University of Tennessee, Knoxville
TRACE: Tennessee Research and Creative Exchange

Doctoral Dissertations | Graduate School

12-2016

Development and Validation of the Statistics Assessment of Graduate Students

Dammika Lakmal Walpitage
University of Tennessee, Knoxville, dwalpita@vols.utk.edu

Follow this and additional works at: https://trace.tennessee.edu/utk_graddiss

Part of the Applied Statistics Commons, Educational Assessment, Evaluation, and Research Commons, Science and Mathematics Education Commons, and the Social Statistics Commons

Recommended Citation

Walpitage, Dammika Lakmal, "Development and Validation of the Statistics Assessment of Graduate Students. " PhD diss., University of Tennessee, 2016.
https://trace.tennessee.edu/utk_graddiss/4113

This Dissertation is brought to you for free and open access by the Graduate School at TRACE: Tennessee Research and Creative Exchange. It has been accepted for inclusion in Doctoral Dissertations by an authorized administrator of TRACE: Tennessee Research and Creative Exchange. For more information, please contact trace@utk.edu.

To the Graduate Council:

I am submitting herewith a dissertation written by Dammika Lakmal Walpitage entitled "Development and Validation of the Statistics Assessment of Graduate Students." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Educational Psychology and Research.

Gary J. Skolits, Major Professor

We have read this dissertation and recommend its acceptance:

Jennifer A. Morrow, Ralph S. McCallum, Hamparsum Bozdogan

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Development and Validation of the Statistics Assessment of Graduate Students

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Dammika Lakmal Walpitage

December 2016

Copyright © 2016 by Dammika Lakmal Walpitage

All rights reserved.

DEDICATION

To my mother, father, and grandparents, who introduced me to the joy of learning and prepared me with unconditional love to chase my dreams.

To my wife and two daughters, whose endless affection and encouragement made me able to achieve such success and honor.

ACKNOWLEDGEMENTS

I wish to thank my committee members who were more than generous with their expertise and precious time. A special thanks to Dr. Gary Skolits, my advisor and committee chairperson, for his countless hours of advising, constructive comments, encouragement, and patience throughout my entire journey of doctoral studies. Thank you Dr. Jennifer Morrow, Dr. Steve McCallum, and Dr. Hamparsum Bozdogan for providing me with continued support and encouragement.

I also must thank Dr. Melisa Martin, Kelly Smith and Maya Mingo, and members of Psycho-educational studies research group. I specifically thank Dr. Kent Wagoner for his invaluable assistance with item development for this project.

I express my appreciation to all my friends for helping and inspiring me to achieve more. I also extend my heartfelt thanks to my family. Thank you so much for always being there for me.

ABSTRACT

This study developed the Statistics Assessment of Graduate Students (SAGS) instrument, and established its preliminary item characteristics, reliability, and validity evidence. Even though there are limited number of assessments available for measuring different aspects of statistical cognition, these previously available assessments have numerous limitations. The SAGS instrument was developed using Rasch modeling approach to create a new measure of statistical research methodology knowledge of graduate students in education and other behavioral and social sciences. Thirty-five multiple-choice questions were written with stems representing applied research situations and response options distinguishing between appropriate use of various statistical tests or procedures. A focus group meeting with upper level graduate students was held in order to revise the initial instrument. Then, a six-person expert panel reviewed the revised items for content validity and to improve the quality of the instrument. The finalized SAGS instrument with 25 cognitive questions and demographic questionnaire was administered online, and 132 participants fully completed the instrument. Results showed that, one SAGS item was not consistent with the Rasch model. This item and distractors of two other items were flagged to be modified during future administrations. Reliability indices, separation indices, constructs maps, and known group comparisons provided the supportive evidence for reliability and validity. Preliminary simulation study conducted with higher order IRT models rejected three parameter logistic (3 PL) model and indicated no impact of guessing parameter when describing the observed data. A simulation study further provided positive evidence towards using ICOMP type model selection criteria that guard against correlations of parameter estimates when choosing the best model among a portfolio of IRT models. Sample independent parameter estimates obtained using Rasch and IRT approaches in this study open an avenue to develop customizable yet psychometrically sound statistical research methodology assessments.

TABLE CONTENTS

CHAPTER ONE: INTRODUCTION AND GENERAL INFORMATION 1
Problem Statement 3
Purpose of the Study 8
Significance of the Study 10
Limitations of the Study 12
Definitions of Key Terms 12
Organization of the Study 13
CHAPTER TWO: LITERATURE REVIEW 15
Graduate Students and Statistics Education 15
Why Statistics is Taught at the Graduate Level? 15
Importance of Selecting an Appropriate Statistical Procedure/Test 17
Statistics Education and Statistics Education Research 18
Importance of Assessing Cognitive Outcomes 20
Assessing Statistical Constructs 21
Defining the Cognitive Constructs Associated with Statistics Education Research 22
Review of Previous Studies on Developing Statistics Skills Assessment Scales 24
Statistical Reasoning Assessment (SRA) (1998) 24
Quantitative Reasoning Quotient (QRQ) (2003) 26
Instruments in (ARTIST) Project and its Relatives, CAOS (2002) 27
Instruments in (ARTIST) Project and its Relatives, ARTIST Topic Scales (2002) 29
Goals and Outcomes Associated with Learning Statistics (GOALS) (2012) 30
Basic Literacy in Statistics (BLIS) (2014) 31
Statistics Concept Inventory (SCI) (2002) 32
Statistical Literacy Survey (SLS) 33
Development of the SAGS Instrument 34
Brief Review of Classical Test theory (CTT) 41
Item Response Theory (IRT) against Classical Test Theory (CTT) 44
Review Basics of Item Response Theory 45
IRT Parameters 46

Assumptions of IRT 48
Three Common IRT Models for Dichotomous Data 49
Goodness of Fit of IRT Models 52
Reliability in IRT 53
Rasch Modeling as a Subset of IRT: Mathematically 56
Differences in IRT and Rasch Modeling: Philosophically 56
Sample Size Requirements for Rasch Modeling 57
Important Pieces of Rash Analysis 58
Chapter Two Summary 66
CHAPTER THREE: MATERIALS AND METHODS 67
Review of the Problem 67
Study Purpose and Objectives 68
Research Design 70
Study Population and Sample 70
Instrument Development 71
Measures 73
Procedure 74
Data Analysis 76
Assess the Rasch/IRT Assumptions: Unidimensionality and Local Independence 76
Evaluating Rasch Model Fit 77
Estimating Parameters and Information Function of SAGS Items 77
Establishing Reliability and Validity Evidence 78
Examining the Model Fit of the SAGS Items to Higher Order IRT Models 78
Chapter Three Summary 79
CHAPTER FOUR: RESULTS 81
Data Cleaning 81
Initial Analysis and Participant Characteristics 82
Rasch and IRT Modeling 86
Test for Violations of Essential Unidimensionality and Local Independence 86
Evaluating Rasch Model Fit 89
SAGS Rasch Statistics: Item Difficulty Estimates (b) 91

SAGS Rasch Statistics: Person Ability Estimates (θ) 91
Variable/Construct/Wright Map 93
SAGS Overall Test Performance 95
Distractor Analysis 95
Establishing Reliability Evidence 98
Establishing Validity Evidence 102
Construct and Predictive Validity 102
Validity Evidence Using Group Mean Comparisons 103
Convergent Validity 104
Post-Estimation of Rasch Model 106
Evaluation of 2 PL and 3 PL IRT Models and Performance of Information Criteria 107
Comparison of Parameter Estimates: IRT Estimates Vs. CTT Indices 110
Chapter Four Summary 112
CHAPTER FIVE: DISCUSSION 114
Summary of Study Purpose, Objectives, and Method 114
Implementation and Results of SAGS Development 118
SAGS Results – Alignment with Previous Research 124
Item Difficulty Parameters and Most Used and Least Used Statistical Procedures 124
Sources for Content Validity Evidence 126
Sources for Item Construct Validity Evidence 127
Item Construction Elements 128
Performance of Various Model Selection Criteria in Selecting Best IRT Models 128
SAGS Results – Expanding Upon Previous Studies 129
Filling the Measurement Gap 129
Expanded Content Coverage and Improved Item Quality 130
Strengthening the Psychometric Accuracy of Assessing Statistics Constructs 131
Exploring Performance of Novel Model Selection Criteria 131
Practical Implications 132
Practical Implications of SAGS for Graduate Students 132
Practical Implications of SAGS for Statistics Educators 133
Limitations of Present Study 135

Future Research 136
Improving SAGS Items and Expansion 137
Additional Testing for Item Parameter Stability and Validity 137
Methodological Advances with Information Theoretic Model Selection Criteria 138
Final Summary 138
LIST OF REFERENCES 140
APPENDICES 163
Appendix A 164
Appendix B 169
Appendix C 171
Appendix D 182
Appendix E 183
Appendix F 185
Appendix G 187
Appendix H 189
Appendix I 191
Appendix J 193
Appendix K 199
Appendix L 200
Appendix M 201
Appendix N 202
Appendix O 203
VITA 205

LIST OF TABLES

Table 2.1 Currently Available Cognitive Statistics Assessments ……. 35

Table 2.2. Sample Size Requirements for Rasch Modeling …….. 58

Table 4.1. Descriptive Statistics of SAGS Instrument ……... 83

Table 4.2. Frequency distribution of Completion Time ……. 84

Table 4.3. Background Characteristics of Participants and Group Specific Total Scores ……..……..... 85

Table 4.4. Statistics Exposure and Group Specific Total Scores ……... 87

Table 4.5. Classical Test Theory and Rasch Approach Difficulty Parameter Estimates ……..……. 92

Table 4.6. Correlation of Response Options and Persons' Ability Level …….. 97

Table 4.7. Response Option Frequencies and Percentages for the Top and Bottom 25% of Participants …….. 99

Table 4.8. Construct and Predictive Validity through Known Group Comparisons …...…….. 105

Table 4.9. IRT Based Model Evaluation Summary Statistics …… 108

Table 4.10. Table of Item Difficulties and Discriminations, CTT Vs. IRT ……. 111

Table 5.1. Summary of Methods by Research Objective ….……. 116

Table A.1. Course Descriptions …….. 164

Table B.1. Commonly Used Statistical Procedures …… 169

LIST OF FIGURES

Figure 2.1. Example: Item Characteristic Curve (ICC) …….. 47

Figure 2.2. Example: Item Characteristic Curve (ICC) for Two Items …….. 54

Figure 2.3. Example: Item Information Functions (IFF) …… 54

Figure 2.4. Example: Test Information Functions (TIF) ……..... 55

Figure 2.5. Item-peon Map of a Latent Trait (Construct) ……... 63

Figure 4.1. One factor Model (Prior to Estimation) for Assessing Unidimensionality …… 88

Figure 4.2. Distribution of Items and Persons on the Common Scale …… 94

Figure 4.3. Test Characteristic Curve of the SAGS Assessment …… 95

Figure 4.4. IIF's of Items in SAGS with Different Difficulties ……. 101

Figure 4.5. TIF of the SAGS Instrument …… 102

CHAPTER ONE

INTRODUCTION AND GENERAL INFORMATION

This chapter introduces the study and describes the problem being investigated, the purpose and the significance of the study. The context of statistics education and the types of statistics education research available will be outlined as well as previous attempts to assess students' statistics knowledge and skills. The proposed research objectives and methods to achieve these objectives will also be discussed in addition to definitions of key terms, assumptions, and limitations of the study.

Statistics is a quantitative approach to the analysis of empirical data for the purpose of making decisions and drawing conclusions in the presence of variability (Montogomery & Runger, 2013). As a discipline, statistics plays an important role in the conduct of applied research, as it facilitates making judgments from available data and information representing dynamic real-world scenarios (Dowdy, Wearden, & Chilko, 2011; Healey, 2014). It is not surprising that statistics education is receiving increased attention, as evidence based and data based quantitative approaches are characterized by growing credibility for making conclusions and educated decisions based on empirical research (Cumming, 2013; LoBiondo-Wood & Haber, 2014). As beginning researchers, most students are exposed to conducting empirical research in graduate school (Agre, 1997). Typically, many students in behavioral and social sciences complete a thesis or dissertation requiring the substantial use of statistics (Karadağ, 2010; Onwuegbuzie, 2002). However, some of these students find it difficult to grasp statistical concepts taught in the classroom and apply them to solve their research problems, and this may ultimately lead to frustration completing their academic work (Alccaci, 2012; Chiesi & Primi, 2010).

Currently, statistics courses are offered in university programs at the graduate and undergraduate levels across many disciplines (Ben-Zvi & Garfield, 2004). Students perusing degrees that require statistics are usually mandated to enroll in introductory statistics courses at the beginning of their degree programs (Feinberg & Halperin, 1978; Henry, 2013, Onwuegbuzie & Wilson, 2003). This is especially true for students attending graduate programs that are traditionally quantitative, as in education and other behavioral and social sciences disciplines where introductory statistics and/or quantitative research methodology courses are usually required (Chiesi & Primi, 2010). Even though graduate students may be required to enroll in statistics and research methods courses, Statistics has become an anxiety inducing subject for many students (Hannigan, Hegarty, & McGrath, 2014; Onwuegbuzie, Da Ros, & Ryan, 1997). In social and behavioral science disciplines statistics anxiety is observed among 80% of students (Onwuegbuzie & Wilson, 2003) creating a major concern for statistics educators (Perepiczka, Chandler, & Becerra, 2011).

Students who have had statistics at the undergraduate level encounter different types of statistics courses which results in students with varying level of knowledge and skills. The manner in which each statistics course is taught can be dependent on the field/discipline of study (Bryce, Gould, Notz, & Peck, 2001; Tarpey, Acuna, Cobb, & De Veaux, 2000). Courses offered to students in behavioral sciences disciplines tend to have a more applied approach while courses offered to engineering and other hard science majors offer a more mathematical and computational approach (American Statistical Association, n.d.; Bryce et al., 2001; Society for the Teaching of Psychology, 2014; Trapey et al., 2000). Also, statistics instructors use a variety of ways to teach statistical concepts and this diversity may result in students having different knowledge and skills in statistics (Davis, 2004; Knypstra, 2009; Williams, 2010). Thus, when

undergraduates move to a graduate level class, they represent a diverse range of prior statistics knowledge that can vary considerably from student to student (Haapala, 2002; Welch, et al., 2015; Onwuegbuzie, 2003; Pagano, 2006).

Problem Statement

Even though the learning objectives and skills developed through different statistics courses might vary from one course to another, in general, the overall goal of most applied statistics courses is to develop students' statistical problem solving skills in order to prepare them to deal effectively and efficiently with applied research problems outside the classroom (Samuels, Witmer & Schaffner, 2012; Yilmaz, 1996). In order to apply the concepts learned in the classroom to real world situations, it is very important for students to have sound basic statistics knowledge, skills, and experience along with the self- confidence and interest (Finney & Schraw, 2003; Healey, 2014; Williams, 2010).

Understanding the building blocks of statistics correctly is vital for students to develop sound foundational knowledge and promote their ability to successfully apply statistics (Garfield, 1995; Hanushek & Jackson, 2013). As a science for collecting, organizing, summarizing and analyzing data, statistics theory provides a wide range of tools and functionalities. In statistics, descriptive techniques/statistics include collecting, organizing and summarizing information that can be used to describe a sample and the associated variables under study (Trochim, 2006). Descriptive statistics is easily taught and learned as it consists of minor graphing and simple summarizing (Noether, 2012; Peck, Olsen, & Devore, 2015). Also, descriptive statistics is the basis for most statistical investigations. Thus most of the introductory level courses and textbooks cover these topics at the start (Trochim, 2006; Noether, 2012; Peck, et al., 2015). Next, introductory courses cover probability which is considered as important component for

understanding the mechanics of inferential statistical procedures (Hawkins, Jolliffe, & Glickman, 2014). Inferential statistics which includes statistical tests and other various statistical procedures are used to test various research hypotheses which can then be used to generalize results from sample data to larger populations (Healey, 2014, Noether, 2012; Peck, et al., 2015). However, especially in the case of statistics courses offered for students not majoring in statistics, important content of probability is usually avoided or given less attention within introductory courses (Healey, 2014; Noether, 2012; Peck, et al., 2015). Thus, students in applied disciplines who take statistics as service courses (Gordan, 2004; "Learn and Teach Statistics and Operations Research", 2013) may learn inferential statistics without a solid mathematical foundation which would be necessary to develop deeper statistics knowledge that can efficiently applied to solve practical problems (STATtr@K, 2012; Vance, 2015). Moreover, statistics courses offered as a service are normally designed to cover relatively broader content in a more superficial manner. Thus instructors have to skim over some topics (Noether, 2012; Peck, et al., 2015) which cause students to develop imprecise conceptual knowledge and lower ability to link statistics with practical applications (Yilmaz, 1996).

Even though the quality of statistics teaching is growing at all educational levels, more than two thirds of students in behavioral and social sciences believe statistics courses are difficult and an unpleasant subject to learn (Ben-Zvi & Garfield, 2004; Berk & Nanda, 1998; Garfield & Ben-Zvi, 2008; Onwuegbuzie, 2003), and they encounter an uncomfortable level of anxiety with statistics courses (Onwuegbuzie & Wilson, 2003). Ben-Zvi and Garfield (2004) described some of the difficulties that have been recognized when teaching and learning statistics. First, motivating students to learn statistics is challenging as the statistical concepts and rules are complex and difficult in nature. Second, many students find learning statistics difficult as they

lack the required knowledge in underlying mathematical theories. Third, students tend to become confused when selecting appropriate statistical tests or procedures to answer given problems. Thus, they tend to rely on their teacher to select an appropriate statistical procedure. Fourth, some students expect one correct answer and interpretation for each statistical problem, and find it challenging to deal with messy data and different interpretations according to different assumptions. Although there are these numerous challenges, students conducting applied research are required to use statistical methods (Healey, 2014; Harris & Jarvis, 2014; Devore, 2015). Thus to avoid confusion when selecting appropriate statistical methods these students usually have to seek help from statistics experts (Alacaci, 2014; Kirk, 1991; Vance, 2015). Unfortunately, providing such expert support incurs considerable cost to universities or to the student (Vance, 2015).

Some students are likely to express frustration about learning abstract statistical concepts and applying what they learn in the classroom to authentic situations (Garfield, 1995). Often, students face difficulties with deciding between various statistical procedures and tests to use with different research or practical scenarios (Alacaci, 2012; Bessant, 1992). For example, although students learn how to test hypotheses on mean comparisons using t tests and Analysis of Variance (ANOVA) in a classroom, a major challenge most students face is that they do not know when to apply it given a new situation or a dataset. Vanhoof et al. (2006) found that undergraduate students have relatively negative attitudes towards using statistics in their field of study, even though they have relatively positive attitudes towards the statistics courses they are taking. Furthermore, Ben-Zvi and Garfield (2004) pointed out students' dissatisfaction on their ability in applying statistics even after formally studying statistics at the college and graduate level.

The above discussion highlights the importance of students both knowing the fundamentals of statistics and having the ability to apply them in real life or research situations. But a statistics course that has perfect balance between training students on these aspects has rarely been observed (Peck, Olsen, & Devore, 2015). Statistics educators have identified that developing and teaching statistics to students who come from different disciplines is a challenging task (Delucchi, 2014; Dunn, Smith, & Beins, 2012; Pagano, 2006). According to Tishkovskaya and Lancaster (2012) another problem for teaching statistics with students from diverse backgrounds is attributed to deficiencies in basic statistical knowledge. They also emphases inadequacies in prerequisite mathematics skills as another problem associated with effectively teaching statistics. Moreover, these authors mentioned that statistics courses given as 'service teaching' (Gordan, 2004; “Learn and Teach Statistics and Operations Research”, 2013) often teach statistics with no link to any specific subject area, or by subject-specific specialists who are not statisticians which cause problems (Tishkovskaya & Lancaster, 2012). Finally, they identified several other factors, ‘Math-phobia', 'statistics anxiety', negative attitude towards statistics, pre-dispositions against statistics, and lack of interest displayed in students from other disciplines as obstacles for statistics education. Even with such challenges instructors teach classes and students continue to take statistics courses at the undergraduate level. Thus, entering graduate students reflect a wide range of fundamental statistics skills related to applying statistics (Griffith et al., 2012) to solve real life or research problems.

Many undergraduate students will pursue graduate education to address their interests in research (Hathaway, Nagda, & Gregerman, 2002) as well as career employment interests (Hawkins, et al., 2014; Zeph, 1991). As occupations requiring a graduate degree will increase by around 20% in 2020 (Sommers & Franklin, 2012), graduate student population are increasing in

colleges and universities (Allum, 2014; Allum & Okahana, 2015). At the graduate level most of the behavioral and social sciences students will likely be asked to a take statistics class to support doing their empirical research (Feinberg & Halperin, 1978; Henry, 2013, Onwuegbuzie & Wilson, 2003). Therefore, most of the departments/colleges conducing graduate programs offer a series of statistics courses for their graduate students. In some cases, students take courses from a statistics department to fulfill their learning needs if their college departments do not offer particular statistics courses (Schmidhammer, n.d.). However, as students enter graduate programs from various disciplines and with different levels of statistics knowledge and skills, selecting an appropriate statistics course to take at the graduate level becomes a dilemma for students as well as faculty teaching or mentoring those students (Dunn, Smith & Beins, 2012; Gelman., Carlin, Stern, & Rubin, 2014). At the graduate level, some students may need to begin learning statistics with basic courses, while for other students basic courses might only provide a repeat of similar content that they already have mastered. Thus, it is important to evaluate students' baseline knowledge in applying statistics for research (Barlow, 2014). Then students can be place in the most suitable statistics course to advance their knowledge in efficient way.

Assessing students' ability related to conducting applied research in a scientific manner is important and it will enhance the teaching and learning statistics (Bidgood, Hunt, & Jolliffe, 2010; Garfield & Franklin, 2011). However, little research has been conducted on developing assessments that measure students' statistics knowledge required for conducting empirical research (Barlow, 2014), even though there is a growing trend toward recommending students take basic and advance statistics classes to address their research needs (Feinberg & Halperin, 1978; Henry, 2013, Onwuegbuzie & Wilson, 2003). In statistics education literature, there are assessments available for measuring college and school levels students' statistical reasoning,

literacy and thinking (Allen, 2003; DelMars, Garfield, Ooms, & Chance, 2007; Grafield & Zieffler, 2012; Garfied, 1998a; Garfileld, 1998; Pfannkuch, & Wild, 2004; Stone, et al., 2003; Sundre, 2003; Ziegler, 2014). But in-depth review of these instruments shows difficulties in using them for assessments with graduate student populations, as these were developed for introductory level courses. Also these instruments focus on assessing conceptual knowledge about statistical procedure rather than assessing skills in applying statistics to solve applied research problems. Further, most of these instruments were developed to measure a narrow range of statistics knowledge that is not sufficient for addressing research questions answered in master thesis and doctoral dissertations in behavioral and social sciences disciplines (Curtis & Harwell, 1998; Hsu, 2005). Thus, to address these deficiencies it is important to develop a valid and reliable instrument to measure statistics knowledge specific to conducting empirical research.

Purpose of the Study

The purpose of the current study was to develop an instrument intended to measure statistical research methodology knowledge for conducting quantitative research for graduate students in education and other behavioral and social sciences (Statistics Assessment of Graduate Students). Developing such an instrument includes determining reliability and validity, with the ultimate goal of providing a valid and reliable measure for assessing graduate students' statistics knowledge for doing empirical research. Initially establishing preliminary item characteristics and validity evidence was necessary for this instrument entitled Statistics Assessment of Graduate Students (SAGS) instrument.

Thus, one objective of the study was to investigate the efficacy of Rasch modeling to develop SAGS. Further the study investigated the potential of using Item Response Theory models, 1 parameter (1 PL), 2 parameter (2 PL), and 3 parameter (3 PL) IRT for the development

of the SAGS item inventory/question bank which provided an opportunity to develop tests that will facilitate wider practical applications such as developing test targeted towards identification of high ability student to offer scholarships or identification of low ability students to give additional academic support. Moreover, the proposed study compared and contrasted the methodological advantages of using models selection criteria that guard against the interdependency of item parameters when developing assessments using IRT approach.

Overall, the study addressed four research objectives aligned with the main purpose of the study:

1. Establish content validity evidence of the SAGS instrument
2. Examine the model fit of the SAGS items to a Rasch model
    a. Test the assumptions of unidimensionality and local independence.
    b. Identify item difficulties and analyze the item information/test information of the SAGS instrument.
    c. Analyze the quality of item distractors of the SAGS instrument.
3. Examine the reliability and validity evidence of the SAGS instrument
    a. Assess the reliability of the SAGS instrument through the analysis of various reliability and separation indices.
    b. Assess construct, predictive, and other types of validities of the SAGS instrument through construct maps and known group comparisons.
4. Examine the model fit of the SAGS items to 1 PL, 2 PL and 3 PL IRT models based on simulated data.

a. Investigate the performance of novel information complexity criteria *(ICOMP)* over other model selection criteria for determining the best fitting IRT model.
b. Identify item difficulty, discrimination and guessing parameters.
c. Compare person ability and item location estimates (difficulty, discrimination, and guessing) from IRT models to those of traditional Classical Test Theory (CTT) indices.

