures for checking assumptions, and model fit should also be addressed.


APPENDIX C. SURVEY WEBSITES

Moreover, questions on estimation methods, procedures for checking assumptions and model fit, should also be addressed.

C.1 Pre-Survey On Attitude and Engagement in Statistics for Statistics 401D-XW

- Pre-Survey Link: http://tinyurl.com/aqwb5n8

Full Pre-Survey Form (below)
Questionnaire on Attitude and Engagement in Statistics Course

Goals of this Questionnaire

- To gain an understanding of student attitude toward the course.
- To evaluate students level of engagement in the course.
- To collect baseline data for future comparison.
- To investigate the effect of the use of visualization technology in teaching graduate-level introductory statistics.

Questionnaire Content

1. Demographic and Academic Background Information (12 questions)
2. Questions on Attitude related to Statistics (6 questions, 24 sub-questions)
3. Questions on Engagement related to Statistics (5 questions, 25 sub-questions)
4. Questions on Technology/Software Experience (4 questions)

Thank you for your cooperation in completing this questionnaire. Your answers will be kept confidentially and no individual can be identified.

* Required

I. Demographic and Academic Background Information (12 questions)

1. Are you a Graduate or Undergraduate student? *
   - [ ] Graduate
   - [ ] Undergraduate

2. What degree are you seeking? *
   - [ ] Bachelors
   - [ ] Masters
   - [ ] Doctorate
   - [ ] Other: ____________________
3. **What is your major or academic discipline? Second major (if applicable)?** *

4. **What is your gender?** *
   - Female
   - Male

5. **Which best describes you?** *
   - Hispanic
   - Asian/Pacific Islander
   - White
   - Black or African American
   - Multi-racial
   - Other: [ ]

6. **What is your age group?** *
   - 18-29
   - 30-39
   - 40-49
   - 50-59
   - >60

7. **Are you registered as an on-campus (section D) or off-campus (section X W) student?** *
   - on-campus
   - or off-campus

8. **Are you enrolled full time or part time student?** *
   - Part time
   - Full time
   - Other: [ ]
9. **Why are you taking this course/this specific section?** *

- Required Course.
- Fits my interest.
- Fits my schedule.
- Online course.
- Heard that the professor is good.
- Other: 

10. **What challenges do you anticipate in this course?** *

*Please express your anticipation in a few sentences.*

11. **What grade do you expect to receive in this course?** *

12. **What best describes your learning style?** *

- Visual (reading, viewing images, videos, etc)
- Auditory (listening to lecture, class discussion, etc), or
- Tactile (learning by doing, hands-on activities)
- Other: 

### II. Questions on Attitude toward Statistics (24 questions on 6 aspects)

Response options:

<table>
<thead>
<tr>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Neither Disagree nor Agree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

**Questions on Attitude Toward Statistics** *

1. Strongly
3. Neither
5. Strongly
<table>
<thead>
<tr>
<th>Statement</th>
<th>Disagree</th>
<th>2. Disagree nor Agree</th>
<th>4. Agree</th>
<th>Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. I can learn statistics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. I understand statistical equations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. I find it difficult to understand statistical concepts</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. I have trouble understanding statistics because of how I think</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. I like statistics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. I feel insecure when I have to do statistics problems</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. I am afraid of statistics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. I enjoy taking statistics courses</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. I will not need statistics in my profession</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. Statistics is irrelevant in my life.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. Statistical skills will make me more employable.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. I plan to work hard in my statistics course.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14. Most people have to learn a new way of thinking to do statistics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15. Learning Statistics requires a great deal of skills and discipline</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16. I plan to spend</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### III. Questions on Engagement

1. In a typical 7-day week, about how many hours do you spend reading for your courses? (Q7)
   
   Response options: 0 hours, 1-5 hours, 6-10 hours, 11-15 hours, 16-20 hours, > 20 hours

   - [ ] 0 hours
   - [ ] 1-5 hours
   - [ ] 6-10 hours
   - [ ] 11-15 hours
   - [ ] 16-20 hours
   - [ ] > 20 hours
2. During the current school year, how often have you done the following? *
*Response options: Very often, Often, Sometimes, Never

<table>
<thead>
<tr>
<th>Activity</th>
<th>Very often</th>
<th>Often</th>
<th>Sometimes</th>
<th>Never</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asked questions or contributed to course discussions in other ways</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prepared two or more drafts for a paper or assignment before turning it in</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Come to class without completing readings or assignments</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asked another student to help you understand course material</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Explained course material to one or more students</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prepared for exams by discussing or working through course material with other students</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Worked with other students on course projects or assignments</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gave a course presentation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attended an art exhibit, play or other arts performance (dance, music, etc.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3. Which of the following have you done or do you plan to do before you graduate? *

<table>
<thead>
<tr>
<th>Activity</th>
<th>Done or In progress</th>
<th>Plan to do</th>
<th>Do not plan to do</th>
<th>Have not decided</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participate in an internship, co-op, field experience, student teaching, or clinical placement</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hold a formal leadership role in a student</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Participate in a learning community or some other formal program where groups of students take two or more classes together

Participate in a study abroad program

Work with a faculty member on a research project

Complete a culminating senior experience (capstone course, senior project or thesis, comprehensive exam, portfolio, etc.)