Significance of the Study

Even though instructors of undergraduate statistics classes typically assess end of course knowledge gains (Delmas, 2002; Garfield & Delmas, 2010), few of them assess students' knowledge at the start of the course. These types of assessments are seldom observed in literature associated with graduate level statistics (Barlow, 2014). In general, individuals and organizations working on innovative educational techniques have suggested that students' basics knowledge and skills should be assessed prior to teaching the class to facilitate the assessment of effective instructional activities (Carnegie Melon Eberly Center, n.d.). This would appear to be extremely important for statistics teaching at the graduate level, considering the diversity of students in typical statistics class offered as a service course (Pagano, 2006; Yilmaz, 1996). However, accurately assessing students' basic statistics knowledge for conducting empirical research remains a challenging task. The current literature does not reveal any assessments that measure students' ability to select appropriate statistical procedure to analyze given research scenarios (Allen, 2003, Delmas, Garfield, et al., 2007; Garfield & Zieffler, 2012; Grafied, 1998a; Garfiled, 1998; Garfield et al., 2002, Stone, et al., 2003; Sundre, 2003; Zeigler, 2014). However, the literature revealed several closely related measures that have been developed to assess

students' statistics knowledge and skills (Biostatistics and Clinical Epidemiology Skills assessment, by Barlow (BACES) (2014); Statistical Reasoning Assessment by Garfield (2003); Comprehensive Assessment of Outcomes in a First Statistics Course (CAOS) by Delmas et.al (2002); Assessment Recourse Tool for Improving Statistical Thinking (ARTIST) Topic Scales by Delmas et.al (2002); Goals and Outcomes Associated with Learning Statistics (GOALS) by Garfield et al. (2012); Basic Literacy in Statistics (BLIS) by Zieffler (2014)). Moreover, there are tests such as the knowledge and ability test for mathematical statisticians by Statistics Canada (Statistics Canada, n.d.) as well as United Kingdom government statistical service assessment centers' written test, developed for basic knowledge assessments to select candidates for statistics jobs. However, all above mentioned measures have many limitations when used or adopt to measure graduate students statistics knowledge and skills for doing empirical research. Therefore, the present study aims to develop and validate a new measure (the SAGS) to assess graduate students knowledge in selecting appropriate statistical tests or procedures.

Such baseline assessment clearly provides opportunities to identify individual students' strengths and weaknesses (Heitman, Olsen, Anestidou, & Bulger, 2007). Therefore, this assessment could help instructors identify students who need supplementary instructional support, and help them with changes they need to make on their instructional techniques to effectively teach course content (Carnegie Melon Eberly Center, n.d.). Students could use such assessment to self-evaluate their knowledge and make adjustments to their statistics training. With regards to graduate level statistics education, this baseline assessment could be used to select appropriate statistics courses, whether introductory, intermediate or advanced depending on students' prior statistics knowledge. As Statistics Assesment of Graduate Students (SAGS) instrument measures the statistical knowledge and skills for doing empirical research, this

instrument could be a screening tool for employers seeking candidates for their applied research oriented job positions. As a reliable and valid measure, SAGS will provide accurate estimates of students' statistics knowledge related to doing research; therefore, such a measure can be used to make better educational decisions associated with student statistical ability.

Limitations of the Study

The major limitation of the study was the SAGS administration was in a rather uncontrolled environment. Tests was administered using "Qualtrics" on-line survey management system, but no time constraint was set for participants to compete the SAGS cognitive questions Further, SAGS was administered to a purposive sample and the compositions of the participants were relatively unknown, which limits the generalizability of the study. The other limitation of the study is that the IRT based (especially 2 PL and 3 PL models) analysis was based on simulated data.

Definitions of Key Terms

Education and Other Behavioral and Social Sciences (EBS). Branch of science that deals primarily with human actions and disciplines including education. Other disciplines include but are not limited to psychology, sociology, anthropology, social work, political science, demography and geography.

Statistical Research Methodology Knowledge. Individuals' ability to select appropriate statistical test/procedure among several other statistical test/procedures to answer given research questions/situation.

Statistics as a Service Course. A "Service Course" is a course (in statistics) for students who are not majoring in Statistics or Mathematics, but majoring in some other subject, such as Business or Medicine or Education. For some students it is a terminating course and they will

never take a statistics courses again. For some students it is the precursor to further applied statistics courses (Gordan, 2004; "Learn and Teach Statistics and Operations Research", 2013).

Cognitive Items. Cognitive items are the questions that measures the statistics research methodology knowledge. These items designed to assess what students know about statistics, about various statistical test and procedures, and when to use these statistical tests or procedures.

Item Response theory (IRT). IRT refers to associated mathematical models that relate person's abilities quantified using a latent construct (θ) and item qualities quantified using various item parameters, namely difficulty (b), decimation (a), and pseudo-guessing (c) to the probability of response to items on the assessment. Item response theory models describes the observed responses and the relationships between person's abilities and item qualities are specified through 1 PL (1 parameter logistics), 2 PL, and 3 PL in IRT framework (Furr & Bacharach, 2014; Templin, 2104).

Rasch Model. The Rasch model is the simplest IRT model, and the discrimination parameter equals the value 1. However, philosophically, the Rasch model is taken as a criterion, specification or statement for the structure of the responses, rather than a mere statistical description of the responses (Brown, Templin, & Cohen, 2014; Wright, 1992; Linacre, 2016).

Reliability. Consistency of differences in respondents' observed scores with differences in their true scores (Furr & Bacharach, 2014).

Validity. The degree to which an instrument measures what it is supposed to measure (Furr & Bacharach, 2014).

Organization of the Study

Chapter one briefly introduced the problem under investigation, its context, four primary study objectives, and the methodological components that the study used to address these objectives. This chapter has also highlighted the limitations, and key definitions for the study.

Chapter two will present a complete review of the literature that informs the present study as well as the theoretical framework on which it is based. Chapter three will describes the details of the study's methodology for developing the SAGS instrument as well as administering and analyzing the results. Chapter four provides the results from collected data, and chapter five contains a discussion of these findings in detail as well as the study's implications and recommendations for future research.

CHAPTER TWO

LITERATURE REVIEW

The purpose of this study is to develop and validate Statistics Assessment of Graduate Students (SAGS) instrument which is designed to measure basic statistical research methodology knowledge for conducting quantitative research by graduate students' in education and other behavioral and social science disciplines. This chapter includes information about graduate statistics education, statistics education research, constructs associated with statistics educational community, research related to assessing statistics knowledge, previous studies on developing statistics knowledge assessments, a summary of available statistics assessments, review of Item response theory and Rasch modeling, and a brief description on how the SAGS that will be developed under the present study differs from currently available scales.

Graduate Students and Statistics Education

Why Statistics is Taught at the Graduate Level?

The graduate student population is increasing in colleges and universities (Allum, 2014; Allum & Okahana, 2015) and it is expected that the occupations requiring a graduate degree will increase by around 20% in 2020 (Monthly Labor Review, 2010). Given that the primary purpose of graduate schools is to prepare graduate students to assume professorial responsibilities, greater emphasis is given by curricular to develop their research skills as it plays an important role in order to generate practice knowledge apply in the field (Gilmore & Felton, 2010; Meerah et al., 2011). Research generates large volume of information (Dubois & Gershon, 2013; Murdoch and Detsky, 2013) and statistics has been the supporting science by promoting the analysis of research data in many disciplines (Healey, 2014; Harris & Jarvis, 2014; Devore, 2015). Moreover, as a tool for learning from data through data collection, analysis and interpretations, and as a science for dealing with uncertainty (American Statistical Association, 2015) statistics

plays a key role to make sense of and interpret a great deal of information (Joy, 2007; Healey, 2014; Keller, 2015). In particular, use of statistics is a common practice in the social science disciplines and good statistical knowledge is compulsory (Henry, 2013) for quantitative researchers to reach full potential as social scientists (Healey, 2014). Thus, for social and behavioral sciences graduate students it is mandatory to take at least one statistics course and/or a quantitative-based research methodology course as component of their degrees (Feinberg & Halperin, 1978; Henry, 2013, Onwuegbuzie & Wilson, 2003).

Even though many students enrolled in statistics and research methods courses, statistics has become the most unwanted and anxiety inducing subject for many students (Hannigan, Hegarty, & McGrath, 2014; Onwuegbuzie, Da Ros, & Ryan, 1997;). Statistics anxiety is observed among graduate students in many disciplines (Macher, Paechter, Papousek, & Ruggeri, 2012; Onwuegbuzie, 1998; Zeidner, 1991) and its prevalence is more in the social sciences (Davis, 2004). In social and behavioral science disciplines statistics anxiety is observed among 80% of students (Onwuegbuzie & Wilson, 2003) suggesting indicating major concern for statistics educators (Perepiczka, Chandler, & Becerra, 2011). Several studies have reported the negative attitudes towards statistics (Lalayants, 2012; Maschi et al., 2007) in the forms of anxiety, fear and resistance (Lalayants, 2012; Perepiczka et al., 2011). These facts cause feelings of inadequacy and low self-efficacy for statistics related activities which linked to performance in statistics and research methods classes (Blalock, 1987; Beurze, Donders, Zielhuis, de Vegt, and Verbeek, 2013; Dillon, 1982; Onwuegbuzie, 1999). Since statistical analysis is an integral part when conducting research (Alacaci, 2012) lack of statistics skills influences students' poor performance in research (Davis, 2004; Zanakis, & Valenzi, 1997) and in the case of graduate

students this delays and leads to complications in completing their thesis or dissertation (Onwuegbuzie, 1997; Onwuegbuzie, 1999; Rudestam, & Newton, 2014).

Importance of Selecting an Appropriate Statistical Procedure/Test

Identifying an appropriate statistical methods or techniques for a given research problem is essential for students for completing quantitative research for doctoral dissertations (Alacaci, 2012). To facilitate statistical research descriptive and inferential statistical methodologies are essential and will be helpful for designing, understanding, evaluating and carrying out research at the dissertation level (Dunn et al., 2012; Jala & Reston, 2011).

Students take statistics courses could possibly do particularly well in those courses by memorizing the content taught or applying the procedures to familiar or well-known problems (STATtr@k, 2012). However, they may lack the training to model the new research problems (Dunn et al.,2012; Marino, 2014), which can be considered as an important skill that is helpful to conduct their own research and perform various data analysis related to their own disciplines (Marusteri & Bacarea, 2010). Gardner and Hudson (1999) and Quilici and Mayer (1996) have documented the need for students to reach this level of understanding as an outcome of statistics education. These authors have identified one reason for the lack of these skills as the computational technology (Alacaci, 2012). Nowadays, most of the statistics courses are taught using user friendly statistical packages, which can be used to perform data analyses quickly (Thiese, Arnold, & Walker, 2014; Higazi, 2002). However, these statistical packages are not capable of identifying correct statistical tests or techniques for a given research problem and easily enable a student to conduct a wrong application of an inappropriate statistical test (Alacaci, 2012; Larwin & Larwin, 2011; Marino, 2014).

Alacaci (2012) has clearly identified that two type of expert knowledge is required during statistical data analysis. One is computational expertise, which is the advanced computational ability of using statistical software. As mentioned before the other very important type is statistical expertise, which is considered as the skill of selecting the appropriate statistical techniques and drawing sensible conclusions from results (Hand, 1984). However, the definition of statistical expertise seems not clearly operationalized in the literature and several authors defined statistical expertise using several constructs such as statistical literacy, reasoning and statistical thinking while showing some overlap among these definitions (Alacaci, 2012; Ben-Zvi & Garfield, 2004; Gibbons & MacGillivray, 2014; Taplin, 2003). Thus, in order to better understand statistical expertise, it is important to distinguish between these constructs in the context of statistics educational research and next section is devoted for this purpose.

Statistics Education and Statistics Education Research

Statistics is considered as a subject associated with higher level of anxiety for many students in their courses of study (Baharun & Porter, 2010; Pan & Tang, 2004). Studies have shown that difficulty faced by the students in learning statistics could be compared to that of leaning a foreign language (Onwuegbuzie, 2003). Due to the fact that statistics is used for decision making across wide areas of fields, the number of students' enrollment in statistics courses at college level has increased (Peris & Beh, 2012, AMST New letter, 2015). With the diversity of students' population, teaching has become more and more in demand and this can be challenging for statistics instructors (Hulsizer & Woolf, 2008; Tishkovskaya & Lancaster, 2012). To overcome these learning difficulties, statistics educators have always being looking for the best strategies to improve student learning and they have shown their willingness to conduct research for exploring and solving their problems (Zieffler, et al., 2008).

The primary goal of educational research is improving instructions which can ultimately lead to effective student learning (Raudenbush, 2005; Consortium for Educational Research and Evaluation, n.d.). Thus, the goal of statistics education research can be defined as improving teaching statistics, which lead to improve student learning (Zieffleret al., 2008). Also, Consortium for the Advancement of Undergraduate Statistics Education (CAUSE) emphasized that statistics education research has direct implication on classroom instruction along with providing the opportunity for developing new research questions in this area. Statistics education research has been recognized as an interdisciplinary but distinct field of study (Tishkovskaya & Lancaster, 2012; Zieffler et al., 2008). Researchers from diverse backgrounds such as educational psychology, psychology, statistics, statistics education and mathematics education have been involved with research in this area and have provided valuable contribution for the advancement of statistics education (Consortium for the Advancement of Undergraduate Statistics Education, n.d.).

According to Zieffler et al. (2008) the studies available in the literature and the contributions made by statistics education researchers can be classified into four major categories: 1) Studies that explore about misconception and inaccurate statistical knowledge, 2) studies that focus on assessing cognitive outcomes, 3) studies that focus on assessing non-cognitive outcomes, and 4) studies focus on teaching statistics at college level. Looking at these four categories it is clear that all these are connected together and they all have important contributions to the advancement of field of statistics education. Studies on the second and third components focus on assessment and provide statistics educators with important tools to evaluate their students and then based on that information to improve their instructions. Such studies have addressed the important measurement issues in statistics education. Further, these studies are

both quantitative as well as qualitative (Gordon, 1995; Groth & Bergner, 2005; Mathew and Clark 2003; Reid & Petocoz, 2002; Zieffler et al., 2008). They have taken approaches to identify quality or attribute to be measured, define a set of operations by which the attribute may be manifest or perceived, and then to established sets of procedures or definitions for translating observations into quantitative statements of degree of amount (Throndike, 2004).

Importance of Assessing Cognitive Outcomes

Measuring students' cognitive outcomes and abilities is crucial to enhance learning and teaching process. Many assessment tools are used for this purpose (Delmas et al., 2007; Earl, 2012; Gathercole, Pickering, Knight, & Stegmann, 2004; Gold & Harris, 2013). Assessment provides a measure of whether the students are learning properly, acquiring the necessary knowledge, and developing the skills that are stated with the course objectives (Barnes, 2015; Bennett, 2011; Wright, 2008). It is also an indicator of the students' success on achieving the required competencies (Bremner, Blake, Long, & Yanosky, 2014; Stiggins, 2005). Thus, assessment is an integral part of instruction, as it determines whether or not the goals of education are being met (Garcia, 2013; Wright, 2008). Assessment affects decisions about grades, placement, advancement, instructional needs, and curriculum (Gonsalvez et al., 2013; Lehmann, 2014). Moreover, assessment results will indicate the strengths and weaknesses of the students (Educational Testing Services, 2003; Ennis, Lane, & Oakes, 2011) and it will be helpful to organize the learning process by giving more weight to the areas or sections that are difficult to particular students (Fuchs and Fuchs, 2006; Will, 1986).

According to the publication titled, "Linking classroom assessment with students learning" by Educational Testing Services (2003) and several other online sources, assessment provides important feedbacks for teachers to design instructional process more effectively. From

the teacher's perspective, it provides information on how the teaching instruction grasp by each student. Therefore, teachers will be able to identify difficulties faced by each student and it can be used to increase student leaning (Hasbrouck & Tindal, 2006). Using assessments teachers are able to evaluate their teaching (Astin, 2012) and are able to make curriculum modifications and instructional improvements to meet individual learning needs (Jonassen & Grabowski, 2012; Kuh, Jankowski, Ikenberry, & Kinzie, 2014).

Assessing Statistical Constructs

The importance of assessment and its implications that have been mentioned before is directly applicable to statistics classrooms and generally for statistics education (Garfield & Chance, 2000). Since statistics is being recognized as a difficult subject for students in many disciplines, cognitive assessment in statistics may contribute to improve teaching and learning process. In the case of cognitive statistics assessments, several researchers have focused on construct associated with measuring statistics knowledge, such as statistical reasoning, statistical thinking, and statistical literacy (Dani & Joan, 2004). There are several quantitative studies related to the cognitive assessments. The development of a 20 item multiple choice Statistical Reasoning Assessment (SRA) (Grafied, 1998a); development of an assessment with 16 multiple choice items and open-ended items for statistics (Hirsch & O'Donnell, 2001); development of the 40 item Comprehensive Assessment of Outcomes in First statistics Course (CAOS) and ARIST topic scales for testing specific areas (Delmas, et al., 2007); development of Quantitative Reasoning Quotient (Sundre, 2003), Statistical Literacy Assessment (Sahin, 2012) are major studies reported in the current literature (these will be discussed in detail later in this chapter). Also, there are few qualitative studies that develop instruments to measure cognitive outcomes. Groth and Burger (2005) conducted a study to use metaphors to measure abstract sampling

knowledge of students. Also, the clinical interviews have used as a measurement instrument by Mathew and Clark (2003) to measure students' knowledge about descriptive statistics.

Non-cognitive assessments measure the beliefs and feelings about statistics and it's utility. Several instruments exist in the literature that measure different constructs such as attitudes, self- efficacy and anxiety associated with statistics. Finney and Schraw (2003) developed two instruments, Current Statistics Self-efficacy (CSE) and Self Efficacy to Learn Statistics (SELS), to measure statistics self-efficacy. For measuring attitudes, there are instruments available in the literature. They include Survey of Attitudes Towards Statistics (STATS) by Schau, Stevens, Dauphinee, and Del Vecchio (1995), Attitudes Towards Statistics (ATS) scale by Wise (1985), and Statistics Attitude Survey (SAS) by Roberts & Bilderbase (1980). However, there are many other studies and instruments to measure statistical related constructs (Delmas et al., 2007; Grafield et al., 2012; Grafield, 1998a; Stone, et al., 2003; Sundre, 2003; Zeigler 2014). Descriptions of all these studies and instruments indicate that they are meaningful and appropriateness for particular groups of participants. Looking at the cognitive and non-cognitive assessments, there are very few instruments available to measure cognitive outcomes (Zieffler et al., 2008). Therefore, there is a need to develop a high quality instruments to asses important and agreed upon learning goals to advance the field of statistic education (Zieffler et al., 2008). However, to efficiently develop assessments, the constructs that are associated with statistics education should be clearly defined.

Defining the Cognitive Constructs Associated with Statistics Education Research

The main objective of statistics instruction is to facilitate students' ability to construct reasoned descriptions, judgments, and inferences and opinions about data (Garfield, 2003). To effectively perform such statistical activities, students require different levels of cognitive

abilities. Beyth-Marom, Fidler, and Cumming (2008), Chance (2000), Delmas (2002), Garfield (2002) and Rumsey (2002) loosely introduce these cognitive processes or constructs as statistical reasoning, statistical thinking, and statistical literacy. However, there is no consistency among statistics educators or researchers about definitions of statistical reasoning, thinking, or literacy; they all use different terms interchangeably and understandings of these cognitive processes (Ben-Zvi & Garfield, 2004). Even though the relationships between these constructs are complex in nature (Tempelaar, 2004) summarizing those articles, Ben-Zvi and Garfield (2004) came up with relatively more concise definitions highlighting that no formal agreement was available at that time point.

Statistical reasoning is defined as "The way people reason with statistical ideas and make sense of statistical information" (Ben-zvi & Garfield, 2004, pg 7). Activities that fall under reasoning includes the ability to understand statistical procedures, explain these procedures, and comprehensively interpret results obtained by performing such procedures. Examples of statistical reasoning are making interpretations based on raw data or statistical summaries of data. Further, they indicate connecting one concept to another (central tendency and variability), or combine ideas about data and chance as statistical reasoning.

Statistical thinking has been broadly defined as the "understanding of why and how statistical investigations are conducted and the "big ideas" that underlie statistical investigations" (Ben-zvi & Garfield 2004, p.7). Statistical thinking is attributed to understanding and utilizing the context of a problem in developing statistical investigations. Also, selecting and using appropriate methods of data analysis are the main component of statistical thinking. Further, drawing conclusions, and recognizing and understanding the entire statistical research process is the central part. Critiquing and evaluating results of a problem solved or a statistical study is

another important aspect. Performing statistical research by understanding and utilizing the central ides such as variability, correlations/causation, and effects sampling is another attribute of statistical thinking.

Statistical literacy has been considered as the understanding of concepts, vocabulary, and symbols, along with understanding of probability as a measure of uncertainty (Ben-zvi & Garfield, 2004). Having the understanding of basic and important skills that may be used in recognizing statistical information or research results is considered as a major component of statistical literacy. Examples of such skills comprise of being able to organize data, construct and display tables, and work with different representations of data.

There are similarities and differences among these three constructs of statistical reasoning, statistical thinking, and statistical thinking. Recognizing these constructs are important to effectively formulating learning goals for students, designing instructional activities, and evaluating learning by using appropriate assessment instruments.

Review of Previous Studies on Developing Statistics Skills Assessment Scales

Even though the current literature reveals several studies about developing assessments to measure students' statistical cognition in terms of statistical reasoning, statistical literacy and statistical thinking, studies on developing instruments for assessing students' statistics knowledge for doing applied research are few. Only eight instruments that specifically relate to assessing students' statistics ability can be currently found in literature, but they have their own limitations and/or generalizability issues.

Statistical Reasoning Assessment (SRA) (1998)

When different statistical constructs became popular among education community in late nineties, there were no instrument existed to assess high school students' ability to understand

statistical concepts and apply statistical reasoning (Garfield, 2003). Addressing this need, Garfield and Konold developed and validated the Statistical Reasoning Assessment (SRA) as a part of NSF funded "Chance Plus Project" for evaluating the effectiveness of a new statistics curriculum for the U.S. high schools (Garfield, 1996; Garfield, 1998(a); Garfield, 1998(b), Garfield, 2003; Konold, 1989). The items in SRA consist of 20 multiple choice items where each item describes statistics and probability problems. Most response options explain the rationale for particular choice which associated with different types of reasoning skills, and students have to select the best answer that matches there thinking. A special scoring method was adapted by the users to provide two types of scores: correct reasoning which includes 8 subscales and common misconceptions which include 8 subscales.

Items in the SRA instrument was carefully developed to represent correct and misconception or incorrect reasoning while going through a long revision process. Content validity was established by reviewing these items by expert (Garfield, 1998a; Garfield, 2003) before going through pilot testing. However, the field administration of this instrument did not produce expected level of reliability and validity evidence. Although limited empirical evidences are available (Garfied, 1998b; Grafied & Chance, 2000; Liu, 1998), the administration indicated a low internal consistency reliability which indicates that there was no single trait of statistical reasoning was measured by this instrument. All empirical studies analyzed the above mentioned total subscales scores and test-retest reliabilities turned out to be 0.70 for correct reasoning and 0.75 for common misconceptions. Also, Liu (1998) examined the performance of SRA through a cross cultural study using USA and Taiwanese college students and found out that there is differential country effect and gender effect, which may due to non- existence of discriminant validity of the instrument. Further, Tempelaar (2004) identified a limitation of SRA

with discrimination validity and suggested to change the difficulty level of SRA instrument to allow for efficient discrimination in reasoning abilities in the context of using it with different populations and subgroups of college students.

The SRA instrument has provided an important step towards developing instructionally friendly assessment tools (Sundre, 2003), and it is one of the few objective instruments for assessing students' statistical reasoning abilities (Tempelaar, 2004). The SRA was formatted as a paper-and–pencil instrument that is an easy to administer, and it gives a useful measure of statistical reasoning ability, but it is not recognized as a comprehensive measurement tool (Garfield, 2003) and only represent a small subset of reasoning skills and strategies. Also, this was developed targeting the pre-college level students and introductory level college statistics students. Thus, the scope of the instrument only covers the skills associated with basic and commonly used statistical procedures. In addition to the weakness with comprehensiveness and above mentioned reliability and validity aspects, it was criticized by Tempelaar (2004) for measuring mathematical reasoning rather than statistical reasoning. With those limitations, using SRA is problematic and there is room for devolving new instruments and large item bank for assessing statistical reasoning.

Quantitative Reasoning Quotient (QRQ) (2003)

The Quantitative Reasoning Quotient (QRQ) instrument is a revision of Garfield's 1998(a) 20-item Statistical Reasoning Assessment (Sundre, 2003). The author took Garfield's 1998(a) advice and modified the instrument to alleviate recognized limitations. Low internal consistency of SRA was addressed through creating additional items from existing alternative response options in SRA items. When using SRA, due to the selection of multiple answers there were possibilities that students may be identified for having the correct reasoning on one concept

at the same time identified as having misconceptions about the same concept. But using additional items and asking students to only agree or disagree to those in QRQ, students were able to effectively distinguish the correct reasoning and misconception on a particular concept. Also with doubling size of instrument to 40 items through additional items, QRQ covered a lager content area by addressing another limitation of the SRA. Finally, this instrument addresses the scoring difficulty of SRA through developing a computerized scoring system for QRQ.

The final QRQ instrument consists of correct reasoning and misconception components as the SRA, but it includes 3 additional subscales for correct reasoning while 7 additional subscales for misconceptions. Content validity of the QRQ was established through reviewing items by several panels of faculty with continuous refinements. During the two cycles (semesters), the authors administered this new instrument to a large number of students at a large college campus and finally reported the descriptive statistics for each of the 26 subscales for those samples of students. The original study of developing the QRQ reported reliability evidence in the form of internal consistency reliability for two cycles as 0.55 and 0.62, but no other validity evidence was reported. Further, authors indicated that faculty identified the reliabilities to be low for QRQ, and they suggested increasing the items to cover more content coverage. Thus, psychometric properties and faculty opinions still question to what degree the QRQ can be improved for assessing statistical reasoning skills over SRA and still this can be considered as imperfect measurement tool for statistical reasoning.

Instruments in (ARTIST) Project and its Relatives, CAOS (2002)

The Comprehensive Assessment of Outcomes in a First Statistics course (CAOS) test was developed by Delmas et al., in 2004 as part of Assessment Resource Tools for Improving Statistical Thinking **(**ARTIST) project. ARTIST project which addressed the evaluation

challenges in statistics education identified by the researchers in this field (Grafield & Gal, 1999) had unique assessments to evaluate students' knowledge and skills covering a variety of topic areas in statistics. In addition, CAOS test was an important component in ARTIST system as it was designed as an instrument that would assess students' statistical thinking and reasoning after any first course in statistics (non-mathematical). Rather than focusing on computation and procedures, the CAOS test focuses on statistical literacy and conceptual understanding, with a focus on reasoning about important concept of variability.

The test was developed through a three-year process of acquiring and writing items, testing and revising items, and gathering evidence of reliability and validity. According to the ARTIST website CAOS test shows very impressive reliability and validity evidence. Based on a sample of 10287 students, an analysis of internal consistency of the 40 items on the CAOS posttest produced a Cronbach's alpha coefficient of 0.77. Also, the 2006 paper reports internal consistency reliability of 0.82 based from a sample of 1470 introductory students taught by 35 instructors from 33 higher education institutions from 21 states.

The content validity of the CAOS items was established through carefully designing the items based on existing items and with continuous reviews and refinements with experts in the subject areas. Also, the final version of the CAOS test which is improved through identifying the weakness observed by administering the earlier versions of the CAOS 1 through CAOS 3. Further, the websites report that there was unanimous agreement by a set of 18 expert raters that CAOS 4 measures important basic learning outcomes, and 94% agreement that it measures important learning outcomes. Based on this evidence, the assumption was made that CAOS 4 is a valid measure of important learning outcomes in a first course in statistics.

However, there are few criticisms about the CAOS test. First, the empirical studies by the authors revealed that when the CAOS test is administered as pre-test and post-test, even though higher gain was expected, in most CAOS administration there was a small gain of scores. Average percentage of correct increased by 9%, but was marginally statistically significant (Delmas et al., 2007). Considering the fact that this study was conducted with multiple states, colleges, and with multiple instructors, this study is generalizable than results attributed to one instructors or college where a possibility for underestimating the gain. Therefore, this minimal gain may be attributed to the problems with the CAOS instrument. Further, some raters indicated topics that they felt some important items were missing from the test. There was no agreement among these raters about the topics that were missing, which provide evidence on incompleteness of the instrument. Also, in the CAOS test fewer than half of the items that measure statistics literacy. Thus authors have created the Artist Topics scales that include more items that measure statistical literacy (Ziegler, 2014). As these tests are developed more than decade ago and considered not to be aligned with the modern introductory statistics courses. Thus, the CAOS test considered to be outdated (Grafield, et al., 2012; Zieffler, 2014).

Instruments in (ARTIST) Project and its Relatives, ARTIST Topic Scales (2002)

These scales cover 11 topics each consisting of 7-15 multiple-choice items to assess student reasoning and literacy in those particular topics. These topics areas are: Data collection, Data representation, Measure of center, Measure of spread, Normal distributions, Probability, Bivariate quantitative data, Bi-variate categorical data, Sampling distributions, Confidence intervals, and Significance tests, which are the focus of most introductory statistics courses for undergraduates.

ARTIST Topic Scales have advantages over the CAOS test. With the increase number of items, these scales were able to cover more content areas that ultimately enhance the measurement scope. Also these tests can be easily administered due to the existence of fewer items and reduce administration time. Moreover, these can be administered to test knowledge gain after teaching a particular topic as a formative assessment. Thus, tests show more practical usability in the classroom. However, these scales still show some opportunity for improvement. Even though, the number of literacy items has been increased, they are limited to definitions and simple calculations (Ziegler, 2014). Also, the reliability and validity evidence for these tests were never published, which leads to question the quality of the measurement and reduced practical usability. As similar to the CAOS test these topics scales are considered outdated and are inappropriate for evaluating the students' statistics ability with respect to modern day statistics courses (Grafield et al., 2012; Zeigler, 2014).

Goals and Outcomes Associated with Learning Statistics (GOALS) (2012)

Goals and Outcomes Associated with Learning Statistics (GOALS) instrument is one of the two primary instruments developed under the Evaluation and Assessment of Teaching and Learning about Statistics (e-ATLAS) to evaluate the effectiveness of reforms associated with teaching and learning of introductory statistics at tertiary level (The University of Minnesota, n.d.) According to the data collected in the 2005-2011 period in the USA, average CAOS test results have not indicated that the reforms have not lead to improve students' outcomes after the first statistics course (Grafield et al., 2012). Thus, statistics educators experimented with curricular with different leaning objectives, which has made the demand for new assessments and GOALS fulfilled this requirement (Chan & Ismail, 2014; Grafield et al., 2012).

GOALS instrument includes 20 forced-choices items and 3 open-ended items specifically designed to measure statistical reasoning. Finalized GOALS instrument (Goals v.2) includes 19 forced-choices items, which measure knowledge of ideas of samples and sampling, inference and p-values, levels of confidence, study design, and covariance. Four items in the GOALS assessment were identical to the corresponding CAOS items, and 12 items are based on modifications while remaining 7 items are specifically addressing the learning goals associated with simulation based content of the courses. However, the authors haven't still published any reliability and validity evidence, which would be useful to researchers who are willing to use it in the field.

Basic Literacy in Statistics (BLIS) (2014)

The content, taught in introductory statistics courses, has changed over the past 10 years and more and more simulation based methods such as randomization test and bootstrapping were included with the conventional parametric methods (Grafield et al., 2012; Ziegler, 2014). Thus, the use of assessment instruments such as CAOS test and ARTIST topics scales have become obstacle. Addressing this limitation Zeigler (2014) as her doctoral dissertation research developed the Basic Literacy in Statistics (BLIS) assessment to measure students' ability to read, understand, and communicate statistical information associated with the modern day statistics courses.

Initial items for the BLIS assessment were drafted based on the content available in textbooks used for introductory statistics courses that includes simulation. Also, the items in the GOALS assessment were merged with initial items to form the instrument. These items were then went through an extensive review by getting the comments from within 6 statistics educators at two stages (BILS and BLIS 1), students interview responses, and pilot testing. The

finalized test had 32 individual items and 5 testlets. The instrument was administered to 940 students from 34 introductory statistics course and collected data was analyzed using both classical test theory and item response theory methods. The BLIS 3 showed very good internal consistency with coefficient of alpha being 0.83. Also, confirmatory factor analysis results showed that items measure the single construct of Statistical Literacy showing strong evidence for construct validity. Even though the authors have stated that this instrument has good reliability and validity, they did not provide multiple versions of reliability evidence that would strongly support the utility of this assessment. Further, since this instrument is targeted to the simulation based courses, it will reduce the practical usage in the conventional statistical courses in different disciplines.

Statistics Concept Inventory (SCI) (2002)

The SCI is a multiple choice instrument developed to assess students' understanding of fundamental statistics concepts mainly targeting the engineering and mathematics based undergraduate students. Items for the test were developed by first identifying the important topics to be covered using modified Delphi method. Utilizing the questions in statistics text books, educational literature, and authors experience items for the test were created. The test was piloted during the fall 2002 semester at the University of Oklahoma with 139 students (Stone, et al., 2003). Following an extensive revision process, which included focus groups and individual expert reviews the test was finalized with 33 items.

Extensive psychometric analysis of SCI (Allen, 2006; Allen, Reed-Rhoads, and Terry, 2006; Allen, Reed-Rhoads, Terry, Murphy, and Stone, 2008; Allen, Stone, Reed-Rhoads, and Murphy, 2004; Stone, 2006; Stone, et al., 2003) did reveal acceptable reliability and validity evidence. Internal consistency raged from 0.57 to 0.71 for pre-tests and 0.58 to 0.86 for post-

tests. In his original study Stone addressed the concurrent validity of the SCI though correlating the SCI scores with scores on Survey of Attitudes Towards Statistics (SATS). Further, SCI showed criterion validity evidence as SCI scores significantly correlated with self-rated confidence in their answer, and structural validity evidenced through results of factor analysis. Thus, SCI has higher reliability and validity compared to the previously considered instruments. Close review of the individual items in the SCI showed that there is diversity of items that address graphical and descriptive data analysis, probability calculations, making inferences from samples, and selecting best statistics test and procedures. However, SCI items available in literature show more mathematical orientation. These items test more mathematical knowledge but not limited to of basic probabilistic, conditional probabilities, and distribution theory that might provoke anxiety to a student form non-mathematical discipline.

Statistical Literacy Survey (SLS)

Schield constructed an inventory about "Reading and Interpreting Tables and Graphs Involving Rates and Percentages" and developed it into "Statistical Literacy Skills Survey" (Schield, 2006). This survey consisted of both cognitive and non-cognitive items as it collected self-reported data on statistics skills as well as cognitive questions (Statistical Literacy Inventory). The instrument consisted of 55 cognitive questions and they were mostly based on interpreting statistics in graphs and tables. The survey was administered to diverse respondent groups including college students, college teachers, and professional data analysts. In literature there are no indications of a validity study for this survey and any reliability data, which restricts the use of SLS. However, data analysis revealed important information for the research community. The item- total score correlations, percentage of questions which were answered right were calculated, and by modeling different number of questions, Schield asserts that the

improvement of the instrument can be possible by eliminating some of the questions (Schield, 2008b, Schield, 2010). Moreover, he concluded that students showed difficulties in decoding the graphs, tables of rates and percentages, and comparing the data, as well as addressing this is an important component in statistics education. However, current literatures review unable to reveal any psychometric properties of SLS instrument. Summary for SLS and all the other assessment mentioned in this section is given in Table 2.1.

Development of the SAGS Instrument

Previously, in this chapter I described, (1) the statistical skills and knowledge needed to successfully perform research in educational and behavioral sciences, (2) the difficulties students have in leaning statistics and applying statistics when conducting their research, (3) students' lack of skills for selecting correct statistical procedure to address their research question as one of the major issues faced by the researchers, and (4) the need for assessments to help to measure the lack of knowledge and skills, which will allow instructors to diagnose limitations to better educate students and increase student outcomes. Furthermore, scholars engaged in statistical education research suggest the timely need to develop high quality instruments to assess important, and agreed upon learning goals to advance the field of statistic education (Zieffler et al., 2008). Thus, developing assessments that measure the statistics knowledge for conducting applied research (selecting appropriate statistical test/procedure) will an address important measurement gap, and it will be useful to advance the field of statistics education.

Table 2.1. Currently Available Cognitive Statistics Assessments.

Scale Description	Content	Psychometric Properties	Merits, Limitations and Suggestions
Name: Statistical Reasoning Assessment (SRA) **By:** Joan B. Garfield (University of Minnesota) **Year:** 1998 **Goals:** Assessing high school students' ability to understand statistical concepts and apply statistical reasoning. **Target Population/s:** High School, Undergraduate	**Items:** 20 Multiple choice Items on probability and Statistics **Responses:** Selected response options from the list of alternatives were considered for correct reasoning or misconceptions **Scoring:** Special scoring method attached to each response option for each question to identify correct reasoning (through 8 subscales) and misconceptions (through 8 subscales)	**Content Validity:** Established through extensive expert reviews along with pilot testing items **Reliability:** Low internal Consistency, Test-retest; .7 for correct reasoning and .75 for misconceptions. But most of the empirical studies did show low reliabilities. Low discriminant validity	Evidence against measuring a single trait by this instrument. Does not cover broad range of the content Not very good reliability and validity evidence Difficulty in scoring But this is the first objective instrument for measuring statistical reasoning abilities
Name: Quantitative Reasoning Quotient (QRQ) **By**: Donna I. Sundre (James Madison University) **Year:** 2003 **Goals:** measuring college students' quantitative reasoning ability. **Target Population/s:** Undergraduate	**Items:** 40 Multiple choice Items on probability and statistics. **Response:** Compared to SRA, all response options from the list of alternatives were considered for correct reasoning or misconceptions **Scoring:** Special scoring method (as SRA) attached to reach response option for each question	**Content Validity:** reviewing items by several panels of faculty with continuous refinements **Reliability:** Low internal Consistency (Ranges from .55 to .62).	Based on the Garfield advice This instrument was developed to address the limitations of SRA Faculty expert panel suggested increasing the number of items in the instrument to cover more content.

Table 2-1. *(Continued)*

Scale Description	Content	Psychometric Properties	Merits, Limitations and Suggestions
	to identify correct reasoning (through 11 subscales) and misconceptions (through 15 subscales).		
Name: Comprehensive Assessment of Outcomes in a First Statistics Course (CAOS) of ARTIST project **By:** Robert Delmas, Joan Garfield, Ann Ooms and Beth Chance **Year:** 2002 **Goal:** Assess students' statistical reasoning/literacy after any first course in statistics **Target Population/s:** Secondary and Tertiary level	**Items:** 40 Multiple choice Items on statistics. **Responses:** one correct or incorrect response for each question. **Scoring:** Conventional scoring methods for MCQ's.	**Content Validity:** Established through adapting items from established tests and extensive expert reviews along with pilot testing items. **Reliability:** Acceptable Internal consistency (Ranges from 0.77 to 0.82) No other reliability (such as alternate form) validity evidence was reported.	Unanimous agreement of 18 Expert statistics Faculty on content validity of finalized CAOS test items for first course in statistics Pre and Post administration of this test did not produce expected gain. And this can be attributed to problem with CAOS items. Still show incompleteness with content coverage Presence of relatively large number of difficult items Limited for first course in statistics. Outdated as this was developed more than decade ago

Table 2.1. *(Continued)*

Scale Description	Content	Psychometric Properties	Merits, Limitations and Suggestions
Name: Assessment Recourse Tool for Improving Statistical Thinking (ARTIST), Topic Scales **By:** Robert Delmas, Joan Garfield, Ann Ooms and Beth Chance **Year:** 2002 **Goal:** Assess students' statistical reasoning/literacy with respect to particular topic areas in introductory statistics course **Target Population/s:** Secondary and Tertiary level	**Items:** 7-15 Multiple choice Items **Responses:** one correct or incorrect response for each question. **Scoring:** Conventional scoring methods for MCQ's.	**Content Validity:** Established through adapting items from established tests and extensive expert reviews along with pilot testing items. Additional Psychometric proprieties for the these individual topic areas were never published	Easily administer (less time) compared to CAOS test Provide support for formative assessment during the introductory statistics course with respect to different topic areas Even though has more literacy items they are limited to definitions and calculations Outdated as this was developed more than decade ago
Name: Goals and Outcomes Associated with Learning Statistics (GOALS) **By:** Joan Garfield, Robert Delmas, Andrew Zieffler, and Dennis Pearl	**Items:** 19 Multiple choice Items **Responses:** one correct or incorrect response for each question. **Scoring:** Conventional scoring methods for MCQ's.	**Content Validity:** Supportive evidence on content validity as items adapted from CAOS test Additional Psychometric proprieties for the these individual topic areas are not available	More suitable with the modern curriculums Especially tailor to courses that contained simulation based simulation based content.

Table 2.1. *(Continued)*

Scale Description	Content	Psychometric Properties	Merits, Limitations and Suggestions
Target Population/s: Tertiary level			Only covers the basic areas of Statistics content that does not address graduate level research
Name: Basic Literacy in Statistics (BLIS) **By:** Andrew Zieffler **Year:** 2014 **Goal:** Assess students' ability to read, understand, and communicate statistical information (statistical literacy) associated with the modern day statistics courses **Target Population/s:** Secondary and Tertiary level	**Items:** 32 Multiple choice Items and 5 test lets **Responses:** one correct or incorrect response for each question. **Scoring:** Conventional scoring methods for MCQ's.	**Content Validity:** Established through adapting items from established tests and extensive expert reviews along with pilot testing items. **Construct validity:** Successful Confirmatory factor analysis results existence of single construct **Reliability:** Very good internal consistency (0.83**)**	More suitable with the modern curriculums Some items of GOALS assessment are included in BLIS Especially tailor to courses that contained simulation based course content. Only covers the basic areas of Statistics content that does not address graduate level research similar to GOALS
Name: Statistics Concept Inventory (SCI) **By:** Andrea Stone **Year:** 2006 **Goal:** assess students' conceptual understanding of	**Items:** 33 Multiple choice Items on statistics. **Responses:** one correct or incorrect response for each question. **Scoring:** Conventional scoring methods for MCQ's.	**Content Validity:** Established through adapting items from text books and extensive review though focus group and independent expert reviews. **Reliability:** Acceptable Internal consistency	Items coming from that address graphical and descriptive data analysis, probability calculations, making inferences from samples and selecting best statistics test and procedures

Table 2.1. *(Continued)*

Scale Description	Content	Psychometric Properties	Merits, Limitations and Suggestions
fundamental statistics introductory statistics course. **Target Population/s:** Undergraduate students (Engineering/Math)		(Ranges from 0.57 to 0.71 for pretests, and 0.58 to 0.86 for posttests)	Considerable portion of SCI items contains more mathematical language oriented question stems and response options. Reliability and validity testing is based on Students' taking courses from Math and Engineering.
Name: Statistical Literacy Survey (SLS) **By:** Milo Schield **Year:** 2002 **Goal:** assess statistical literacy with respect to interpreting parentages and ratio in graphs and tables. **Target Population/s:** College students, College teachers, Professional data analyst.	**Items:** 55 Multiple choice Items (in Statistical Literacy Inventory) on statistics. **Responses:** Correct or incorrect or response (Original response categories in the inventory was Yes, No, Don't Know: These can be classified to Correct or Incorrect) for each question. **Scoring:** Conventional scoring methods for MCQ questions	Current literature review unable to locate any published psychometric proprieties of the instrument	Suggested to remove some items to improve the quality of instrument.

A review of statistic education literature, indicates that most of the previous studies related to assessing students' statistics knowledge are not specifically targeted for the population of graduate students in education and other behavioral science and social sciences (Delmas et al., 2002; Garfield, 1998 (a); Garfield et al., 2012; Schield, 2002; Stone, 2006; Sundre, 2003; Ziegler, 2014). Moreover, these instruments are focused on assessing conceptual knowledge rather than assessing skills of applying statistics to find answers to research questions. And in particular these instruments do not address measurement of knowledge of selecting correct statistical test or procedure. Therefore, the present study focused on developing and validating a new instrument (SAGS) to assess statistics knowledge of graduate students in education and other behavioral science and social sciences, with the careful consideration of the statistical skills that graduate students in these disciplines are supposed know. The SAGS instrument is intended to address several limitations of currently available instruments. Therefore, the SAGS could potentially serve as a basis for further development of interventions and strategies for enhancing graduate students statistics knowledge, statistics self-efficacy, and performance.

Items of the SAGS instrument were based on statistical educational goals and curriculum guidelines for statistics education. Even though such guidelines are available for high school and undergraduate level statistics courses (American Statistical Association, n.d., Bryce et al., 2001; Cannon et al., 2002; Garfield et al., 2000; Moore, 2001; Ritter, Starbuck, & Hogg, 2001; Tarpey, et al., 2000), none have been formally defined for graduate education. Also the curriculum guidelines seem to be different from discipline to discipline (Society for the Teaching of Psychology, n.d.). Thus, using established goals and guidelines to base the item creation becomes infeasible. Alternatively, looking at the main topic areas covered in graduate statistics courses are identified as one solution. However, it is a common fact that the curriculums of

statistics programs are different from one university to another (Bryce et al., 2001). Therefore, graduate level courses falling under education and other behavioral and social science disciplines in universities that belong to South Eastern Conference (SEC) were reviewed to identify the commonly taught statistical topics. Table in Appendix A provides a summary of such courses.

Moreover, to identify the important topics for item creation, commonly used statistical procedures in education and behavioral research were taken into consideration. The review of literature found several studies that focused on identifying commonly used statistical procedures published in articles, doctoral dissertation, and master thesis. These studies are listed in the Table in Appendix B along with commonly used statistical tests and procedures.

Review of the assessment development methodologies shows that the early assessments development efforts in statistics education literature (such as SRA and CAOS) used classical test theory (CTT) approach. But later assessment developed after 2010 such as BLIS utilized Item Response Theory (IRT) approach (Zieffler, 2014). Item Response Theory (IRT) had been experiencing an exponential growth during last few decades and become preferred tool for developing and validating new assents (Clark & Watson, 1995; De Champlain, 2010; Stage, 1998). Specifically, IRT is used to create custom test forms. Such customized tests can be used to accurately measuring the ability level most desirable for educators needs (DeMars, 2010; Hays, Morales, & Reise, 2000; Fliege et al., 2005). Thus to deeply investigate the advantages of using IRT based methods for the current study, rest of the literature review was directed towards identifying salient features of classical test theory and IRT modeling.

Brief Review of Classical Test theory (CTT)

In classical test theory, which is also known as *true score theory,* subject's observed score on the entire instrument is the focus (de Ayala, 2009). Further, subject's latent trait score is

viewed as a function of observed score on a measurement instrument and measurement error (de Ayala, 2009). Thus, CTT models the total score or observed score for subject *s:* Y_s

$$Y_s = T_s + e_s \text{, where } T_s \text{: True score and } e_s \text{: Measurement error}$$

CTT assumes that e_s values are (a) random and not related to one another and (b) not related to true score on the latent variable. Further, in CTT items are assumed to be interchangeable and are not part of the model for creating a latent trait estimate. In Classical test theory the latent trait estimate is the total score, which is problematic when making comparisons across different test forms (Boone, Staver & Yale, 2014; Templin, 2014). In addition, the Item difficulty parameter and Item discrimination parameters defined below is sample dependent.

Item Difficulty. Measure how difficult this item to be answered correctly. Difficulty index is the proportion of subjects who got a particular item correct. Well-tuned test shows item difficulty indices between 0.3 –0.5. Difficulties less than .2 and greater than .8 are considered too hard or too easy, respectively (Aiken, 1997; Ebel & Frisbie 1972).

Item Discrimination. Measures how well an item differentiates between examinees who possess different ability levels. The discrimination index may be calculated by correlating the responses to each item with the total test core on the test. Poorly discriminating items will show a correlation near 0 (Aiken, 1997; Ebel & Frisbie 1972).

In addition to calculating difficulty and discrimination indexes, CTT considers analysis of each of the incorrect answer for items. This is called distractor analysis and item response frequencies are analyzed for the top 25% and bottom 25% of examinees to identify poor distractors (Ebel & Frisbie, 1972; McGahee & Ball, 2009).

Reliability is one of the two important concepts associated with instrument development (Aiken, 1997; Colton & Covert, 2007). Major types of reliabilities used (but not limited to) in

CTT test development are 1) internal consistency reliability 2) test-retest reliability, and 3) alternate from reliability. The internal consistency reliability that will be used in this study is defined as follows:

Internal-Consistency Reliability. This refers to how the degree to which items on the instrument relates to one another. This is usually measured by Cronbach alpha which ranges from 0 to 1, and higher alpha indicate more reliable instrument (Furr & Bacharach, 2014).

These reliabilities measure how consistently an instrument estimates an examinee's score on the latent variable of interest (DeVellis, 2012). Reliabilities are calculated using measurement errors: only one such reliability, which is constant over all trait levels is calculated in CTT.

Validity is the other important concept associated with instrument development. In the context of test development different types of validity are considered: 1) content validity, 2) construct, 3) concurrent validity and 4) convergent validity. Types of validities related to this study are defined as follows:

Content Validity. Content validity asserts how closely the content covered in a test matches the content that should be included in the test (Furr & Bacharach, 2014). This type of validity evidence is normally gathered before administering the test through expert reviews (Aiken, 1997).

Construct Validity. This is the extent to which the items on the test measure the domains it is supposed to measure (Devellis, 2012: Furr & Bacharach, 2014). Construct validity is established through employing multiple procedures such as: 1) experts' judgement, 2) internal consistency**,** 3) studying the relationships, in both experimentally contrived and naturally occurring groups which is known group validity (Aiken, 1997; Devellis, 2012).