4. During the current school year, how much has your coursework emphasized the following? *

<table>
<thead>
<tr>
<th></th>
<th>Very much</th>
<th>Quite a bit</th>
<th>Some</th>
<th>Very little</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memorizing course material</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Applying facts, theories, or methods to practical problems or new situations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analyzing an idea, experience, or line of reasoning in depth by examining its parts</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluating a point of view, decision, or information source</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forming a new idea or understanding from various pieces of information</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Connected ideas from your courses to your prior experiences and knowledge</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
5. During the current school year, about how often have you done the following? *
Response options: Very often, Often, Sometimes, Never

<table>
<thead>
<tr>
<th>Activity</th>
<th>Very often</th>
<th>Often</th>
<th>Sometimes</th>
<th>Never</th>
</tr>
</thead>
<tbody>
<tr>
<td>Talked about career plans with a faculty member</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Worked with a faculty member on activities other than coursework (committees, student groups, etc.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discussed course topics, ideas, or concepts with a faculty member outside of class</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discussed your academic performance with a faculty member</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

IV. Questions on Visualization Tools

1. Have you ever heard of Professor Hans Rosling and his data visualization tool Gapminder? If yes, please explain in which context/situation you encounter Gapminder. *

State the context in which you encountered Gapminder!

2. What is your favorite educational gadget or software? *

3. In terms of using statistical data analysis software such as JMP, SAS, SPSS, or R
are you a beginner? Intermediate user? Or advanced user? *

4. Given a choice of a point-click menu-driven statistical software or a program-statement-and-command interface software, what would you choose? *

Thank you for your co-operation in completing this Questionnaire.
C.2 Post-Survey (Six-Week) survey on Attitude and Engagement in Statistics for Statistics 401D-XW

- Post-Survey Link: http://tinyurl.com/cxrnax

- Full Post-Survey Form (below)
# Six-Week Questionnaire for Statistics 401 D-XW

## Part I: Demographic and Academic Background Information (12 questions)

**Q37** *Six-Week Questionnaire for Statistics 401 D-XW*

This questionnaire is conducted in an effort to evaluate our teaching effectiveness. Please take time to complete the survey. It consists of four parts, Demographics, Attitude, Engagement, and Technology with 28 questions.

Your participation is greatly appreciated. A 1% extra credit will be awarded when the group response rate reaches 90% by midnight on Thursday 03/07/2013.

---

### Q1. Are you a Graduate or Undergraduate?

- [ ] Graduate
- [ ] Undergraduate
- [ ] Other

### Q2. What degree are you seeking?

- [ ] Bachelors
- [ ] Masters
- [ ] Doctorate
- [ ] Other

### Q3. What is your major or academic discipline? Second major (if applicable)?


### Q4. What is your gender?

- [ ] Male
- [ ] Female

### Q5. Which best describes you?

- [ ] Asian/Pacific Islander
- [ ] Black/ African American
- [ ] Hispanic
- [ ] Multi-racial
- [ ] White
- [ ] Other
Q6. What is your age group?

- 18-29
- 30-39
- 40-49
- 50-59
- 60 or older

Q7. Are you registered as an on-campus (section D) or off-campus (section XW) student?

- On-campus
- Off-campus

Q8. Are you enrolled as a full-time or part-time student?

- Full-time
- Part-time
- Other

Q9. Have you ever taken any statistics courses before?

- Yes. Please explain or list the name of the course.
- No

Q10. What challenges did you face at this juncture of the course? Have your expectations changed over the course of the first six weeks?

Q11. Did your grade so far match your expectation? Is it higher or lower than your expectation?

Q12. What best describes your learning style?

- Visual (reading, viewing images, videos, etc)
- Auditory (listening to lecture, class discussion, etc)
- Tactile (learning by doing, hands-on activity)
- Other

Part II: Questions on Attitude related to Statistics (6 questions)
### Q13. Attitude toward statistics

<table>
<thead>
<tr>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Neutral</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>I now see the value of learning statistics.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>My attitude about statistics become more positive as the course progresses.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>My anxiety about statistics has reduced so far.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I am enthusiastic about using statistics.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I can learn Statistics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Statistics is a complicated subject.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I am interested in learning statistics</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I am interested in using statistics to understand the world and global issues.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Q14. Your feeling about statistics at this juncture of the course.