Convergent Validity. This is established through examining whether the test scores have high correlation with other measures of the same construct (Aiken, 1997).

According to Aiken (1997), validity of a test is influenced by both measurement error and others systematic (constant) errors. A test may be reliable without being valid. However, a test cannot be valid without being reliable. Thus, when developing a test using the CCT approach examination of both the reliability and validity is essential.

Item Response Theory (IRT) against Classical Test Theory (CTT)

Currently, Item Response Theory (IRT) is becoming increasingly dominant and for many is the preferred approach for test construction and item banking due to its advantages over CTT (De Champlain, 2010; Fan, 1998; Furr & Bacharach, 2013). In CTT approach item statistics such as estimates of item difficulty and item discrimination are sample (examinee) dependent (i.e. the quality of the students in the sample). Also, in CTT analysis, the students observed score (total score) for a particular test form depends on the items in this particular test form (Hambleton & Swaminathan, 1985). These theoretical difficulties of CTT hiders the utility of item parameter estimates in measurements situations such as: 1) Building the larger item inventory with linking new items, 2) Developing testing system into computer adaptive testing system, 3) Identification of biased items, and 4) Various test form equating (Demirtas, 2002; Fan, 1998).

In CTT, the major focus is the test-level information. Thus, the CTT approach addresses the reliability and validity evidence for the calibrated items as one assessment. In the case of developing item inventory or a test item bank, the items are of greater interest. Alternatively, researchers are more focused on examining the characteristics of items and how well these items measure different ability levels. Item response theory has been developed as a platform for examining important features of the items in order to measure the levels of latent.

IRT is a family of associated mathematical models that relates the probability of particular responses on an item to overall examinee ability (Camilli & Shepard, 1994; de Ayala, 2009). In other words, IRT presents a model describing how examinees with different ability levels answer the items. Therefore, in IRT, parameter estimates are not sample dependent (quality of examinees) and examine ability estimates are not test (quality of items) dependent (Embretson & Reise, 2013; Hambleton & Swaminathan, 1985). This is called the invariance property of item parameters; which pertains to the sample-free nature of its results. It also means that groups, as well as individuals, can be tested with a different set of items that are appropriate to their ability levels and the scores will be directly comparable (Anastasi & Urbina, 2002).

Another benefit of IRT is that its treatment of reliability and error of measurement are computed for each item (Lord, 1980). In contrast CTT, reliability and measurement error estimates are obtained for test as a whole. IRT reliability estimates takes attributes of each item into account and shows the measurement efficiency of the item at different ability levels. Thus, researchers are able to identify items best suited to measure particular level of examinee ability and improve the reliability of the test (Templar, 2014). These functions provide a sound basis for choosing items in customized test construction (e.g. to select high ability students to recruit for summer program/award, and to identify students who need additional tutoring).

Review Basics of Item Response Theory

Item Response theory (IRT) is an alternative to classical test theory (CTT) and emerged in early 1950's (Lord, 1953) and was more firmly established in late sixties (Lord & Novick, 1968). Even though it has long history it uses emerged relatively recently as a way to analyze measurement in education and other social and behavioral sciences (Furr & Bacharach, 2013). Currently IRT has become the increasingly dominant and preferred approach for test

construction due to its advantages over CTT and traditional Item analysis (De Champlain, 2010; Fan, 1998; Furr & Bacharach, 2013).

IRT Parameters

Item response theory is based on the fact that person's response to particular test item is influenced by qualities of the individual and by the qualities of the item (de Ayala, 2009; Furr & Bacharach, 2013) In traditional sense IRT cannot considered to be a theory because it is does not explain why person provides particular response to a given item (de Ayala, 2009; Falmagne, 1989). However, IRT is characterized as a psychometric theory which includes a family of associated mathematical models. These models relate individual qualities quantified using latent construct denoted by theta (θ) and item qualities quantified using various item parameters, namely difficulty (b), decimation (a), and pseudo-guessing (c) to the probability of response to items on the assessment. The Item characteristic curve (ICC) is this primary IRT concepts and illustrated in Figure 2.1 to represent a general ICC for one multiple choice test item (with one correct and multiple incorrect answers) intended to measure particular ability of interest.

Moreover, the following section explanations various parameters in these IRT models (Baker, 2001; Barlow, 2014; de Ayala, 2009; DeMars, 2010; Furr & Bacharach, 2013).

Latent Trait Distribution (θ). The spread of a quantified latent trait an instrument intended to measure. These trait levels are measured on an interval along the horizontal axis with a mean of 0.0 and standard deviation of 1.0 comparable to a z-score distribution. For example, an individual with an average trait level would be located at $\theta = 0.0$, and the majority of individuals θ will fall between -3.0 to 3.0.

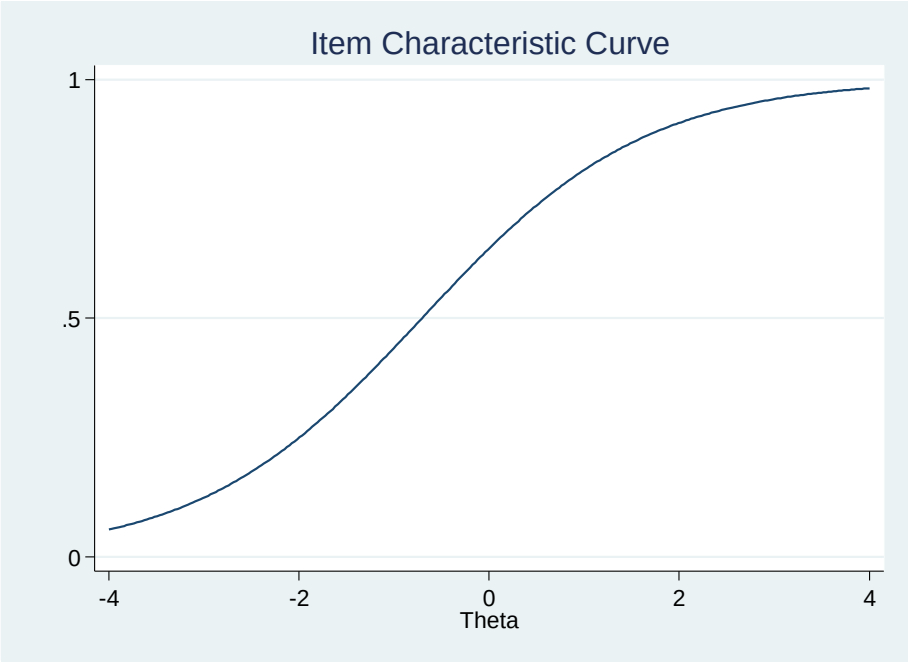


Figure 2.1. Example: Item Characteristic Curve (ICC)

Item Difficulty (b). The location on the latent trait continuum (θ) where a person has a 0.5 probability or 50% chance of giving the correct answer. The higher the"b" parameter, respondent need have higher level of latent trait is needed in order to correctly answer the item. Like θ the item difficulties parameters (b's) will fall between -3.0 to 3.0 and this is expressed in terms of trait level.

Item Discrimination (a). The slope of the ICC line at the location of item difficulty. The value of the discrimination parameter indicates the relevance of the item to the trait measured by the test. Further it indicates how strongly an item is related to the latent trait and is comparable to a factor analysis loading. High discriminating items are better at discriminating respondent where their trait level is closer to item difficulty.

Item Pseudo-Guessing (c): Represented as the lower-asymptote on an ICC, which is, "The value the function approaches as θ approaches negative infinity" (DeMars, 2010, p. 13). Inclusion of Pseudo-guessing (c) parameter suggest that respondents with very low trait level (low θ) will answer an item correctly given chance alone (Barlow, 2014, p.8)

A variety of models have been developed from the IRT perspective for the purpose of developing and validating instruments. These models differ from each other in terms of 1) item characteristics or *number parameters* and 2) response option format. The following section describes only the basic IRT models designed to be used with dichotomous items such as correct or incorrect items in multiple choice exams.

Assumptions of IRT

Item response theory has two primary assumptions: 1) the test data must contain single dimension (unidimensionality), and 2) the data must be locally independent (DeMars, 2010, Waller et al., 2013). Unidimensionality refers to the need that only one latent trait, θ, is

measured by the items on the test (DeMars, 2010, Boone, et al., 2014). Even though this assumption is key to IRT modeling, the literature reveled that there is some degree of violation of this assumption in most of the given testing situations (de Ayala, 2009, DeMars, 2010). Hays and colleagues suggested the concepts of "essential unidimensionality" where satisfying the assumption at acceptable level. Thus, items in a test are considered to be unidimensional when a single factor or trait accounts for a substantial portion of the total test score variance (Templin, 2014). However, the serious violation of this assumption could bias several item and ability parameter estimates (Templin, 2014).

The second assumption, the local independence, assumes that any two items will be unrelated to one another after controlling for θ (DeMars, 2010, Boone et al., 2014). Moreover, once you know the subject's θ level his or her responses to the items are independent of one another. If one trait determines success on each item, then the subject's latent trait value θ is the only thing that systematically affects item performances. Thus, local independence is closely related to unidimensionality (Barlow, 2014; DeMars, 2010, Templin, 2014). Local independence leads to statistically independent probabilities for item responses.

$$P(Y_{is} = 1, Y_{i's} = 1 \mid \theta_s) = P(Y_{is} = 1 \mid \theta_s) P(Y_{i's} = 1 \mid \theta_s),$$

Here, *s* denotes examinee *s*, and *i* and *i'* denotes items *i* and item *i'* respectively. This formulation is the basis for the construction of the likelihood function which is important for the estimation of Item and ability parameters in IRT.

Three Common IRT Models for Dichotomous Data

Three most commonly used IRT modes for dichotomous data, in order of mathematical complexity are: 1) Rasch model or 1-Parameter Logistic (1 PL) model), 2) 2- Parameter Logistics (2 PL) model, and 3) 3- Parameter Logistic Model (3 PL).

One-Parameter Logistic Model and Rasch Model. The Rasch model is the simplest IRT model. According to this model specification response to a dichotomous item is determined by individual's trait level and only a single item characteristic or parameter that is the item's difficulty. According to the model, the probability that the individual will correctly respond to the items depends on the subject's trait level (θ_s) and the items difficulty (b_i). These trait levels and difficulties are standardized so that their means are 0 and their standard deviations are 1. This model is represented as (Emberston & Reise, 2000):

$$P(X_{is} = 1 \mid \theta_s, b_i) = \frac{e^{a_i(\theta_s - b_i)}}{1 + e^{a_i(\theta_s - b_i)}}$$

Where, P refers to conditional probability that the individual will correctly respond to the items depends on the subject's trait level (θ_s) and the items difficulty (b_i).

X_{is} refers to a particular response X made by individual *s* to item *i*, $X_{is} = 1$ refers to correct response.

θ_s refers to the trait level of subject *s*.

b_i refers to the difficulty of item *i*.

a refers to the discrimination of item *i*, (*a=1* in Rasch model).

e is the base of natural logarithm.

Due to its simplicity 1 PL model requires estimation of only one parameter for each item and one common discrimination parameter. Estimation of 1 PL model requires marginally lower sample size than either 2PL or 3PL models. Further citing Lord (1983), de Ayala Lord (2009) highlights that the Rasch model provides more stable parameter estimates than 2PL and 3PL models with the sample sizes less than 200.

Two-Parameter Logistic Model (2 PL model). The 2 PL model has two parameters. Addition to the difficulty parameter in Rasch model this includes varying discriminant parameters. In this models discrimination parameters varies among items. According to this model response to dichotomous item is determined by individual's trait level (θ_s), the item's difficulty (b_i), and the item discrimination (a_i).This model is represented as:

$$P(X_{is} = 1 \mid \theta_s, b_i, a_i) = \frac{e^{(a_i(\theta_s - b_i))}}{1 + e^{(a_i(\theta_s - b_i))}}$$

Where a_i refers to the difficulty of item *i*.

Considering the trade-off between sample size and accuracy of the parameter estimates 2 PL model is considered to be superior model than Rasch,1 PL or 3 PL model and require between 200 and 500 subjects to fit the model (Drasgow, 1989; Yen, 1981; Stone, 1992 as cited in de Ayala).

Three-Parameter Logistic Model (3 PL model). The 3 PL model has three parameters. Addition to the difficulty and discrimination parameter in 2 PL model this includes the pseudo-guessing parameter. In this model pseudo-guessing varies among items.

According to this model response to dichotomous item is determined by individual's trait level (θ_s), the item's difficulty (b_i), the item discrimination (a_i) and pseudo-guessing (c_i) .This model is represented as:

$$P(X_{is} = 1 \mid \theta_s, c_i, b_i, a_i) = c_i + (1 - c_i)\frac{e^{(1.7a_i(\theta_s - b_i))}}{1 + e^{(1.7a_i(\theta_s - b_i))}}$$

Where c_i refers to the pseudo-guessing of item *i*.

An advantage of 3 PL model is that it takes into account non-random guessing that typically occurs in multiple choice exams with low-ability examines (DeMars, 2010). As the

most complex model compared to 2 PL and Rasch models, 3 PL requires substantially larger sample to efficiently estimate parameters. According to de Ayala (2009) to obtain stable estimates pseudo-guessing parameters needs sample of at least size 1000. When quality distractors are used to minimize guesting the 2 PL model is preferred over the 3 PL model (DeMars, 2010).

Goodness of Fit of IRT Models

Prior to estimating the parameter in an IRT approach observed data should be tested against the different IRT models. According to Templin (2014) there is no one best way to assess fit in IRT models. Techniques typically used are classified into following categories.

Absolute fit/Global Fit. Use of this measure allows model-based hypothesis test. Chi-squared test falls under this category but this is only for small number of items (10-15). Relative entropy measures also fall under this category but this is hard to interpret in some occasions.

Relative fit. Relative fit of nested models is evaluated using chi-squared (deviance) test. Here non-significant change in log-likelihood is tested. Also, a portfolio of models is evaluated using information theoretic model selection criteria such as Akaike Information Criteria (AIC) and Schwartz's Bayesian Criteria (BIC/SBC) which are used to test the relative model fit. The model with lowest values is identified as the best fitting model.

Item Fit. This compares the model predicted and observed frequencies of all items in the test marginally (univariate fit) and evaluates using the chi-squared statistic. This is not very useful in IRT as most items fit. Another type of item fit, bivariate fit compares the model predicted and observed frequencies of responses using chi-squared statistic for all pairs of items in the test.

Reliability in IRT

Recall that in CTT, one reliability coefficient is calculated for the test as a whole, and this reliability value cannot be decomposed to a value that is attributed to an individual examinee (Stage 1998). From the IRT perspective, reliability is measured in the form of item information and test information. These measures indicate how much a researcher can be certain of a person's location along the latent trait θ.

Item Information Function (IIF). For each item, the amount of information is proportionate to the standard error of estimate (SEE) for each possible θ location (de Ayala, 2009). A smaller SEE indicates a stronger certainty in the estimate of θ and therefore provides more information about individuals with that particular θ value. According to theory, an item provides its highest amount of information near its difficulty value ("b") because there is the least amount of variability (error) near this value (DeMars, 2010). Similarly, an item with a high discrimination value will provide a large amount of information over a short range of θ whereas the flatter line of a poorly discriminating item will provide less information over a lengthier range of θ values (DeMars, 2010).

Any item in a test provides some information about the ability of the examinee, but the amount of this information depends on how closely the difficulty of the item matches the ability of the person. In the case of the 1PL model this is the only factor affecting item information, while in other models it combines with other factors. To further examine the nature of item information functions, two items item characteristic curve are given in the Figure 2.2 and corresponding item information functions are given in Figure 2.3 to observe salient features. Item q1 has a steeper slope (discrimination) than item q6. Thus, q1 shows greater item information than q6.

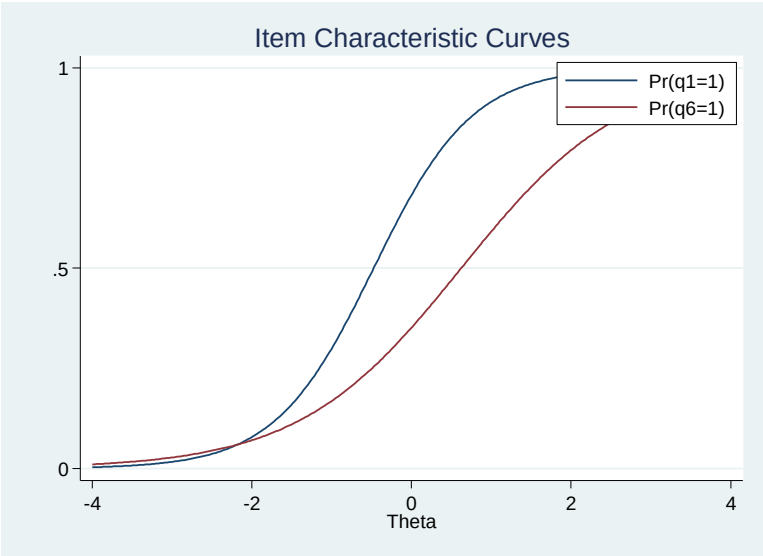


Figure 2.2. Example: Item Characteristic Curve (ICC) for Two Items

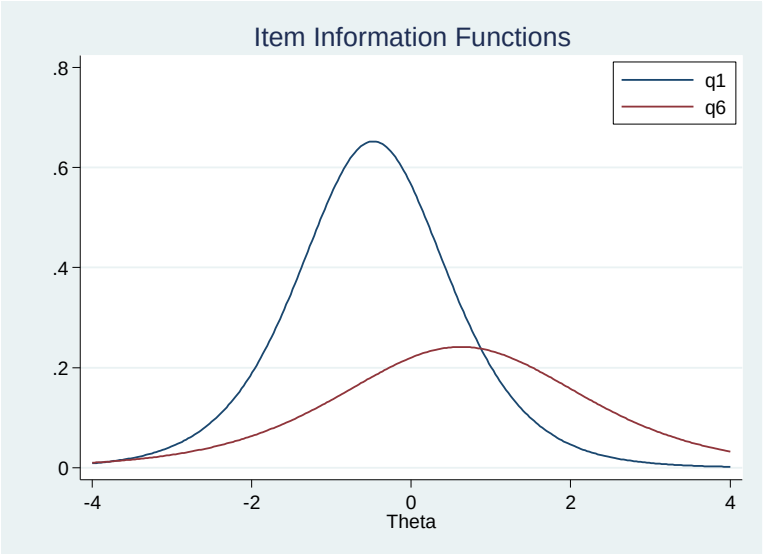


Figure 2.3. Example: Item Information Functions (IFF)

Test Information Function (TIF). Item information functions are summed to obtain a test information function (TIF). The TIF plot tells us how accurately the instrument can estimate person locations (Templin, 2014). Test information can be easily recalculated according to the items chosen for a particular test/form and this leads to test being having stronger properties for some examines and weaker for some other examinees (Furr & Bacharach, 2008). As an example, test geared towards novice students may be designed to be more accurate at determining differences among lower ability levels, so specific items can be chosen which provide the most information in such lower ability range (Barlow, 2014). Sample test information function is illustrated in Figure 2.4.

The most important thing about the test information function is that it predicts the accuracy to which we can measure any value of the latent ability. Even though person ability cannot observe directly, using test information functions we could obtain an estimate of what level of accuracy of a particular test expects to achieve at any ability level.

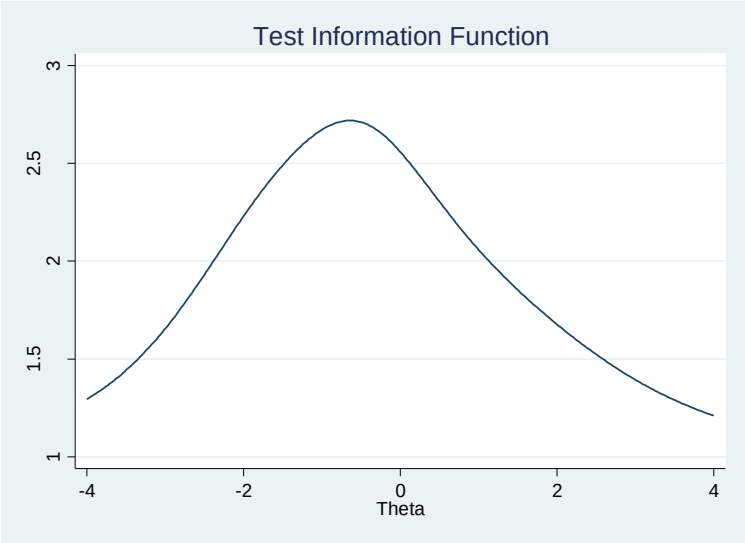


Figure 2.4. Example: Test Information Function (TIF)

Rasch Modeling as a Subset of IRT: Mathematically

Three IRT models described earlier in this chapter has close mathematical relationship as lower level models are nested within higher level models. The 2PL model is nested within 3PL model which is expressed as:

$$P(X_{is} = 1 \mid \theta_s, c_i, b_i, a_i) = c_i + (1 - c_i)\frac{e^{(1.7a_i(\theta_s - b_i))}}{1 + e^{(1.7a_i(\theta_s - b_i))}},$$

as it can be obtained by setting guessing parameter $c_i = 0$, for all *i*. The 1 PL model is nested within 2 PL and 3 PL models, as it can be obtained by setting the discrimination parameter $a_i = a$, a common value. The Rasch model is a special case of 1 PL model where $a_i = a = 1$ (Brown, Templin, & Cohen, 2014). Thus, Rasch model is the simplest of item response theory (IRT) model having the minimum of parameters for the person (just one), and just one parameter corresponding to each item (in case of dichotomous items). Additionally, the Rash model converts the item difficulty parameters and the examinee ability parameters to the same logit scale.

Differences in IRT and Rasch Modeling: Philosophically

Item response theory models are designed to imitate data, so they are data driven. If data do not fit the model perhaps a better model is needed. But, the Rasch model is derived to define a measurement but is not designed to fit any data. Thus, the Rash model is theory driven, defined a priori, and if data does not fit there is a need to get better data (Brown, Templin, & Cohen, 2014; Wright, 1992). Thus the Rasch model is taken as a criterion, specification or statement for the structure of the responses, rather than a mere statistical description of the responses.

According to Wright (1992) maintaining the additivity and linearity is an important concept in measurement construction and these concepts can be loosely defined as the condition

of one more unit is always to be the same amount (equal interval property). The Rash model formulation is a sufficient and necessary condition for construction of linear and objective measures. Raw scores have unknown spacing between them but Rasch analysis estimates true intervals of item difficulty and person ability by creating linear measures. Also, in Rasch model the total score summarizes completely a person's standing on a variable, arises from a more fundamental measurement requirement of comparison of two people who took different forms of tests (Boone, et al., 2014; Wright, 1992).

Compared to lose specifications in higher order IRT models Rasch model has tight specification. Rasch model does not account for guessing parameters for items as guessing is considered to be dependent on person characteristics. Also, different discrimination parameters for each item (crossed item characteristic curves) give rise to a measurement system that changes for different levels of ability similar to meter-stick ordering different size objects differently when measured in two occasions. Use of such measurement systems have problems when consistently defining the construct being measured, thus in Rasch as a measurement strategy does not accept varying discrimination. According to Rash philosophy items should not have different discrimination if the researcher knows the construct (Templin, 2014).

Sample Size Requirements for Rasch Modeling

According to Linacre (1994) minimum sample to give useful item calibrations is important question when conducting Rasch analysis. In measurement studies item calibrations are expected to be similar enough to maintain a useful level of measurement stability. Similar enough is defined in such a way that items have difficulty stable with in certain logit values difference under certain level of confidence. Citing Lee (1992) Linacre (2016) states that if item calibration is stable within a 1 logit, it is considered to be targeted at a correct grade level. In

addition, measures based on item calibration with random deviations up to 0.5 logit are considered to be free from any bias even if the test is short (less than 30items). Sample size requirement for Rasch analysis is summarized in Table 2. 2.

Table 2.2. Sample Size Requirements for Rasch Modeling.

Item Calibration Stable within	Confidence	Minimum Sample size range	Size for most Purposes
+/- 1 logit	95%	16 - 36	30
+/- .5 logit	95%	64 - 144	100
+/- .5 logit	99%	108 -243	150

Here the lower end sample size value is given when the sample is targeted for items with a 40%-60% success rate. Higher end value is given when the sample obtains success rates more extreme than 15% or 85% (Wright & Stone, 1979). In addition, there should be at least 8 correct responses and 8 incorrect responses are needed for reasonable confidence that an item calibration is stable within 1 logit. In case of un-modeled measurement disturbance, such as different testing conditions or alternative curricula it is advised to inflate these sample sizes by 10%-40%.

Important Pieces of Rash Analysis

Person Measures/ Item measures. Rasch model assumes that the probability of a given person/item interaction (in terms of rating high or low) is only governed by the difficulty of the item and the ability of the person, that are determined by the item locations on the presumed

latent variable along with the rating scale structure. The mathematical formulation of the Rasch model is:

$$P(X_{is} = 1 \mid \theta_s, b_i) = \frac{e^{(\theta_s - b_i)}}{1 + e^{(\theta_s - b_i)}} \quad \text{or} \quad \theta_s - b_i = \ln\left(\frac{P_{is}}{1 - P_{is}}\right).$$

Where b_i is the difficulty parameter of item i

and

θ_S, is the ability parameter of perosns

Additionally, Rasch model puts persons and items on the same scale with the equal-interval property.

Person Measures. Person measures are the estimated ability of the person which is represented by θ_s.

Item Measures. Item measures are the estimated difficulties of test items which is represented by b_i.

Point-Measure Correlation. A point measure correlation indicates the degree to which the item is aligned with the abilities of persons. According to Rasch philosophy the higher person measure implies higher rating on items (correct scored as 1 against incorrect scored as 0) and higher item measure implies higher rating on persons. Point measure correlation reports the degree to which this relationship is true. Medium to large positive correlations are acceptable and negative and close to zero correlations are problematic (Linacre, 2016). Small positive correlations may need further investigations. Rasch analysis reports PT-Measure: CORR (PTMACORR) as the observed correlation and PT- Measure: EXP as the expected correlation. When the data fit the Rasch model, these values will be the same. When PTMACORR is greater than EXP, the item is over-discriminating between high and low performers. When

PTMACORR is less than EXP, the item is under-discriminating between high and low performers. When EXP is near to zero, then the item is very easy or very hard. It is off-target to the person distribution (Boone at el., 2014; Linacre, n.d.; Linacre, 2016).

Fit Evaluation Using Local Fit Statistics. Fit describes how well the data agree with the Rash model. INFIT and OUTFIT statistics (local fit) are the most widely used diagnostic Rasch fit statistics and these are obtained for both items and persons.

Item Fit. Item fit can be better explained using "misfitting" items. A "misfitting" item would be an item that would be correctly answered by low performing students (Not all but some). Similar misfit would be an easy item that would be incorrectly answered by high performing students, and this type of misfit is called item outfit. Outfit statistic is a chi-square statistic which represents the association between data and Rasch model, specifically how well data fit the model. OUTFIT statistics is generally sensitive to outliers and more diagnostic when item measures are far from the person measures. INFIT is another fit statistics and it is more diagnostic when item measures are close to the person measures (Linacre, n.d; Linacre, 2016).

According to Boone at al. (2014), in Rasch analysis several fit indices are provided for evaluating Item fit and Person fit, these are: Item outfit MNSQ, Item Outfit ZSTD, Person Outfit MNSQ, Person Outfit ZSTD, Item Infit MNSQ, Item Infit ZSTD, Person Infit MNSQ, and person Infit ZSTD. Linacre (n.d) advised fit evaluation should be started with Item outfit since identifying outliers is helpful to identify and correct the issues with misfit. MNSQ is a chi-square statistic (which measures level of association). First Outfit and Infit statistics with the fit MNSQ values are to be investigated. Values range between 0.7- 1.3 are reasonable for MCQ's, while the general acceptability of items MNSQ should from 0.5-1.5 range to create productive measurement. If the MNSQ fall outside this range, it is advice to look for respective ZSTD

value. ZSTD is a *t test* statistic measuring the probability of extreme MNSQ occurring by chance. Values of ZSTD ranged between -2.0 and 2.0 are acceptable and it will justify the data have reasonable predictability. In case of fit analysis (investigating MMSQ and ZSTD) show which items misfit such items are to be individually explored to identify the misfitting person response. Next, these person responses will be deleted and new fit statistics should be obtained for the second round of Item fit evaluation. On the other hand, the items which do not fit the Rasch model indicate multidimensionality and need for modification or discarding of the item. Also misfit is an indicator that the construct theory needs revising. The items that fit Rasch model are likely to be measuring the single dimension intended by the construct theory (Baghaei, 2008).

Wright Map, Item-Person Map, or Construct Map. When data fit Rasch model, person measures and items measures can be computed with higher level of confidence and both these measures are expressed in same equal-interval scale. Item –person map also known as wright map or construct map is a graphical presentation of person measures and items measures which is useful to evaluate the validity of the measurement. In item-person map, items are ordered from lowest difficulty to highest difficulty on one side of plot (middle line, expressed in logit scale) and persons are ordered from lowest ability to highest ability. In a good measurement system there should be some items with different difficulty levels which are consistent with full ability range of the respondent sample. As an example, there should be difficult items to accurately measure high performing students and there should be easy items that should accurately measure low performing students. Existence of difficult items (top of the Item-person map) without persons aligned is and alarm bell indicating such items is too difficult and not useful to construct good measures. Item-person map is used to evaluate the consistency of items

and person and this concept is called item targeting. In addition, Item-person map is used to establish the construct validity and predictive validity of the measure. Example of Item-Person map is given in Figure 2.5.

In item-person map, items are ordered from lowest difficulty to highest difficulty on one side of plot (middle line, expressed in logit scale) and persons are ordered from lowest ability to highest ability. In a good measurement system there should be some items with different difficulty levels which are consistent with full ability range of the respondent sample. As an example, there should be difficult items to accurately measure high performing students and there should be easy items that should accurately measure low performing students.

Existence of difficult items (top of the Item-person map) without persons aligned is problematics indicating such items are too difficult and not useful to construct good measures. Item-person map is used to evaluate the consistency of items and person and this concept is called item targeting. In addition, Item-person map is used to establish the construct validity and predictive validity of the measure.

Linacre (2016) provides guidelines to explain Item-Person maps. Left-hand column locates the person ability measures along the variable (latent construct being measured) and the person's measures often have a normal distribution. Right-hand column pinpoints the item difficulty measures along the latent construct. Items arranged by measure and hierarchy of item description should indicate a meaningful construct from easiest at the bottom to hardest at the top. For dichotomous items, an even spread of items along the variable (the y-axis) with no gaps is preferred. Gaps in the map indicates poorly defined or poorly tested regions of the construct. A better constructed tests usually have the items targeted (lined up with) the persons.

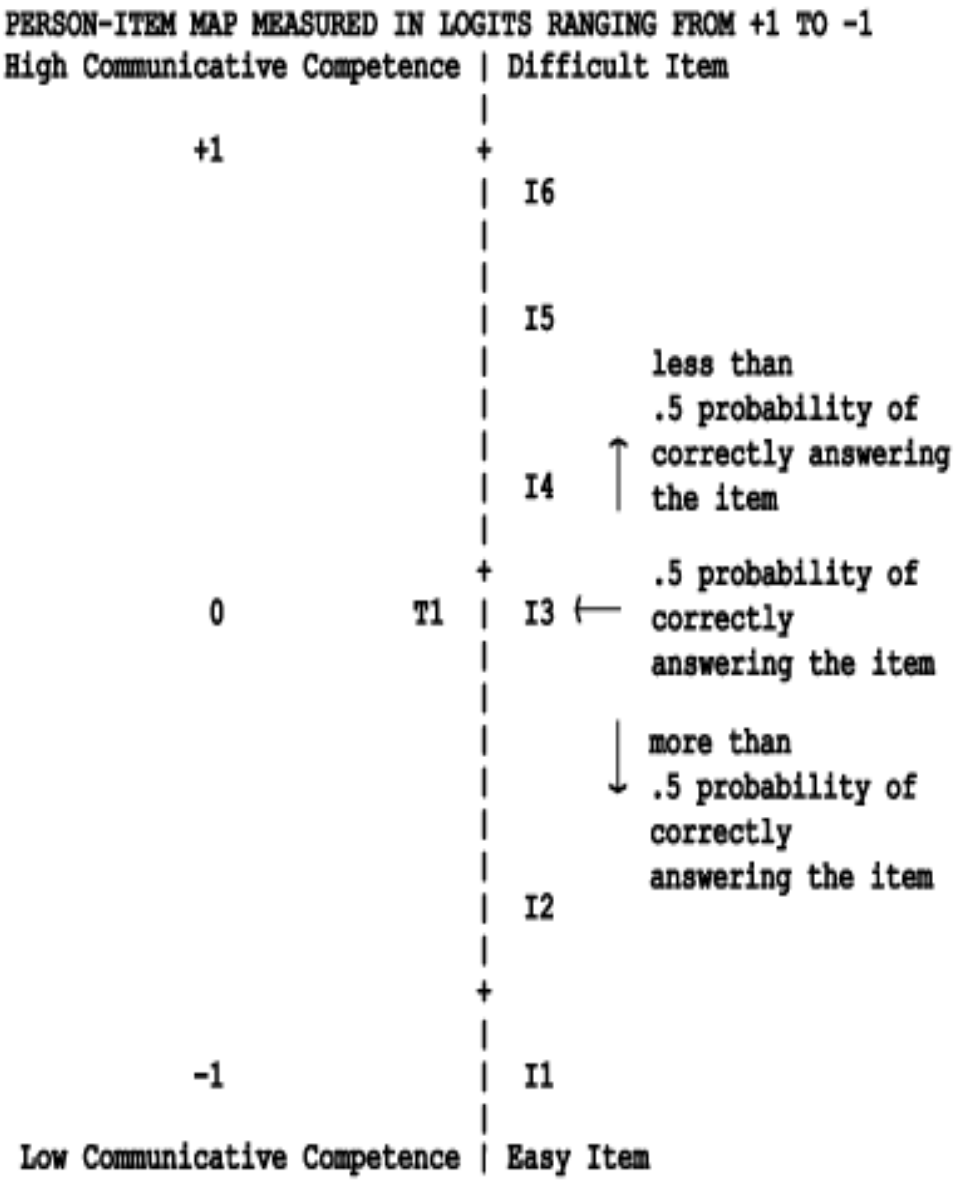


Figure 2.5. Item-peon Map of a Latent Trait (Construct)

Distractor Analysis. Rasch approach distractor analysis was conducted using the original item response data (ungraded). Poor performing item distractors are identified using point bi-serial correlation between the data code (0 or 1) and the examinees Rasch scores. Ideally, almost correct distractor has somewhat positive point bi-serial correlation (PTMACORR) values, mostly wrong distractor has zero or negative value, completely wrong distractor has highly negative value, in addition to the correct response has positive point bi-serial correlation values (Linacre, 2016).

Construct Validity in Rasch. Construct validity is established by observing the item hierarchy in Item-person map. Items should be ordered as would be expected based on what the researcher intends to measure (theory). Known group comparisons are also used to justify the construct validity. Means scores of group that are believed to have a higher level of the construct is compared with mean scores of groups that are believed to have a lower level of the construct (Barlow, 2014; Bone et al., 2014, Linacre, 2016)

Predictive Validity in Rasch. Predictive validity is established by observing person ordering according the construct and other background information about them. Such as investigating whether the experienced respondents (measured by demographic variable) have higher person measures. Alternatively, person measures and other information are correlated to establish predicted validity numerically. Predictive validity is also tested using known group comparisons (Barlow, 2014; Bone et al., 2014, Linacre, 2016).

Reliability analysis in Rasch. Rasch reliability is consistent with the usual the general concept of reproducibility. Rash analysis reports reliability and separation indexes for items and persons and those have different applications and implications (Boone et al. 2014, Linacre, n.d.; Linacre, 2016).

Reliability Indexes. Person reliability index in Rasch analysis corresponds to conventional reliability. Low values point out a narrow range of person measures, or a small number of items. Values of .5 indicates that respondents sample can discriminate only into 1 or 2 levels while values of 0.8 and 0.9 indicates sample can be discriminated into 2 or 3 levels and 3 or 4 levels respectively (Boone et al., 2014).

Item reliability index has no equivalent statistics in classical test theory. Low values indicate a narrow range of item measures, or a smaller sample size. In general, low item reliability indicates the need of gathering data from a larger sample to obtain more stable item parameter estimates.

Separation Indexes. Item separation is used to confirm the item hierarchy. Low item separation (< 3 = high, medium, low item difficulties, item reliability < 0.9) suggests that the person sample is not large enough to confirm the item difficulty hierarchy (construct validity) of the measure (Bonne et al. 2014: Linacre, 2016). However, Duncan, et al., (2003) suggest that item separation index of 1.50 is acceptable when no respondent groupings is considered for analysis.

Person separation is used to classify people, and this represents a measure of ratio between true person variance and error variance. Low person separation (< 2) and person reliability (< 0.8) with an appropriate person sample indicates that the instrument may not be not sensitive enough to differentiate between high and low ability examinees. Thus, more items may be needed. However, Duncan, et al., (2003) states that person separation index of 1.50 as an acceptable level of separation.

Chapter Two Summary

When it comes to graduate degrees in behavioral and social sciences, students are required to take statistics courses as statistics is essential tool to successfully complete their research component. However, for these students statistics is considered as one of the most anxiety providing subjects during their course of studies. Even though students able to apply statistics to common type research problems taught in the classes, they lack the ability to select appropriate statistical procedure to answer their own research questions by modeling it to new research problems. Moreover, previous research in statistics education consistently concludes that there is a generally low and variable level of statistics knowledgebase among students who complete their undergraduate degrees. Statistics educators who teach graduate level have to deal with students who have diverse statistics knowledge. Thus, by measuring each students statistics skills in terms of ability to select appropriate statistical procedure will help them to identify most effective approaches to teach their graduate statistics class.

After statistics education emerged as new filed in late nineties number of attempts were made to develop assessments to measures students' statistics knowledge and skills. Almost all these instruments were intended to measure major constructs, statistical reasoning, statistical thinking and statistical literacy of undergraduate and high school students. However, there is no validated assessment that is directly targeting graduate students and measuring their statistical test/procedure selection ability. Review of meta-analysis revealed that student completing graduate degrees have used set of common statistical procedures and review of graduate statistics curriculums revealed that departments offer statistics course that teach these common procedures. Thus the items in the new assessment will include these procedures and tests. The next chapter describes the methodology for development of new SAGS instrument.

CHAPTER THREE

MATERIALS AND METHODS

The sections of this chapter include review of the problem, study purpose and objectives, as well as the research design of the present study. It also addresses the population of interest and sample of students that data were collected from, as well as instrument development within the study, measures that were used, study procedures, and data analysis methods.

Review of the Problem

Chapter two reviewed the status of statistics education research and fundamental problems on which the current study is focused. The essential use of quantitative methods in Education and other Behavioral and Social science (EBS) research have placed a premium importance on educating graduate students in these disciplines to transle empirical evidence to answer research question (Gilmore & Feldon, 2010; Meerah, et al., 2012). Knowledge of statistics is an essential component to comprehending much empirical evidence (Devore, 2015; Healey, 2014; Harris & Jarvis, 2014), but several studies from past few decades have shown a consistently low, variable knowledge and higher level of anxiety in statistics among graduate students in EBS disciplines (Grácio & Garrutti, 2006; Hannigan, et al., 2014; Onwuegbuzie & Wilson, 2003; Onwuegbuzie et al., 1997). Also, reviews of literature have shown a steady increase in the frequency of use and complexity of statistical methods over the last few decades and potential for them to grow in the future (Jackman, 2009; Kline & Santor, 1999; Vance, 2015; Wright & London, 2009).

Identifying appropriate statistical methods or techniques to analyze a given research problem is essential for students completing graduate level research (Alacaci, 2012). Thus, graduate level statistics courses are designed to train students on developing such selection skills while providing conceptual understanding (Eastern Illinois University, n.d.; The University of

Tennessee, Knoxville, n.d.). Students may take statistics courses applying procedures to familiar or well-known problems, yet they can be greatly challenged to apply statistics to their own research (Dunn et al., 2012; Marusteri & Bacarea, 2010). Thus, measuring students' ability to select appropriate statistical procedure (which could be defined as statistical research methodology knowledge) is crucial to enhance learning and teaching process in graduate classroom, and assessments are one of the tool used for this purpose (Earl, 2012; Delmas et al., 2007; Gathercole, Pickering, Knight, & Stegmann, 2004; Gold, & Harris, 2013). However, the current literature does not reveal any assessments that measure such ability of gradute students in ESB (Delmas et al.,2004; Grafield et al.,2012; Grafield, 1998a; Stone, et al., 2003; Sundre, 2003; Zeigler 2014).

Study Purpose and Objectives

The purpose of the proposed study was to develop an instrument; Statistics Assessment of Graduate Students (SAGS) that measure students' statistical research methodology knowledge through establishing preliminary item characteristics and validity evidence. In addition, the proposed study investigated the efficacy of Rasch modeling and Item Response Theory (IRT) to develop a novel statistical research methodology knowledge assessment for graduate students in EBS disciplines.

There were four research objectives guiding this study:

1. Establish content validity evidence of the SAGS instrument
2. Examine the model fit of the SAGS items to a Rasch model
    a. Test the assumptions of unidimensionality and local independence.
    b. Identify item difficulties and analyze the item information/test information of the SAGS instrument.

c. Analyze the quality of item distractors of the SAGS instrument.

3. Examine the reliability and validity of the SAGS instrument

    a. Assess the reliability of the SAGS instrument through the analysis of various reliability and separation indices.

    b. Assess construct, predictive, and other types of validities of the SAGS instrument through construct maps and known group comparisons.

4. Examine the model fit of the SAGS items to 1 PL, 2PL and 3PL IRT models based on simulated data.

    a. Investigate the performance of novel information complexity criteria *(ICOMP)* over other model selection criteria for determining the best fitting IRT model.

    b. Identify item difficulty, discrimination and guessing parameters.

    c. Compare person ability and item location estimates (difficulty, discrimination, and guessing) from IRT models to those of traditional Classical Test Theory (CTT) indices.

The following section of this chapter outlines the *Research Design* including the study population of interest as well as the sampling procedures that were used to select participants. Next, the *Instrument development* section will illustrate the process by which the SAGS assessment instrument was developed including the process used for initial item generation. This section also provides the detailed description about the measures associated with the study. The remaining portion of the chapter is allocated to the study *Procedure*. First, a description of the data collection procedures is presented and it will include details on selecting participants for the

various stages of study. Finally, the statistical methods that will be used to analyze data are explained according to each research objective.

Research Design

This study used a cross sectional survey design, as it facilitates collecting data from a fairly large number of individuals (Colton & Covert, 2007) and a simulation study. Data were collected through an online instrument, which included items to assess students' ability to select statistical procedures along with relevant participant demographic items. Since the assessment is designed to collect data from a fairly large number of students, the students were asked to answer multiple choice questions/items with four response options. In addition to demographic variables such as field of study/major, gender, and year of study the assessment also consisted of other relevant variables such as number of statistics courses respondents have taken at the graduate level, undergraduate level, and high school level. Also, perception about their own statistics knowledge and frequency of using statistics was collected, and these data were used in support of efforts to assess validity.

Study Population and Sample

The SAGS instrument was developed to target graduate students in education and behavioral and social science disciplines. Thus, the population of interest was graduate students who were enrolled in graduate level quantitative course in departments that fall under ESB disciplines at universities in United Sates. To meet Rasch analysis requirement, sample size ranged from 64 to 144 was proposed for the study (Linacre, 2016). Subjects for the study were recruited through purposive sampling approaches. The instructors who teach graduate level quantitative courses at the University of Tennessee, Knoxville and others universities were contacted initially through personal communications and then a using snowball approach

(Morrow, 2013). Instructors who agreed to support data collection were requested to forward a flyer, which introduced and described this study, offering a link to the online SAGS instrument to students in their current and previous classes. To reach additional participants, additional announcements of the study were posted on various websites, social media pages, and academic discussion boards. Also, the study announcement was sent through several listserv mostly subscribed by graduate students in education and other behavioral and social sciences.

Instrument Development

Items for the assessment were developed through a four stage process. During the first stage, initial items were constructed (Phase 1 of the study) by reviewing the contents of graduate statistics courses (Appendix A) offered in education department in universities who are members of southeastern conference universities ("SEC Universities", n.d.). Also, the most commonly used statistical procedures in doctoral dissertations and master's thesis (as presented in refereed meta-analysis articles) in behavioral and social science disciplines were reviewed to identify relevant content areas that should be included in the instrument (Appendix B). A "Brainstorming" session was conducted with group of peer students to generate list of ideas (Morrow, 2013; Wikipedia, n.d) for developing the initial set of SAGS items. Thirty-five questions (items) with multiple choice answer options were developed initially, covering the application level of the Bloom's Taxonomy (Bloom, 1956; Krathwohl, 2002). Also, items were developed to represent the wider range of difficulty level starting form descriptive statistics items, followed by ANOVA or mean comparison type items, and then on to Regression items, Multivariate items, and ending with higher level statistical modeling items such as Structural Equations Modeling items and Multilevel Modeling items. One correct answer and three Item

distractors per each question were carefully written to capture the difference in mastery level of the working knowledge with the particular statistical concept or procedure.

During the second stage (Phase 2 of the study), a focus group meeting with graduate students (who have taken upper level statistics classes) was conducted to initially review the items. Focus group participants consisted of students coming from various departments including Educational Psychology, Theory and Practice of Teacher Education, Educational Leadership and Policy Studies and Social Work. Furthermore, some of these students have completed undergraduate and master's degrees in psychology. Focus group feedback was used to modify the items stems as well as to change the item distractors. In stage three, an expert review was conducted (Phase 3 of the study) to establish the content validity and to further improve the quality of the SAGS instrument (Morrow, 2013). Expert review panel which included content area experts (5 faculty members who teach graduate level statistics classes for students in EBS and one statistician at academic consulting center) was asked to review the SAGS items individually and provide suggestions regarding the face and content validity of the drafted instrument. Moreover, in both stages reviewers of the items were asked to provide:

- feedback on the clarity of instructions, the item stems, and the quality of the item distractors.
- suggestions regarding any potential new items that should be added, existing items that should be deleted or revised, or issues regarding spelling and/or grammar, and appropriateness of the ordering of the items.
- other comments that would help improve the instrument, and
- estimates on the approximate time that their students would take to completely answer the SAGS instrument.

Next, using these expert panel comments and suggestions, the instrument was finalized for the pilot administration. Instrument was then informally pilot tested with peer group of five students. Responses from the pilot test were analyzed to make necessary modification before the final administration (Phase 4 of the study).

Measures

The instrumentation developed through this study included two types of measurement items: a) cognitive items that measure ability to select appropriate statistical test or procedure, and b) demographic items that were used to describe the sample and conduct validity analyses. To collect data for the study, the two measures were combined together as one instrument and administered to students in an on-line (using "Qualtrics" survey management system) setting.

Cognitive Items for SAGS Instrument. The cognitive items in the assessment ask students to select the correct or best answer to given statistics problems. Questions/items of this assessment were developed with multiple choice answers providing *four* response options (Appendix C). The instrument seeks to measure the underlying ability of respondents to select an appropriate statistical procedure (defined as statistical research methodology knowledge) to address specific research situation common to graduate students.

Demographic Items for SAGS Instrument. A set of demographic items was also developed to gather information about individual students involved in the study. These demographic items were used to identify the attributes of the students and student groups in support of different types of validity evidence. Furthermore, demographic information was used to describe the sample and to assess the representativeness of the sample. These variables were also used to compare students' basic statistics knowledge across demographic groups during the post assessment development stage. The gathered students' information such as their gender,

major/ field of study, year of study, level of exposure to statistics at graduate level, undergraduate level, and high school level, and their perception on their own statistics knowledge and usage of statistics (Appendix C).

Procedure

First, all the required study materials were submitted to the University of Tennessee Institutional Review Board (IRB) to receive institutional approval. After getting the IRB approval, instructors who teach graduate level statistics classes (upper level) in departments that falls EBS disciplines at the university of Tennessee, Knoxville were identified by reviewing the university website. Next, students for the focus group (phase 2 of the study) were recruited by requesting these instructors to forward an announcement about the focus group to their class participants (Appendix D). Once students responded with their willingness to participate in the focus group, the received a detailed explanation of the study and scheduling information. (Appendix E). Prior to the focus group, each graduate student participant was asked to give consent to participate in the study (Appendix F). the focus group protocol was created to direct the discussion on the review of item stems, distractors and instructions to participants (Appendix G). The instrument was modified using focus group comments and then sent to the IRB approval prior to expert review (Phase 3 of the study).

An invitation letter (Appendix H) was send to the expert faculty and statistical consultants to participate in the SAGS expert review. Informed contest (Appendix I) and review materials (Appendix J) were send to the experts who agreed to participate in the study. Expert panel members were asked to conduct the expert review independently and return the review document electronically to the author. The SAGS instrument was further modified based on the experts' comments and informally pilot tested with five of peer graduate students. Using pilot

test data, the instrument was finalized for administration in the phase 4 of the study. Another IRB form was submitted for approval for administering the modified instrument to the target population.

Subsequently, data for the SAGS validation study was collected from a purposive sample of students from the target population of graduate students studying in ESB disciplines. Instructors who teach graduate level statistics courses at the University of Tennessee, Knoxville (UTK) were contacted through personnel communications. Further, they were asked to identify other potential instructors (within and outside UTK) to contact regarding participation in the study. All instructors in the study were requested (Appendix K) to assist with the collection of completed instrument from their students by posting the study flyer (Appendix L) in the blackboard site or sending the flyer to student's e-mail. Also, they were asked to make an announcement about this study in the class, and request their students to complete SAGS instrument. Announcement of the study was send to the moderators of various website, listserv, and social media pages (Appendix M and Appendix N) with the objective of gathering data from participants outside university of Tennessee, Knoxville.