<table>
<thead>
<tr>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Not agree nor disagree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>After experiencing the course work, I plan to spend more time studying for this course.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The course has improved my understanding of statistical concepts</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I feel that if I work hard I can do well in this course.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>This course helps preparing me for my research involving data analysis.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Statistics should be a required part of my professional training.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Learning Statistics requires a great deal of skills and discipline.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I like statistics more than before.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I like statistics teaching methods that emphasizes data visualization.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Q15. List two or three specific items about this course that most stimulated your study of statistics.

- Item 1
- Item 2
- Item 3

### Q16. List two or three specific items about this course that most helped your understanding of statistics.

- Item 1
- Item 2
- Item 3
Q17. Is your attitude towards statistics more positive after six weeks in the course?
- Yes. Please elaborate.
- No. Please elaborate.
- Remains Unchanged. Please elaborate.

Q18. Which of the items below do you find useful in helping you learning statistics? Mark all that apply.
- Using a data visualization tool like Gapminder
- Using statistical software JMP
- Using MyLabPlus
- Collaborating with other students and/or group members
- Collaborating with instructors and/or TA
- Engaging in class assignments and project
- Working with real world issues and real data
- Investigating topic that interests me using Gapminder and JMP
- Sharing successes and concerns openly in class
- Other. Please Explain.

Part III: Questions on Engagement in Statistics (5 questions)

Q19. Approximately how much time a day have you spent studying for this course?
- 1-2 hours per day
- 3-4 hours per day
- more than 4 hours per day
- Other. Explain.

Q20. Approximately how much time did you spend on the Gapminder Project?
- Checkpoint 1
- Checkpoint 2
- Checkpoint 3

Q21. Does being exposed and having used Gapminder increase your appreciation for statistics? Please explain.
**Q22. My level of engagement in the class has increased due to:**

<table>
<thead>
<tr>
<th></th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Neither Agree Nor Disagree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Using Gapminder to investigate and learn about world issues</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b. Using JMP to do data analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c. Participating in discussion forum, study groups, etc.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d. Doing Homework via MyLabPlus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e. Doing Tech assignments</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>f. Collaborating with group members on Gapminder project and Exam.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g. Viewing weekly lab videos</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>h. Viewing Lecture Videos</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i. Using material in MyLabPlus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Q23. Course Learning Engagement**

<table>
<thead>
<tr>
<th></th>
<th>Strongly disagree</th>
<th>Disagree</th>
<th>Neutral</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. I enjoy working with other students on course project, assignments, and exam.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b. I try to find the intuition behind the ideas presented.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c. I ask “Why” questions often.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d. I often revisit or review the concepts taught in class.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e. I often asked questions and contributed to course discussion.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>f. I explained course material to one or more of my peers.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g. I am motivated to get a good grade in this course.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>h. I want to learn how to do data analysis in JMP.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Part IV: Questions on Technology/Software Experience (5 questions)**

**Q24. List two or three specific technology items we use in this course that are most helpful to you in learning statistics.**

**Q25. Did you like the idea of using web-based technology such as Gapminder in learning statistics? Please elaborate.**

Yes

No
Q26. How proficient do you think you are with the following software? (i.e. being able to do what you want to do with them)

<table>
<thead>
<tr>
<th>Software</th>
<th>So so</th>
<th>Good</th>
<th>Very Good</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gapminder</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JMP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MyLabPlus</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Q27. Do you think using JMP rather than SAS or any programming-based statistical software help you in learning statistics in a better manner?

- Yes. Please elaborate.
- No. Please elaborate.
- Other. Please elaborate.

Q28. How proficient do you think you are with JMP so far? Are you able to use JMP to do your homework and create graphics such as histogram, scatter plot, charts, etc?

Q29. On the scale from 1 to 5, from least (1) to most favorite (5), rate the following components?

<table>
<thead>
<tr>
<th>Component</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gapminder software</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JMP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MyLabPlus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blackboard</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gapminder Wiki</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lecture Videos</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lab Videos</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other, please specify</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

THANK YOU for YOUR PARTICIPATION!
C.3 Survey of Attitude Toward Statistics (SATS)


C.4 National Survey of Student Engagement (NSSE)

http://nsse.iub.edu/html/about.cfm
APPENDIX D. GAPMINDER PROJECT ACTIVITIES

D.1 Gapminder Project Website

- https://sites.google.com/site/gapminderproject/

Purpose: provide educators a mini-research project sample using Gapminder. In this project, students will use data sets from Gapminder to investigate a social topic in which they are interested. Through this activity, students will develop the following skills:

1. Research Process (identifying and developing a topic, gather information, data analysis, etc)
2. Quantitative Reasoning (analyzing charts, patterns, trends, relationships)
3. Writing Research Paper
4. Presenting their Findings