To ensure the safety and privacy of the participants, ethical guidelines were followed as outlined in publication titled Ethical Principles of Psychologists and Code of Conduct (American Psychological Association, 2010). An informed consent (Appendix O) was given to the participating students to provide information on the study and to obtain their agreement to participate in the study. Data were collected assuring confidentiality, and only the principal investigator has the access to the assessment data. Completed assessment data was retrieved from the "Qualtrics" system as "SPSS" data file and stored on a password protected computer. Data were made available only to the principal investigator and his dissertation advisor. No references

were made in oral or written reports which could link participants of the study (collected in Focus group, Exert review, and Final Assessment administration) to their responses.

Data Analysis

Once the data for the study were collected, a number of data cleaning procedures were used as described by Morrow and Skolits (2015) to prepare the data for subsequent analysis. First, the data were cleaned for outliers. Percentage of missing values for demographic variables including previous numbers of statistics courses taken at graduate and undergraduate college level as well as exposure to statistics at high school level were also examined prior to the analysis.

Assess the Rasch/IRT Assumptions: Unidimensionality and Local Independence

The cleaned data were first utilized to test the essential unidimensionality and local independence assumption required for the Rasch and Item-Response modeling (Ayala, 2010; DeMars, 2010; Stewart-Brownet al., 2009; Yu, Popp, DiGangi, & Jannasch-Pennell, 2007). The data of observed responses for the 25 cognitive items were graded using "SPSS" program to create binary responses (correct and incorrect). For this binary item response data, a confirmatory factor analysis was performed using tetrachoric correlation (Baglin, 2014; Cook, Kallen, & Amtmann, 2009; Deng et al., 2008; Holgado-Tello, Chacón–Moscoso, Barbero–García, & Vila–Abad, 2010; Uebersax, 2006; Zieffler, 2014). Then the established criteria for absolute and relative model fit indexes were considered (such as Chi-square statistic, GFI, PGFI, REMSA, AIC, BIC) in order to test for one factor solution (Bryne, 2012; Tabachnick, & Fidell, 2013). The one factor solution indicated only a single latent trait is influencing item responses and, thereby, local independence was assumed for the subsequent analysis (DeMars, 2010).

Evaluating Rasch Model Fit

Under the presence of above the mentioned essential assumptions, the data set was exported to “Winsteps” software (Linacre, 2002) for the purpose of fitting a Rash model. According to the Rasch theory (Bond & Fox, 2013; Boone et al., 2014; Wright & Stone, 1979) the model was first evaluated based on the traditional Chi-square goodness of fit statistics to establish global fit of the model. Further, values of the local fit statistics in terms of INFIT and OUTFIT in “WINSTEPS” Rasch terminology were simultaneously examined to identify poorly performing items and respondents (persons in Rasch terminology) that were inconsistent with the theory underlying the Rasch model (Boone at al., 2014; Linacre, 2016). Several individual observations were set to missing values as some items demonstrated misfitting behavior. The modified data set was imported into the “WINSTEPS” program for re-evaluation of Rasch model fit. Following non-significant Chi-square statistic which indicates good fit of the Rasch model, and acceptable local fit statistics, item difficulty parameters and person location parameters were estimated. Furthermore, distractor analysis was performed under both Rash and classical test theory approaches.

Estimating Parameters and Information Function of SAGS Items

Item difficulty estimates were obtained using “WINSTEPS” program and estimates were based on joint maximal likelihood estimation. Further, Rasch graphical analysis using Wright/Construct map, which displays the item difficulty parameters and person ability parameters simultaneously mapped into logit scale (Boone, et al., 2014; Linacre, 2012) was used to identify discrepancy of the items used in the assessment with the ability level of the study participants. All Rash analysis estimates was then used to make decisions about deletion of the existing items or adding new items to the SAGS item bank to make a better instrument. Item

information function for the SAGS instrument based on Rasch model was constructed and reviewed to identify the ability levels that SAGS items measure the person abilities most accurately (Baker, 2001; Baker & Kim, 2004; Partchev, 2004; van der Linden & Hambleton, 2013). Next, these item information functions were summed to examine the total test information provided for 25 item SAGS assessment. Test information function was reviewed to identify ability ranges where complete SAGS instrument give most accurate measures.

Establishing Reliability and Validity Evidence

Persisting with the Rash analysis framework, reliability of the SAGS instrument was evaluated using item reliability, person reliability, item separation, and person separation indices as well as internal consistency reliability measured through Cronbach's alpha (Boone et al., 2014; Lee-Ellis, 2009). Also, three types of validity evidence were examined during this study. First, the construct validity of the instrument was examined using the Wright map and the estimated Rash item difficulty hierarchy (Baghaei, 2008; Boone et al., 2014; Bradley et al., 2010; Conaghan, Emerton, & Tennant, 2007; Linacare, 2016). If the observed item difficulty hierarchy is compatible/makes sense with the known difficulty level (identified through the literature) of the items, construct validity of the instrument is justified. An alternative measure for each person's ability collected through demographic variables and self-reported statistics skills were used to conduct known group compressions to further establish construct and predictive validity (Barlow, 2014; de Ayala, 2009; Linacre, 2016).

Examining the Model Fit of the SAGS Items to Higher Order IRT Models

Utilizing the sample data obtained, larger data sets were simulated using re-sampling techniques (Efron, & Tibshirani, 1994; Yu, 2003) to facilitate the objective of evaluating whether the 1 parameter logistic (1 PL), 2 parameters logistic (2PL), or 3 parameter logistic (3PL) models

performed better than Rasch model for describing the responses of initial SAGS administration data. Furthermore, this simulation study was performed based on mean vectors and covariance matrices to keep the attributes of original data in the simulated larger data sets with varying sample sizes (100, 200, 500, and 1000). Relative performance of the three models (1 Pl, 2PL, and 3PL over Rasch) was evaluated using several model comparison indices (Brown et al., 2015; Kang & Cohen, 2007), i.e., namely change in negative log-likelihood, Akaike Information Criteria (AIC), Consistent Akaike Information Criteria (CAIC), Schwartz's Bayesian Criteria (SBC), Bozdogan's Consistent Alike Information Criteria (CAIC), and Consistant Information Complexity Criteria (CICOMP). Based on the minimum values obtained for the most number of above mentioned model selection criteria's, best fitting model was selected (Bozdogan, 1994; Bozdogan, 2010; Howe, Bozdogan, & Katragadda, 2011). For the best fitting IRT model item difficulty, discrimination and guessing parameters along with the ability estimate for each individual were estimated. Finally, these parameters were compared with the parameters obtained from the classical test theory approach using Pearson correlations.

Chapter Three Summary

Chapter three includes comprehensive description of methods for developing the SAGS instrument. In summary, this cross-section study and preliminary simulation study is guided by four primary objectives: (1) establish content validity evidence of the SAGS instrument; (2) examine the model fit of the SAGS items to a Rasch, model; (3) gather preliminary reliability and validity evidence for the SAGS instrument and 4) examine the model fit of the SAGS items to 1 PL, 2PL and 3PL IRT models based on simulated data through a summation based study. SAGS items were written in a multiple choice with one-best answer format, initially constructed items were revised using the comments made by focus group that involves graduate students.

Further review of the items was performed using six-person expert panel which included statistic teaching faculty and statistics consultants. Content validity of the instrument was also established through the expert review while making further improvements to the instrument based on their comments. After informal pilot testing the instrument, final modification was made to the instrument. Then, on-line SAGS instrument was administered to the graduate students in education and other behavioral and social science disciplines. After collecting the data, preliminary analysis was conducted with the objective of data cleaning and to assess statistical assumptions. Rasch model was then fitted to the dataset to observe the global fit, misfitting items were identified using local fit statistics. Observations causing item misfit were identified and removed from the data prior to the final Rasch model evolution. Given good global and local fit indices final Rasch parameters were estimated. Item distractors are also assessed using both Rasch and CTT approaches. Reliability of SAGS instrument was established through looking at various reliability and separation indices. Construct validity evidence was established through examining construct maps and known group comparisons. Simulation study was carried out to examine the performance of 1 PL, 2PL and 3PL IRT model to SAGS items which will provide additional information about functioning of SAGS items. Next chapter will present the detailed results of the analysis related to the development SAGS instruments mentioned in this chapter.

CHAPTER FOUR

RESULTS

This chapter presents results of the data collection and analysis processes carried out according to the design and procedures introduced in Chapter Three. The chapter begins by describing the procedures used to clean the data prior to quantitative analyses. Similar to the previous chapters, data analyses and associated results will be presented and organized by study objectives.

Data Cleaning

Data of respondents' who completed all cognitive questions were extracted from the "Qualtrics" survey management system for the analysis. The final dataset used for the analysis consisted of 132 observations. At the time of the analysis there were additional 41 partially competed responses available in the "Qualtrics" system, but those were not used for the current analysis. Data were then cleaned in the manner described by Morrow and Skolits (2014), and assessed for the specific assumptions required for Confirmatory Factor Analysis (CFA) and Rasch/Item Response Theory modeling. First, frequency analyses were conducted on each of the variables consisting data for 25 cognitive questions and 13 demographic questions to examine for any coding errors. Participants' responses to open-ended questions about their major area of study and study concentration were cleaned for spelling mistakes.

Next, the data were examined variable-wise for missing values. Since the cognitive questions were set in a forced choice format within "Qualtrics" system, there were no missing data for variables representing these questions. However, some demographic variables (quantitative) had missing values, but the amount of missing data was less than 4% per variable. Five percent or fewer amounts of random missing values in a relatively large dataset is not

considered a serious deficiency as list-wise deletion of observations could be used to analyze the data in the presence of lower degree of missing values (Roth, 1994; Tabachnick, & Fidell, 2013).

The cognitive variables were then cleaned for outliers. Since 99.9% of scores in a standard normal distribution fall between -3.29 and +3.29, according to Tabachnick and Fidell (2013) any score that is outside of that range can be considered as an outlier. Since none of the cognitive variables had z-scores outside the range -3.29 to 3.29, no modification was done to the dataset.

As the final step univariate normality of the cognitive and the demographic variables were assessed by looking at skewness and kurtosis values. Skewness and kurtosis values were less than |2| on all variables, therefore univariate normality was assumed (Tabachnick & Fidell, 2013).

Responses for 25 cognitive questions recorded in the original "SPSS" datasheet retrieved from "Qualtrics" system and cleaned as mentioned in this section was scored using the recode function in "SPSS" software. The correct answer option for a particular question was recoded to value "1" while the 3 distractors were recoded to value "0".

Initial Analysis and Participant Characteristics

Descriptive statistics reflect that on average participants got slightly more than 15 questions correct (M = 15.08, SD = 5.25). Out of 132 respondents, two respondents got all 25 questions correct, while the minimum score earned was 4 out of 25 which was obtained by two respondents. The largest group of respondents got correct answers for 20 questions while second largest and third largest groups got 11 and 12 questions correct. Additional summary statistics of the SAGS instrument are given in Table 4.1.

Table 4.1. Descriptive Statistics of SAGS Instrument.

Summary Statistics	Total Score	Percentages
Mean	15.08	60.30
Median	15	60.00
Mode	20	80.00
Standard Deviation	5.25	21.00
Range	21	84.00
Minimum	4	16.00
Maximum	25	100.00
Skewness	-0.08	-0.08
Kurtosis	-1.02	-1.02
25th Percentile	11	44.00
75th Percentile	20	80.00

Note. Statistics for 132 individuals and 25 items are presented.

Frequency distribution of time that the participants spend competing the assessment are given in Table 4.2. Most of the participants in the sample (54, 40.91%) took between 15 to 30 minutes to complete the assessment. Noticeably there were few participants who competed the assessment by taking more than 90 minutes which influenced on the mean time to complete the survey (mean time to complete the assessment was 77 minutes with standard deviation of 6 minutes). Median completion time was 26 minutes, and this fall under the approximated completion time of 40 minutes when deigning the instrument. In addition, the 5% trimmed mean completion time was 35 minutes.

Table 4.2. Frequency Distribution of Completion Time.

Time	Frequency	Percentage
0 -15 minutes	21	15.91
15 – 30 minutes	54	40.91
30 - 45 minutes	23	17.42
45 – 60 minutes	15	11.36
60 -90 minutes	6	4.55
More than 90 minutes	13	9.85

Descriptive statistics of demographic variables are given in Table 4.3. Participants in the sample were mostly education majors (106, 83.46%). The majority of participants were doctoral students (101, 76.52%), and female (83, 62.88%). Most of the participants were in their second year of study (30, 22.73%), and the remaining participants were distributed as, first year (22, 16.67%), third year (20, 15.15%), and fourth year (29, 21.97%).

Table 4.3. Background Characteristics of Participants and Group Specific Total Scores.

Demographic Variable	Frequency	SAGS Performance
Program		
Education	106 (83.46%)	15.46 (5.39)
Other	21 (16.54%)	14.24 (4.45)
Graduate Level		
Doctoral	101 (76.5%)	16.01 (4.97)
Masters	24 (18.18%)	11.67 (4.95)
Year[a]		
One	22 (16.67%)	11.50 (5.23)
Two	30 (22.73%)	13.90 (4.96)
Three	20 (15.15%)	15.90 (5.68)
Four	29 (21.97%)	17.24 (4.57)
Other	24 (18.18%)	16.75 (4.32)
Gender[b]		
Male	46 (34.85%)	15.59 (5.77)
Female	83 (62.88%)	14.94 (4.94)

Note. [a,b]Seven additional participant for the Graduate Level and Year variables selected "prefer not to answer", three participants selected "prefer not to answer" for Gender variable.

Most of the participants (128, 96.97%) had taken at least one statistics course at the graduate or undergraduate level. Also, the majority (115, 87.12%) had completed a research methodology course. At the time of data collection, a majority (112, 87.50%) of the participants were not taking any statistics class. Participants' exposure to statistics is depicted in Table 4.4.

Rasch and IRT Modeling

Rasch modeling and item response theory is based on a demanding set of assumptions. If the assumptions are not satisfied, the usefulness or accuracy of Rasch and IRT estimates are severely misleading (de Ayala, 2009; Templin, 2014). Thus, the first step in examining the model fit of the SAGS items is the assessment of the data for violations of unidimensionality and local independence assumptions (Boone et al. 2014; de Ayala, 2009; Linacre, 2016).

Test for Violations of Essential Unidimensionality and Local Independence

Dimensionality of item response data is defined as the minimum number of latent traits necessary to achieve LI (local independence), thus unidimensionality refers to dominance of one latent trait on the item responses (Abswoude, van der Ark, & Sijtsma, 2004). Considering the notion that the strictness of the unidimensionality assumption is oftentimes challenging to meet in real-life data, the essential unidimensionality (EU) approach was proposed as an alternative method to analyze item response data (Barlow, 2014; de Ayala, 2009; Zhang & Stout, 1999). The EU procedure identifies the number of dimensions within item response data. Subsequently the test is independently analyzed as a smaller "testlet" attributed to a number of identified dimensions rather than using a complex multidimensional model. As the SAGS items were carefully developed according to Rasch philosophy representing one dimension (Linacre, 2016), the strict unidimensionality (presence of only one dimension) assumption was initially tested as opposed to the EU.

Table 4.4. Statistics Exposure and Group Specific Total Scores.

Demographic Variable	Frequency (Valid %)	Exam Performance
Taken previous Stat Courses		
No courses before	4 (3.03%)	7.50 (3.42)
graduate or undergraduate	124 (96.97%)	15.31 (5.13)
Research methodology		
Took a course	115 (87.12%)	15.30 (5.22)
Not taken any course	17 (12.88%)	13.53 (5.32)
Currently		
Taking a course	16 (12.50%)	13.94 (5.47)
Not taking a course	112 (87.50%)	15.51 (5.07)
Time completed last course		
1 semester ago	38 (29.69%)	16.61 (5.20)
Within one year	24 (18.75%)	14.50 (5.21)
Within 1-2 years	30 (23.44%)	16.90 (4.51)
More than 2 years	36 (28.13%)	13.17 (4.78)

Confirmatory Factor Analysis (CFA) was used to test the unidimensionality assumption through examining the presence of one latent trait (Cook, Kallen, & Amtmann, 2009; Deng et al., 2008; Zieffler, 2014). As the graded cognitive items responses were binary, CFA was conducted using Tetrachoric correlation matrix (Seaver, 2013; Ubersax, 2006). Smaller number of response categories in observed variables (which measure a latent construct) cause weakened correlation among the considered variables (Bonett & Price, 2005, Ubersax, 2006). Thus, in this analysis, Tetrachoric correlations were used to counteract the problem of underestimated correlations (Seaver, 2013; Ubersax, 2006). Utilizing the PROC CALIS procedure in the "SAS" statistical package, a hypothetical one factor model represented in Figure 4.1 was specified by formulating hypothetical equations (Seaver, 2013). In the model, cognitive items responses were predicted by one and only one factor and respective error terms. The distribution free Unweighted Least Square (ULS) estimation method was used to evaluate the model and obtain estimated values for the model parameters and fit indices (Suhr, 2006).

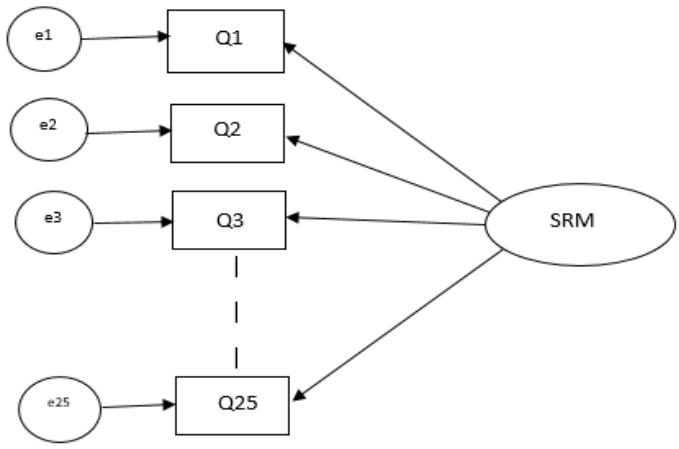


Figure 4.1. One Factor Model (Prior to Estimation) for Assessing Unidimensionality
Note. "SRM' represent Statistical Research Methodology Knowledge, "e" represent error term

While the chi-square goodness of fit statistic was not observable for the ULS estimation method, other available absolute fit indices were examined. The goodness of fit index (GFI) value was 0.90 and thus reached the threshold of 0.90 (Byrne, 2010), while the Parsimonious Goodness of fit index (PGF1) value which is theoretically smaller than GFI was .82 and also demonstrated an acceptable level of model fit (Byrne, 2010). Even though the values of goodness of fit indices showed evidence to support the hypothesized one factor model, caution should be made about this result as factor analysis is a large sample technique (Byrne, 2010; Newsome, 2012; Tabachnick & Fidell, 2013).

Two factor and three factor solutions were also examined to further justify the one factor solution. Items were randomly selected to form two clusters. Next, CFA was conducted by loading these items onto two factors (Abswoude et al., 2004; Deng et al., 2008). This procedure was repeated five times to obtain a more reliable solution. There was no considerable improvement of fit indices where the average GFI value for the 5 repeated runs of CFA's remained at 0.90 and PGFI was 0.83. Similarly, analyses were conducted for the three factors and the average values for the GFI and PGFI were 0.91 and 0.83 respectively. Thus, for further analysis a one factor solution was selected and unidimensionality of item responses was assumed. Unidimensionality of the item responses also justified the satisfaction of local independence assumption (DeMars, 2010).

Evaluating Rasch Model Fit

Following the preliminary analysis and validation of the assumptions, Rasch analysis was conducted to monitor the data quality in the context of Rasch measurement requirements. The Pont Measure Correlation (PTMACORR) ranged from 0.22 to 0.66 and no value was close to zero or there were no negative values which indicated that the items responses were not

contradictory (Linacre, 2016) with the latent construct being measured (participants who have process higher level of latent construct knowledge get higher scores and participants who have process lower level of latent construct knowledge get lower scores). As the items measure the intended construct, fit analysis of the Rasch model was conducted prior to obtaining the item parameter and person ability estimates.

Multiple fit statistics provided in "WINSTEPS" software (Linacre, 2016) was utilized to evaluate the model fit for the study sample. First, Rasch model fit was assessed using the chi-square global fit statistic (Bone at al., 2014; Linacre, 2016). The test statistic value was not statistically significant, implying that the observed data were well fitted to the Rasch model with log-likelihood $\chi^2(3187)=3166.07$, $p=0.55$. Secondly, local fit analyses were conducted for each individual items and respondents. Analysis of Item Outfit identified four items (Q2, Q4, Q18, and Q25) as having Mean Square (MSNQ) statistics values above 1.5, which is not recommended by Wright and Linacre (1994) to construct a productive measurement.

Considering the MCQ format of the SAGS assessment, further investigation was carried out by looking at MNSQ values along with Z-Standardized (ZSTD) values. MNSQ Values outside 0.7 and 1.3 and ZSTD values outside the range -2 to +2 indicates that the model tends to under-fit the data (Boone et al., 2014; Linacre, 2016). Further Item Infit MNSQ and ZSTD indicated there was one item (Q12) flagged as an item that shows an indication of over fit. Detailed observation of 132 individuals' items response Z-residuals (>2) for misfitting items indicated that 16 out of 3300 individual question responses had provided idiosyncratic answers to these items. Thus, a second Rasch analysis was conducted after removing these responses which brought the item misfit statistic to an acceptable level for 24 items except for Q4. According to Wright et al. (1994), the parameter for Q4 was still usable with the Rasch model

definition to create productive measurement. Subsequently, person fit statistics were observed and no misfitting persons were found.

SAGS Rasch Statistics: Item Difficulty Estimates (b)

Once the Rasch model was chosen, all Rasch estimates for item parameters could be assessed. The difficulty estimates for each item are shown in Table 4.5, ordered according to most difficult item to easiest item. The value of "b" (difficulty parameter in Rasch) can be directly compared to the proportion correct ("P" column) to show how items located at a higher level of ability (i.e. a higher "b") translated to a smaller proportion of correct responses (Barlow, 2014; de Ayala, 2009). Overall, for the observed sample, "b" values ranged from -1.59 to 2.47 (M = 0.00, SD = 1.05). Difficulty parameters of a majority (14) of the items were less than 0 (equivalent to P>0.5 in CTT) while the other 11 items had difficulty parameters greater than 0 indicating majority of participants answered these items incorrectly. Out of 132 respondents, most respondents (111, 84.09%) correctly answered Q12 (an easy item) while the least number of respondents (26, 20.63%) correctly answered Q2 (most difficult item).

SAGS Rasch Statistics: Person Ability Estimates (θ)

Person location or ability estimate "θ" in Rasch modeling corresponds to the total score in CTT. The observed sample had two individuals who answered all the questions correctly and Rasch analysis identifies these two along one other participant, who had a raw score of 24 as extreme scores. Overall, person ability estimates (in logit scale) ranged from -1.97 to 3.76 (M = 0.51, SD = 1.27) for 129 non-extreme respondents. The respective average raw score was 14.70 (SD=5.30) for participants identified as non-extreme.

Table 4.5. Classical Test Theory and Rasch Approach Difficulty Parameter Estimates.

Item	CTT Estimate (P)	Difficulty	Rasch Estimate (b)
Q2	0.20	High	2.47
Q18	0.23	High	2.29
Q25	0.29	High	1.84
Q4	0.39	Medium	1.15
Q17	0.48	Medium	0.64
Q23	0.49	Medium	0.60
Q16	0.50	Medium	0.55
Q5	0.51	Medium	0.51
Q10	0.54	Medium	0.31
Q14	0.54	Medium	0.31
Q24	0.57	Medium	0.18
Q8	0.63	Medium	-0.16
Q22	0.66	Medium	-0.33
Q19	0.67	Medium	-0.38
Q1	0.68	Medium	-0.47
Q11	0.71	Low	-0.65
Q15	0.73	Low	-0.75
Q9	0.74	Low	-0.80
Q21	0.74	Low	-0.80
Q6	0.75	Low	-0.90
Q7	0.76	Low	-0.95
Q13	0.76	Low	-0.95
Q3	0.77	Low	-1.00
Q20	0.78	Low	-1.11
Q12	0.84	Low	-1.59

Variable/Construct/Wright Map

One of the strengths of Rasch/IRT modeling over CTT is that both item difficulty and respondent abilities are expressed on the same scale, so the functioning of the items can be explored thoroughly. Wright map in Figure 4.2 is a graphical presentation of common scaled item and person estimates.

Items are presented from difficult, at the top of the map in Figure 4.2, to easy, at bottom of the Wright map. Items near each other are those items that define the construct in a similar manner. Ordering of the items at the top of the map matches the ordering (difficulty level) conceptualized by the researcher at the item development stage. In contrast, two items at the most bottom positions of the map were not consistent with the researcher's belief about easier items in the instrument. Higher difficult items were the items testing knowledge of Coefficient of Variation (Q2), Canonical Correlations (Q18) and Log-linear analysis (Q22). Somewhat surprisingly, the easiest items (at the bottom of the map) were Repeated Measures ANOVA (Q12), followed by Confirmatory Factor Analysis (Q20). However, the next set of easier items were Multiple Linear Regression (Q13), Measures of Variation (Q3) and Factorial ANOVA (Q7). There were nine participants (indicated by numerals of the left portion of the map) above the highest ability level measured by the most difficult items, and six participants were below the ability level measured by easiest item. Average Rasch measure for the group of participants was slightly higher than the average Rasch measure for the set of cognitive items comprising the SAGS instrument.

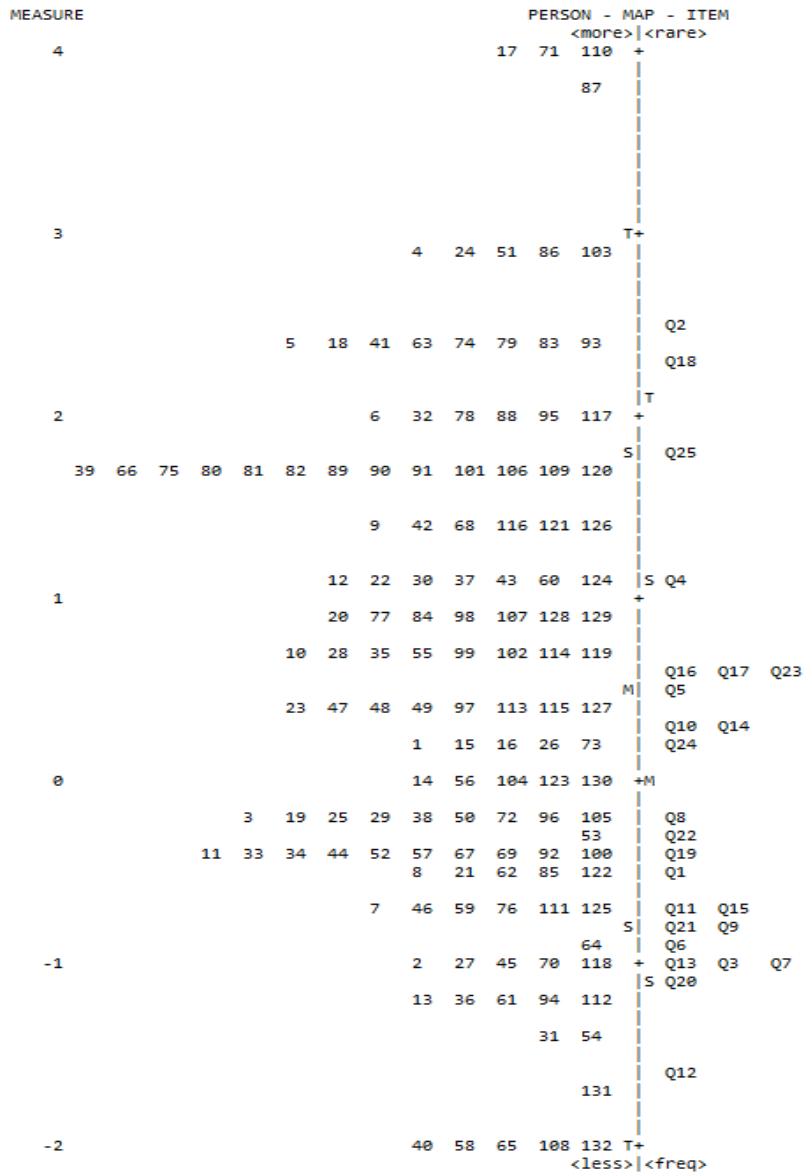


Figure 4.2. Distribution of Items and Persons on the Common Scale
Note. Q1: Central tendency measures, Q2: Coefficient of variation, Q3: Measures of variation, Q4: Correlations, Q5: One sample t-test, Q6: Dependent sample t-test, Q7: Factorial between subjects ANOVA, Q8: ANCOVA, Q9: t-test for independent samples, Q10: Chi-square test, Q11: One-way ANOVA, Q12: Repeated measures ANOVA, Q13: Multiple linear regression, Q14: Multinomial logistic regression, Q15: ANOVA with interactions, Q16: One-way between subjects MANOVA, Q17: Discriminant analysis, Q18: Canonical correlations, Q19: Exploratory factor analysis, Q20: Confirmatory factor analysis, Q21: Cluster Analysis, Q22: Structural equation modeling, Q23: Multilevel modeling, Q24: MANCOVA, and Q25: Log-linear analysis.

SAGS Overall Test Performance

The test characteristics curve (TCC) for the SAGS instrument is represented in Figure 4.3. The TCC was constructed by aggregating all the ICC's, and this particular version of TCC showed the relationship between the respondents' estimated ability level and the expected raw score. The plot indicated that, to perfectly answer SAGS instrument questions, respondents needed to have higher level of ability more than +5 logit of the latent variable (conceptualized as statistical research methodology knowledge).

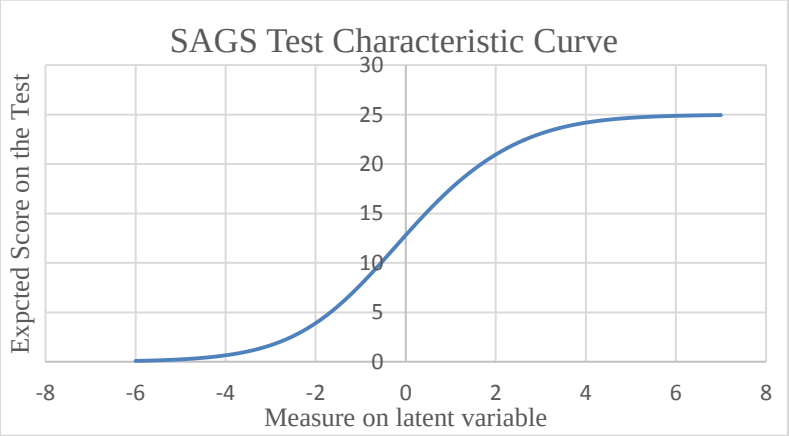


Figure 4.3. Test Characteristic Curve of the SAGS Assessment

Distractor Analysis

Rasch approach distractor analysis was conducted using the original items response data (ungraded). Another Rasch analysis was performed by inputting the test answer key into the "WINSTEPS" control file. The "WINSTEPS" item distractor table was examined to identify ill performing item distractors using point-biserial correlation values (PTMACORR values) (Linacre, 2016). Further, the PTMACORR value represents the point bi-serial correlation between of persons' estimated ability (Rasch measure) and response for each answer option (yes

or no) in a particular item. Given only one correct answer for each question, existence of positive PTMACORR value for correct option and negative values for distractors were examined. Negative PTMACORR values were not observed for SAGS distractors of some items. Response option 4 of Q4 and Q5, and option 2 of Q18 had positive PTMACORR values. Rasch analysis suggested distractor options of these items to be revised. Observed PTAMACORR values and ordering of each response options is given in Table 4.6.

Being more critical about each distractor, ordering of the correlation values was compared with the estimated participant's ability to endorse each distractor. Rasch modeling expects that least difficult distractor to have higher negative PTMACORR value, and PTMACORR value should be increased toward zero (0) or small positive value as the distractors become relatively more correct (Linacre, 2016). Such perfect relationships were observed for majority of SAGS items, but some items and their distractors did not show this relationship. Response option 3 of Q1, option 1 and 2 of Q2, option 1 of Q7, option 1 of Q9, Option 4 of Q13, Option 2 of Q14, Option 2 of Q18, Option 4 of Q16, options 2 of Q20, option 4 of Q21, option 2 of Q22, option 1 of Q23 showed a disrupted ordering. Thus these distractors may be considered for revision to create a near perfectly functioning assessment.

Using a more conventional approach, a distractor analysis was conducted by comparing the answering patterns of the participants belonging to top and bottom 25% ability groups created based total-correct score. Top and bottom 25% groups were created based on the total score they obtained for answering 25 cognitive questions.

Table 4.6. Correlation of Response Options and Persons' Ability Level.

Item	Point Bi-serial (PTMACORR) correlation with person				Response
Key	*1*	*2*	*3*	*4*	Ordering
Q1*(2)*	-0.39	0.53	-0.31	-0.03	*3 1 4 2*
Q2*(3)*	-0.10	-0.10	0.39	-0.08	*1 4 2 3*
Q3*(4)*	-0.22	-0.16	-0.14	0.29	*1 2 3 4*
Q4*(2)**	-0.28	0.24	-0.03	0.07	*1 3 4 2*
Q5*(1)**	0.35	-0.33	-0.16	0.08	*3 2 4 1*
Q6*(2)*	-0.27	0.52	-0.23	-0.31	*1 3 4 2*
Q7*(4)*	-0.14	-0.34	-0.26	0.48	*1 2 3 4*
Q8*(1)*	0.63	-0.37	-0.44	-0.06	*3 2 4 1*
Q9*(4)*	-0.19	-0.22	-0.26	0.42	*3 1 2 4*
Q10*(2)*	-0.30	0.61	-0.34	-0.20	*3 1 4 2*
Q11*(3)*	-0.15	-0.41	0.43	-0.07	*2 1 4 3*
Q12*(3)*	-0.25	-0.27	-0.25	0.48	*2 4 1 3*
Q13*(1)*	0.42	-0.28	-0.23	-0.18	*2 4 3 1*
Q14*(4)*	-0.20	-0.33	-0.26	0.55	*3 1 2 4*
Q15*(2)*	-0.23	0.44	-0.29	-0.17	*3 1 4 2*
Q16*(2)*	-0.32	0.56	-0.15	-0.27	*4 1 3 2*
Q17*(4)*	-0.18	-0.35	-0.34	0.67	*2 3 1 4*
Q18*(1)**	0.41	0.02	-0.21	-0.16	*3 4 2 1*
Q19*(2)*	-0.12	0.29	-0.24	-0.22	*4 3 1 2*
Q20*(3)*	-0.19	-0.25	0.39	-0.19	*4 2 1 3*
Q21*(3)*	-0.19	-0.31	0.38	-0.14	*2 4 1 3*
Q22*(1)*	0.57	-0.37	-0.24	-0.26	*4 3 2 1*
Q23*(4)*	-0.23	-0.28	-0.24	0.51	*1 2 3 4*
Q24*(2)*	-0.32	0.50	-0.22	-0.17	*1 3 4 2*
Q25*(3)*	-0.13	-0.16	0.49	-0.07	*2 1 4 3*

*Distractor has positive correlation with person ability.

Classical test theory based distractors analysis results provided in Table 4.7 shows that respondents in lowest 25% group selected all possible distractors. Analysis showed that only few distractors were endorsed by respondents in higher 25% group than lower 25% group indicating ill functioning of such distractors. Distractor options 1 and 2 of Q2, option 4 of Q4 and Q5, option 4 of Q8, option 1 of Q11, option 2 of Q18, and option 1 of Q25 were identified to be revised using this approach. Rasch and conventional distractor analysis showed the definite necessity of revising option 4 of Q4 and Q5 and option 2 of Q18. Also, both analyses indicated that option 1 and 2 of Q2, option2 of Q14, option 2 of Q20 could be revised to improvement in the assessment. Overall, observed none-zero frequencies for the lower 25% group except for answer option 4 in Q8 indicated that the item distractors performed correctly (almost) in misleading those with relatively little knowledge.

Establishing Reliability Evidence

Person reliability for observed SAGS assessment data was 0.83. This corresponds to conventional CTT reliability and provides information about the capability of ordering group respondents in the same way (reproduce person ability hierarchy) in repeated administration of the instrument. Also, the observed value above 0.80 indicates that respondents can be discriminated into 2 or 3 levels based on their abilities (Boone et al., 2014). Further, the observed person separation index of 2.20 for this sample suggested that the instrument is a good measure, which was sensitive enough to differentiate between high and low ability respondents (Boone et al., 2014, Duncan et al., 2003). Moreover, the item reliability for observed SAGS assessment data was 0.96, and this highly desirable item reliability indicate an acceptable range of item difficulties. However, the corresponding statistic (to item reliability) was not observed in the

Table 4.7. Response Frequencies (Percentages) for the Top and Bottom 25% of Participants.

Item	Frequency and with-in group (Lower or Higher) % of Responses to Each Option									
	1		2		3		4		Omit	
(Key)	25%	75%	25%	75%	25%	75%	25%	75%	25%	75%
Q1*(2)*	12(40)	1(3)	8(27)	34(94)	8(27)	0(0)	2(7)	1(3)	0(0)	0(0)
Q2*(3)*	1(3)	1(3)	18(60)	17(47)	1(3)	15(42)	5(17)	3(8)	5(17)	0(0)
Q3*(4)*	3(10)	0(0)	3(10)	0(0)	6(20)	4(11)	18(60)	32(89)	0(0)	0(0)
Q4*(2)*	14(47)	3(8)	7(23)	19(53)	6(20)	7(19)	2(7)	6(17)	1(3)	1(3)
Q5*(1)*	9(30)	28(78)	16(53)	3(8)	4(13)	1(3)	1(3)	4(11)	0(0)	0(0)
Q6*(2)*	5(17)	0(0)	11(37)	34(94)	4(13)	0(0)	10(33)	2(6)	0(0)	0(0)
Q7*(4)*	2(7)	0(0)	12(40)	0(0)	6(20)	0(0)	10(33)	36(100)	0(0)	0(0)
Q8*(1)*	4(13)	35(97)	13(43)	0(0)	13(43)	0(0)	0(0)	1(3)	0(0)	0(0)
Q9*(4)*	3(10)	0(0)	9(30)	2(6)	6(20)	0(0)	12(40)	34(94)	0(0)	0(0)
Q10*(2)*	9(30)	0(0)	5(17)	36(100)	12(40)	0(0)	4(13)	0(0)	0(0)	0(0)
Q11*(3)*	2(7)	1(3)	12(40)	0(0)	14(47)	35(97)	2(7)	0(0)	0(0)	0(0)
Q12*(3)*	5(17)	0(0)	5(17)	0(0)	16(53)	36(100)	4(13)	0(0)	0(0)	0(0)
Q13*(1)*	13(43)	35(97)	5(17)	0(0)	7(23)	1(3)	5(17)	0(0)	0(0)	0(0)
Q14*(4)*	4(13)	0(0)	14(47)	4(11)	6(20)	0(0)	6(20)	32(89)	0(0)	0(0)
Q15*(2)*	6(20)	0(0)	14(47)	35(97)	5(17)	0(0)	5(17)	1(3)	0(0)	0(0)
Q16*(2)*	12(40)	1(3)	4(13)	32(89)	4(13)	1(3)	10(33)	2(6)	0(0)	0(0)
Q17*(4)*	3(10)	0(0)	13(43)	0(0)	13(43)	2(6)	1(3)	34(94)	0(0)	0(0)
Q18*(1)*	0(0)	14(39)	15(50)	22(61)	5(17)	0(0)	3(10)	0(0)	7(23)	0(0)
Q19*(2)*	8(27)	5(14)	15(50)	30(83)	3(10)	1(3)	4(13)	0(0)	0(0)	0(0)
Q20*(3)*	6(20)	2(6)	5(17)	1(3)	16(53)	33(92)	3(10)	0(0)	0(0)	0(0)
Q21*(3)*	8(27)	3(8)	7(23)	0(0)	14(47)	33(92)	1(3)	0(0)	0(0)	0(0)
Q22*(1)*	8(27)	36(100)	12(40)	0(0)	4(13)	0(0)	6(20)	0(0)	0(0)	0(0)
Q23*(4)*	3(10)	0(0)	8(27)	0(0)	14(47)	7(19)	5(17)	29(81)	0(0)	0(0)
Q24*(2)*	11(37)	0(0)	9(30)	35(97)	4(13)	0(0)	6(20)	1(3)	0(0)	0(0)
Q25*(3)*	6(20)	8(22)	11(37)	3(8)	4(13)	23(64)	3(10)	2(6)	6(20)	0(0)

Note. Participants were grouped based on their SAGS raw score. Percentages were rounded off and may not be added to 100%.

CTT approach. Item reliability provides information about reproducibility of the item hierarchy. Also, the higher item reliability index justified the appropriateness of the observed sample size to conduct Rasch analysis. The item separation index of 4.67 suggested an acceptable level of noise/error variance in SAGS data, and that also implied that the respondent sample was sufficient to confirm the item difficulty hierarchy (construct validity). Additionally, Cronbach's alpha value of 0.86 justified the strong internal consistency among the items in SAGS assessment (George & Mallory, 2003).

Using the features of the "WINSTEPS" software, item information functions were created and visually analyzed to identify ability ranges that maximize its information. In the range that maximize item information the latent trait is measured more reliably and accurately (de Ayala, 2009, DeMars, 2010). Regarding the SAGS instrument, a relatively difficult item, Q25, shows highest information at ability levels above the average ability of the respondent sample. Thus, items such as Q25 are desirable when creating an assessment targeting high ability students. A less difficult item, Q3, most accurately measures below average respondents, while Q10, a medium difficult item can be best used to measure respondents with average statistical research methodology knowledge. Samples of item information functions are given in Figure 4.4.

The test information function reports the statistical information in the data corresponding to each score or measure on the complete test (Linacre, 2016). Simply, the plot of test information function tells how accurately the complete instrument can estimate person locations (Templin, 2014). Theoretically, test information function peak at the point where the test most accurately measures given ability, and the width of test information function is the effective measurement range of the test (de Ayala, 2009; Linacre, 2016; Templin, 2014).

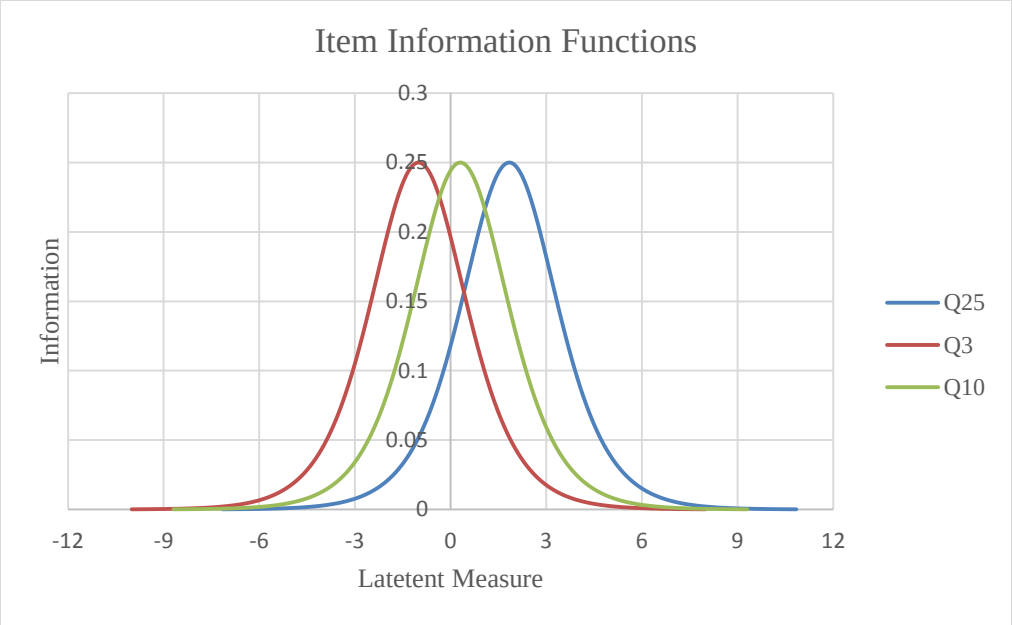


Figure 4.4. IIF's of Items in SAGS with Different Difficulties

Test information function for the SAGS instrument given in Figure 4.5 was obtained by summing all the 25 item information functions. As a complete test, the highest information was observed at an ability level below the average ability level of observed respondent sample. Thus, the SAGS instrument was most accurate to measure participants processing below average ability level.

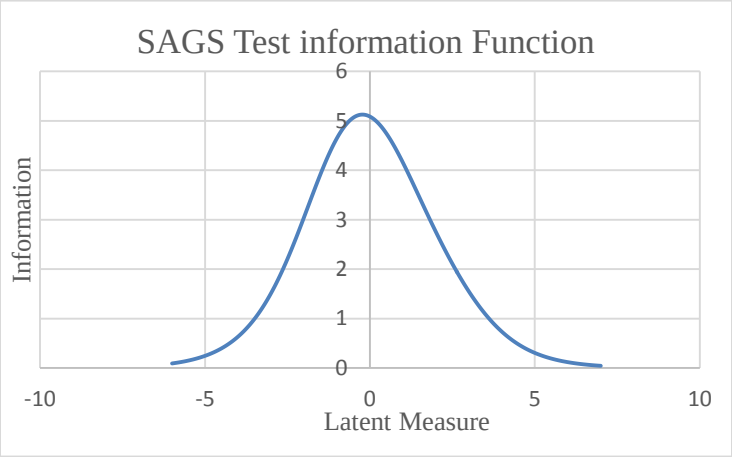


Figure 4.5. TIF of the SAGS Instrument

Establishing Validity Evidence

Construct and Predictive Validity

Initial construct validity evidence was evaluated by observing the item difficulty hierarchy (Boone et al., 2014; Linacre, 2016). Reflecting on the analysis with the Wright map presented in this chapter, item hierarchy was examined for construct validity. Ordering of high difficulty items (at the top of the map) matched the ordering conceptualized by the researcher at the item development stage. Higher difficulty items were the coefficient of variation, canonical correlations and log-linear analysis, this ordering was consistent with the difficulty ordering conceptualize by the researcher when developing items. Items testing the knowledge of

multivariate techniques (MANOVA, Discriminant analysis) and higher-level modeling (Multilevel modeling) were positioned below the most difficult set of items, and also consistent with the researcher's belief. Somewhat surprisingly the easiest items were repeated measures ANOVA, followed by confirmatory factor analysis. The next set of easier items were the items testing the knowledge of multiple linear regression, measures of variation, and Factorial ANOVA, which showed some agreement with what was expected by the researcher. Thus, observed item hierarchy showed satisfactory evidence of construct validity.

Validity Evidence Using Group Mean Comparisons

Construct and predicative validity was also established by examining how the repossess of items measuring the construct was influenced by different factors (Barlow, 2014; Linacre, 2016). Related factors were identified using the collected demographic variables and mean comparisons were performed for different levels of these factors. The researcher developed the SAGS assessment targeting EBS graduate students, so item responses were expected to be relatively similar for education students and students from other behavioral and social science disciplines. Rasch mean scores for education students and students coming from other disciplines were compared using independent sample *t test* to examine whether the discipline had an influence on the way construct had been defined. There was no difference in the mean Rasch scores of Education students (M=0.72, SD=1.48) and students from other disciplines (M=0.36, SD=1.09), $t(125)=-1.06$, $p=.293$. Thus, it was evident that SAGS functions equally for education students and students form other behavioral and social science disciplines. On the other hand, *t test* results showed that the mean score for students who took a research methodology course (M=0.67, SD=1.44) was higher than mean score for students who have not taken such a course (M=0.21, SD=1.24), but mean difference was not significantly different, $t(130)=-1.24$, $p=.217$.

Thus, it was evident that the SAGS measures a construct that goes beyond pure research methodology knowledge.

The mean score for students who took a three or more graduate level statistics courses (M=1.04, SD=1.02) was significantly higher than mean score for students who took only two or less course (M=0.10, SD=1.02), $t(112)=-3.29$, $p=.001$, $d=0.7$. Also, One-way ANOVA results showed that, the mean scores for student groups who have different statistic usage (often to never) were significantly different. Mean scores showed an increasing pattern moving form never used group (M=-0.98, SD=0.62), rarely used group (M=-0.08, SD=0.84), occasionally used group (M=0.98, SD=1.46), to often used group (M=1.43, SD=1.35), $F(3, 128)=19.56$, *Partial* η^2 =0.31. In both cases significantly higher means scores showed more statistics experience, and thus provide construct and predictive validity evidence of SAGS assessment. Summary of group mean comparisons are given in Table 4.8.

Convergent Validity

Convergent validity refers to the degree to which two measures of constructs that theoretically should be related are in fact related (Colton & Covert, 2007, 2007; University of York, n.d.). Specifically, in well planned instrument development projects, convergent validity is established through correlating scores from the new measure with scores form another validated measure related to a similar construct (Colton & Covert, 2007; Pathirage, 2015). However, in this study, a alternative measure (validated instrument) was not administered with SAGS instrument to facilitate convergent validity testing. Assuming that the self-reported confidence in doing statistics related tasks (an item in the demographic questionnaire) as a measure of statistical cognition,

Table 4.8. Construct and Predictive Validity through Known Group Comparisons.

Factor/ Background Variable	*N*	*M*	*SD*	*Test statistic*	*p*	*Effect Size*
Discipline						
Education	10	0.72	1.48			
Other	21	0.36	1.09	$t(125)=-1.06$	.293	
Research						
Taken	11	0.67	1.44			
Not Taken	17	0.21	1.33	$t(130)=-1.24$	.217	
Graduate level						
Two or less	28	0.10	1.02			
Three or more	86	1.04	1.40	$t(112)=-3.29$	.001*	$d=0.7$
Statistics usage						
Never	12	-0.98	0.62			
Rarely	42	-0.08	0.84			
Occasionally	35	0.98	1.46			
Often	43	1.43	1.35	$F(3, 128)=19.56$	.000*	*Partial* $\eta^2=0.31$

Note. *$p < .05$, ** $p < .01$, *** $p < .001$.

validity testing was performed using correlation analysis (University of York, n.d.). Self-reported confidence in ability to conduct statistics tasks and students' Rasch score showed a significant positive correlation, $r(132)=0.66$, $p< .001$. According to Cohen (1988) a coefficient of 0.66 is a large correlation and this evidence justifies the convergent validity.

Post-Estimation of Rasch Model

Rasch models are defined to have equal discrimination for each item. During Rasch model fitting, discrimination parameters are set to be equal, of value 1.0. But empirical item discriminations never are exactly equal to 1, and item discriminations vary from item to item. The "WINSTEPS" Rasch modeling software reports an estimate of those discrimination values as a post-hoc statistic (Linacre, 2016). The amount of the departure from 1.0 is an indication of the degree to which those items are inconstant with Rasch model. According to Linacre (2016), in general, the geometric mean of the estimated discriminations approximates value 1.0 for a good Rasch model. Geometric mean represented the 25th root of the product of all SAGS item (25 items) discrimination values (Costa & Judge, 2010). The geometric mean for SAGS discrimination parameter estimates was 0.93 and it did not show a larger deviance from overall benchmark discrimination value of 1. However, the individual item discrimination values 0.23, 1.55, 0.43 and 1.53 for Q4, Q5, Q17 and Q19 showed relatively higher deviance from the perfect individual discrimination value 1. Due to this slight departure and greater interest on examining the degree of respondents' guessing activities during the assessment, higher order item response theory models (2 PL and 3 PL) were examined for the SAGS instrument data. Fitting higher order models also enabled to identify additional item properties if all the items were not created equally well (Templin, 2014).

Evaluation of 2 PL and 3 PL IRT Models and Performance of Information Criteria

As explained in Chapter two Rasch model is a definition of measurement, philosophically Rasch modeling is different from IRT. It is important to note that the Rasch analysis was used purposefully to create a good measurement using definition of Rasch model (Bone et al., 2014). According to Rasch philosophy, if the person/items responses are not aligned with the Rasch model definition, observations causing misfit of the model are deleted from the analysis. Thus, subsequent parameter estimates are obtained for modified data (Boone et al., 2014). However, in an IRT framework, models were tested in a more exploratory manner using a portfolio of models to best describe the observed item responses without any modification to the original data (Shaw, 1991).

The general requirement of a larger sample size to evaluate higher order IRT models was not satisfied for this study. Thus a preliminary simulation study was conducted to generate larger samples which facilitated IRT modeling. Multivariate normal item data were generated using mean (proportion of correct) and interdependence (correlation matrix) among the items and later dichotomized to obtain binary item responses (Genz et al., 2008; Leisch, Weingessel, & Hornik, 2012). Models were evaluated for hypothetical samples of size 200, 500, and 1000, as the literature indicate the need of minimum of 200 responses for 2 PL model and 500 for 3 PL model (Barlow, 2014; de Ayala, 2009). Ten datasets representing each of these sample sizes were generated. Model evaluation statistics for original data and summaries statistics for the 10 datasets of the same size are given in Table 4.9.

Table 4.9. IRT Based Model Evaluation Summary Statistics.

Sample Size	Model			
Model evaluation statistics	Rasch df=25	Rasch/1 PL df=26	2 PL df=50	3 PL df=75
132				
Log-likelihood	-1837.19	-1836.45	-1777.34	-1746.38
Δ Log-likelihood: Sig[b]		0/1	1/1	1/1
AIC	3724.38	3724.90	3654.68	3642.76
SBC	3704.27	3703.78	3609.59	3572.64
CAIC	3859.31	3865.86	3963.82	4218.97
CICOMP	3828.12	3835.71	3921.39	6966.10
200[a]				
Log-likelihood	-33150.04	-3031.95	-2973.91	-2964.96
Δ Log-likelihood: Sig[b]	-	10/10	10/10	0/10
AIC	6132.75	6115.90	6047.65	6079.92
SBC	6113.04	6095.20	6003.11	6010.22
CAIC	6273.01	6262.12	6347.87	6571.20
CICOMP	6248.18	6228.62	6281.07	8936.67
500[a]				
Log-likelihood	-7553.87	-7536.29	-7412.41	-7403.23
Δ Log-likelihood: Sig[b]	-	10/10	10/10	0/10
AIC	15157.73	15124.57	14924.82	14949.33
SBC	15138.95	15104.78	14881.03	14887.58
CAIC	15317.75	15289.23	15247.15	15449.85
CICOMP	15295.87	15309.19	15197.87	20953.85
1000[a]				
Log-likelihood	-15145.23	-15109.85	-14875.43	-14868.70
Δ Log-likelihood: Sig[b]	-	10/10	10/10	0/10
AIC	29640.47	30271.71	29850.86	29887.40
SBC	30322.38	30252.61	29807.76	29789.80
CAIC	30514.55	30450.51	30201.73	30417.99
CICOMP	30498.69	30429.32	30158.29	32389.66

Note. [a]Samples were generated using original data, [b]Number of likelihood ratio tests that were significant after 10 runs. AIC=”Akaike’s Information Criteria”, SBC=”Schwartz’s Bayesian Criteria”, CAIC=”Bozdogan’s Consistent Information Criteria”, CICOMP=”Consistent Information Complexity Criteria”.

For the original data, Change in log-likelihood values suggested that the Rasch model with common discrimination parameter (1 PL model) was not significantly better fitting than the conventional Rasch model, (p=.224). The model with both difficulty and discrimination parameters for each item (2 PL) showed a better fit than 1PL model (p<.001), while 3 PL model revealed a better fit for observed data than 2 PL model (p<.001). Thus, the 3 PL model was judged to be the best fitting based on the classical likelihood ratio test. Also, lower values for AIC and SBC suggested that the 3 PL model is best fitting, CAIC and CICOMP identified the Rasch model to be a better fit to the the data. Results from likelihood ratio tests, AIC, and SBC were not consistent with the conclusions from the CAIC and CICOMP. Further, looking at simulation results 3 PL model was never identified as the best fitting model. Thus, it was evident that with small sample sizes different models selection criteria behaves differently.

Considering the fact that SAGS items were developed assuming the Rasch model definition, results showed encouraging evidence on using CAIC and ICOMP for IRT model selections with small sample size. When the sample size increases from 200 to 500 in the simulation study, all information criteria selected the 2 PL model as the best. But again for sample size of 1000, AIC and SBC selected two different models, Rasch and 3 PL models, as the best models. Still CAIC and CICOMP provided consistent results as it was for the sample size of 500. Further, there was no evidence toward 3 PL model in simulated data, thus impact of guessing in SAGS administration can be clearly ignore when SAGS scores are used for practical purposes. The majority of the model evaluation statistics observed for different sample sizes identified 2 PL to best describe simulated SAGS item responses.

Comparison of Parameter Estimates: IRT Estimates Vs. CTT Indices

Item parameter estimates 2 PL model with CTT equivalents are given in Table 4.10 for making meaningful comparisons. Estimated IRT parameters of SAGS items and CTT equivalents were compared to increase the validity of overall results as well as to give test developers a broader picture. During the analysis, the item difficulty "a", discrimination "b", and person ability "θ "estimates from the final 2PL model were compared with each item's CTT difficulty and CTT discrimination measured as item total correlation. IRT difficulty estimates shows no items as being overly difficult (i.e. "b" parameters >3.5), But corresponding CTT analysis shows Q2, Q18, and Q25 as difficult items (de Ayala, 2009; Testing Services, 2016). Q2, Q4 and Q18 were flagged as being poorly discriminating (i.e. "a" parameters < 0.40) based on IRT estimates (de Ayala, 2009). However, CTT discrimination indices suggested no poorly discriminating items (i.e. Discrimination index <0.24 or Item total correlations ranges between 0.00 to 0.09). However, item total correlations (i.e. Item total correlations ranges from 0.20 to 0.29) suggested that items Q3, Q4, and Q19 are reasonably good, but they are subject to improvement (Karelia, Pillai, & Vegada, 2013; University of Wisconsin, 2016).

Pearson correlations were calculated to quantify the relationship between CTT and IRT estimates, and discovered that the estimates for "a", "b", and "θ" parameters were strongly correlated with their CTT counterparts. Specifically, CTT difficulty (P) was highly negatively related to IRT item difficulty "b" with $r(23)= -0.925$, $p < .001$, and CTT discrimination (R) was highly positively correlated with IRT item discrimination "a" with $r(23) = 0.88$, $p < .001$. Similarly, "θ" estimates of person ability were highly positively related to CTT total score with $r(130)== 0.98$, $p < .001$.

Table 4.10. Table of Item Difficulties and Discriminations, CTT Vs IRT.

Item	Difficulty CTT[a]	Difficulty Level	Difficulty IRT	High %	Low %	DI CTT[b]	DI IRT[c]	Total Corr[d]
Q1	0.68	Medium	-0.70	94	27	0.67	1.67	0.55
Q2	0.20	High	3.24	42	4	0.38	0.36	0.35
Q3	0.76	Low	-2.32	89	60	0.29	0.54	0.28
Q4	0.39	Medium	2.07	54	24	0.30	0.21	0.22
Q5	0.51	Medium	-0.05	78	30	0.48	0.66	0.34
Q6	0.75	Low	-0.89	94	37	0.57	2.04	0.55
Q7	0.76	Low	-1.04	100	33	0.67	1.54	0.51
Q8	0.63	Medium	-0.45	97	13	0.84	2.50	0.68
Q9	0.74	Low	-1.07	94	40	0.54	1.22	0.43
Q10	0.51	Medium	-0.19	100	17	0.83	2.09	0.65
Q11	0.71	Low	-1.03	97	47	0.50	1.07	0.44
Q12	0.84	Low	-1.14	100	53	0.47	2.94	0.51
Q13	0.76	Low	-1.19	97	43	0.54	1.22	0.43
Q14	0.54	Medium	-0.20	89	20	0.69	1.60	0.58
Q15	0.73	Low	-1.04	97	47	0.50	1.21	0.46
Q16	0.50	Medium	-0.03	89	13	0.76	1.61	0.58
Q17	0.48	Medium	-0.01	94	3	0.91	2.63	0.71
Q18	0.23	High	2.75	39	0	0.39	0.37	0.36
Q19	0.67	Medium	-1.43	83	50	0.33	0.51	0.27
Q20	0.78	Low	-1.42	92	53	0.39	1.08	0.40
Q21	0.74	Low	-1.24	92	47	0.45	0.97	0.38
Q22	0.66	Medium	-0.56	100	27	0.73	2.17	0.61
Q23	0.49	Medium	-0.01	81	17	0.64	1.30	0.52
Q24	0.57	Medium	-0.30	97	30	0.67	1.32	0.51
Q25	0.29	High	1.23	64	15	0.49	0.65	0.42

Note. [a]Difficulty index=% of correct, Difficulty Index of <.3=High To.7=Medium, >.8=Low, [b]DI CTT=Discrimination Index=% correct in high Group - % correct in low group. [c]DI IRT=Discrimination IRT, [d]Total Corr=Item data (binary) correlation with total (raw) score.

Chapter Four Summary

The results described throughout chapter four have shown evidence for how the SAGS instrument has performed as a measure of statistical research methodology knowledge. The assessment established evidence for unidimensionality of the construct (statistical research methodology knowledge) measured by SAGS. Initial Rasch analysis showed a favorable global model fit indicating consistency of a Rasch measurement definition with the observed data. However, five SAGS items showed misfit evidence, thus some item responses of 16 individuals had to be deleted prior to attaining acceptable individual (local) fit for both the items and persons in the sample. Item Q4 (Correlations) was identified as an item diserving further improvements.

When investigating functionality of items through examining Rasch item difficulty estimates, Item Q2 (Coefficient of variation), Q18 (Canonical correlations) and Q25 (Log-linear analysis) were identified as most difficult items, and these were flagged for more attention. Graphical analysis with Wright map suggested adding more items with higher difficulties as well as more items with lower difficulties. Distractor analysis revealed problems with option 4 (Phi correlation coefficient) of Q4, option 4 (Independent samples z-test) Q5 and option 2 (Pearson product-moment correlation) of Q18.

The new SAGS instrument showed good reliability evidence and can be judged as a good measure with an acceptable range of estimated item difficulties (except for three items) capable of differentiating between high and low ability participants. Favorable separation indexes justified that the person sample was sufficient to confirm the item difficulty hierarchy. Item difficulty hierarchy was relatively consistent with researcher's belief about the item ordering justifying more than sufficient evidence concerning construct validity. Known group comparisons conducted using several demographic variables provided additional evidence

towards construct validity, predictive validity as well and convergent validity of SAGS instrument.

Taking more descriptive approach compared to more perspective Rasch approach, higher order IRT models were fitted to simulated SAGS data. Results indicated inconsistent behavior of various model selection criteria when selecting best IRT model. Two parameter model (2 PL) was identified as best to describe simulated SAGS item response data. Simulation results showed no support for 3 PL model providing evidence against significant impact of guessing when completing SAGS assessment. Chapter five will position these findings within the context of the statistics education and assessment development literature, and future of the SAGS instrument.

CHAPTER FIVE

DISCUSSION

Chapter Five is directed to position the results from developing the SAGS instrument within the larger body of statistics education and assessment development literature. Results from chapter four will be reviewed based on the purpose and specific research objectives. Limitations associated with the study and suggestions for future researchers to improve upon these limitations will also be presented in this chapter. Finally, a number of implications for graduate statistics education and establishing good measurement practice will be described.

Summary of Study Purpose, Objectives, and Method

The literature indicates at least one statistics or quantitative course is mandatory for students perusing graduate degrees in education and other behavioral and social sciences (Aiken et al., 2008; Capraro & Thompson, 2008; Henry, 2013; Onwuegbuzie & Wilson, 2003). However, for most students statistics is considered as an anxiety provoking subject (Pan & Tang, 2004; Onwuegbuzie et al., 1997). Students have shown a greater degree of misconceptions, variable knowledge, and in some cases lack of knowledge on higher level statistics (Aiken et al., 2008; Alacaci, 2012; Bessant, 1992; Henson et al., 2010; STATtr@k, 2012), thus for many students selecting appropriate statistical procedure to answer their own research questions has become a dilemma. Measuring the ability to select appropriate statistical procedure is considered to be important, and such measure can be used to better teach students (Alacaci, 2012; Heitman et al., 2007; Marusteri & Bacarea, 2010).

Although there are similar instruments measuring statistical competencies, there is no validated assessment directly targeting graduate students and measuring their statistical test/procedure selection ability (Delmas et al.,2004; Grafield et al.,2012; Grafield, 1998a; Stone,

et al., 2003; Sundre, 2003; Zeigler 2014). Currently available instruments seek to measure major constructs such as statistical reasoning, statistical thinking, and statistical literacy of undergraduate and high school students (Alacaci, 2012; Zeiffler et al., 2008). Further, lack of extensive psychometric analysis and use of Classical Test Theory (CTT) methodologies make these instruments' irrelevant for students outside of their original sample (Boone et al, 2014; de Ayala, 2009; Delmas et al.,2004; Grafield et al.,2012; Grafield, 1998a; Stone, et al., 2003; Sundre, 2003; Templin, 2012). Without a generalizable instrument to measure statistics abilities, statistics educators face difficulties with accurately assessing graduate students' statistics knowledge.

The purpose of the present study was to address this measurement requirement through developing and validating a new assessment. Specifically, this study sought to establish preliminary item characteristics and validity evidence for the Statistics Assessment of Graduate Students (SAGS) instrument. Rather than using CTT to develop the instrument, Rasch and Item Response Theory (IRT) was used in order to offer educators item and person ability parameters that are independent of the sample from which they are estimated (Boone et al., 2014; de Ayala, 2009; Templin, 2012). This invariance trait could provide statistics educators the freedom to customize their test by using selected number of items that function best for the intended examinee sample and the assessment objective. The study specifically aimed to address four primary objectives, and Table 5.1 provides an overview of methods used to address each primary research objective.

Table 5.1. Summary of Methods by Research Objective.

	Research Objective	Primary Method(s)
1	Establish content validity evidence for the SAGS instrument	An expert panel consisted of five faculty members and one statistics consultant reviewed the SAGS items and provided feedback
2	Examine the model fit of the SAGS items to a Rasch model	Chi- Square goodness fit test (Global) was conducted. Local fit was evaluated using MNSQ values and Z-Residuals. Misfitting item responses were deleted to obtain the fit for Rasch model. WINSTEPS program was used.
2a	Test the assumptions of unidimensionality and local independence.	Confirmatory Factor Analysis (CFA) was used to test the existence of one factor. Comparisons were made with two or three factor models (Experimentally) to justify one factor model. This was done prior to Rasch model fitting.
2b	Identify item difficulties, and analyze the item information/ test information of the SAGS instrument	Item parameters were estimated using joint maximum likelihood method. Item information functions were plotted. Item information was summed to examine the total test information.
2c	Analyze the quality of item distracters of the SAGS assessment	Direction and ordering of PTMACORR values of distractors in WINSTEPS were examined to identify distractors. Classical distractor analysis was performed with top and bottom 25% of examinees. Distractors that performed poorly were flagged for review.
3	Examine the reliably and validity of the SAGS instrument	Rasch framework correlations and operations were observed for reliability, and construct maps and demographic data were used to test validity.
3a	Assess the reliability of the instrument through analysis of various reliability and separation indices	Item correlation, Person correlations, item separation, Person separation, and Cronbach's alpha were calculated.
3b	Assess construct, predictive, and other forms of validity of the instrument through construct map and known group comparisons	Construct map was examined to justify the items hierarchy. A combination of descriptive analysis, independent t-tests, and ANOVA were used to compare person ability estimates across different participant groups.

Table 5.1. *(Continued)*

	Research Objective	Primary Method(s)
4	Examine the model fit of the SAGS items to 1PL, 2PL, and 3PL IRT models based on simulated data	Multiple multivariate binary datasets of size 200, 500, and 1000 were simulated using mean and correlation matrices of observed sample.
4a	Investigate the performance of novel information complexity criteria (ICOMP) over other model selection criteria for determining the best fitting IRT model	A program in R software was developed and ICOMP and other model selection criteria were coded. Comparisons were made between AIC, SBC, CAIC, and difference in log likelihood for portfolio of IRT models (Rasch, 1PL, 2 PL, and 3 PL).
4b	Identify item difficulty, discrimination, and guessing parameters	For the best fitting model (2PL), parameters were estimated using the R program. Parameters were estimated using latent variable modeling (ltm) package.
4c	Compare person ability and item location estimates (difficulty, discrimination, and guessing) from IRT models to those of CTT indices	Pearson correlations were used to compare (1) IRT "θ" estimates with CTT total-correct scores, (2) IRT "b" parameters to CTT difficulty index, and (3) IRT "a" parameters to CTT discrimination index.

Implementation and Results of SAGS Development

The SAGS instrument was developed using a 25 one-best-answer format questions (Case & Swanson, 2002). Each question presented the examinee with an applied research scenario with a specific research question, and was written as 4-option multiple choice question (MCQ). Response options were selected specifically to allow educators as well as students to distinguish precisely their misconceptions on a given statistical procedure (Suskie, 2009).

The SAGS instrument was administered online through "Qualtrics" survey management system. Instructors who taught graduate level quantitative courses in EBS disciplines were contacted, and they were asked to send an announcement of the assessment, which included the link to the online assessment. Invitation to participate in the SAGS administration was also send through various listserv, SAGS was also posted in websites, discussion boards, and social media. The SAGS cognitive items were distributed along with background demographic questions that were used to identify the nature of the sample and used for reliability and validity testing. The assessment took approximately 30-40 minutes to be complete by a participant. Once completed, descriptive answer key was presented to the students. At the time of current analysis 173 students participated in the assessment, however only 132 competed the entire 25 cognitive questions. Only the fully competed responses were used for the current study. The primary findings from this study are summarized by research objective as follows:

1. *Establish content validity evidence for the SAGS assessment*
    a. A focus group meeting with upper level graduate students coming from EBS disciplines was held. Focused group members evaluated whether the question stems were applicable to applied research problems in their disciplines. Some questions were revised and some were re-rewritten after the focus group.

b. A heterogeneous panel of six statistics experts critiqued the content and items, which were finalized after the focus group. Expert review panel determined whether the SAGS items covered the scope of inferential statistical tests applicable to graduate level research in education and other social and behavioral sciences.

c. Three reviewers requested to add descriptive statistics questions.

d. Three reviewers made comments about infrequent use of Canonical Correlation and Log-linear analysis. But two of them asked to administer these items during the data collection.

e. Two of the five expert reviewers were informally interviewed for additional follow-up discussions. Decisions were made regarding changes to improve the quality of the items/instrument.

2. *Examine the model fit of the SAGS items to a Rasch model*

Initial Rasch analysis provided good global model fits showing a non-significant chi-square likelihood ratio statistic. However, the local fit analysis identified five (Q2, Q4, Q12, Q18, and Q25 items to be misfitting. Removal of the misfitting person responses for these items resulted good local fit for all items except for Q4 in the second Rasch analysis. Q4 was kept in the SAGS instrument as it was still considered good for creating a productive measurement (Wright et al., 1994).

a. Test the assumptions of unidimensionality and local independence

Confirmatory Factor Analysis (CFA) conducted to test for one factor model concluded that responses of 25-item instrument represented unidimensional latent

contract (Cook, et al., 2009; Deng et al., 2008). Unidimensionality of the item responses also justified the local independence assumption (DeMars, 2010).

b. Identify the item difficulties, person abilities, and analyze the item information and test information of SAGS

   i. Difficulty parameters ranged from -1.59 to 2.47 (*M* = 0.00, *SD* = 1.05). Considering Rasch model as a subset of IRT models, and using benchmarks for IRT difficulties, no overly difficult items were identified (de Ayala, 2009). Most difficult items were Q2 (Coefficient of variation), Q18(Canonical correlation), and Q25 (Log-linear analysis). Surprisingly, most essay items were Q12 (Repeated measures ANOVA) and Q20 (Confirmatory factor analysis). But next set of easy items were Q3 (Measures of variation) and Q13(Multiple regression).

   ii. Person ability estimates (in logit scale) ranged from -1.97 to 3.76 (*M* = 0.51, *SD* = 1.27) for 129 non-extreme respondents. Two individuals got all items correct and one individual got 24 of the items correct, and these individuals were identified as extreme. Respective average raw score was 14.70 (*SD*=5.30) for respondents identified as non-extreme.

   iii. In Rasch modeling value for highest information is 0.25 for every item. and it is observed at the difficulty level of the item. Thus, all items were most and equally reliable at their respective difficulty levels. Highest peak of the Test Information Function (TIF) observed just below the average difficulty level for all items (logit value of 0). Thus, test was most reliable to measure the abilities of individuals below the average ability level of the observed sample.

c. *Analyze the quality of item distracters of the SAGS assessment*

i. Every one of the 100 possible response options was chosen at least once by the participants. Only one distractor was not chosen at all by the individuals in the lowest 25% of raw scores group, which compared to 40 that were not chosen by the highest 25%.

ii. Combination of Rasch and conventional distractor analysis showed the definite necessity of revising option 4 (Phi correlation coefficient) of Q4, option 4 (independent sample z test) of Q5 and option 2 (Pearson product moment correlation) of Q18.

iii. Also, both Rasch and classical approach distractor analysis indicated that option 1 (Range) and 2 (Standard deviation) of Q2, option2 (Multiple Linear Regression) of Q14, option 2 (Canonical correlations) of Q20 could be revised to make improvements.

3. *Examine the reliability and validity evidence for the SAGS assessment*

a. Assess the reliability of SAGS through analysis of various reliability indexes and separation indexes

i. Good person reliability (this corresponds to conventional reliability in CTT) value of 0.83 indicated that respondents could be discriminated into 2 or 3 levels based on their abilities (Boone et al., 2014).

ii. Good person separation index of 2.20 suggested the instrument was sensitive enough to differentiate between high and low ability respondents (Boone et al., 2014, Duncan et al., 2003).

iii. Very good item reliability (no corresponding value in CTT) for observed SAGS assessment data was 0.96, and this highly desirable item reliability indicated acceptable range of item difficulties. Further it was an indication for acceptable sample size (Boone et al., 2014, Duncan et al., 2003).

iv. The item separation index of 4.67 implied that person sample was sufficient to confirm the item difficulty hierarchy (construct validity). (Boone et al., 2014, Duncan et al., 2003).

v. Cronbach's alpha value of 0 .86 justified the strong internal consistency among the items in SAGS assessment (George & Mallery, 2003).

b. Assess the construct, predictive and other relevant validity evidence of the instrument through analysis of construct map and known group comparisons

i. Observed item hierarchy in the construct map showed some agreement with what was expected by the researcher about positioning of the items. This provided satisfactory evidence of construct validity of SAGS.

ii. No difference found in Rasch scores between education majors another majors justified that the SAGS function equally for both these groups.

iii. Rasch scores were higher but not significant for students who took research methodology course compare to who did not took such course. This indicates SAGS measures a construct that goes beyond pure research methodology knowledge.

iv. Respondents with higher statistics experience (Number of graduate level course taken, Frequency of statistics usage outside classes) showed significantly higher Rasch scores. Also, significant and positive correlations

were observed between statistics experience and Rasch scores. These provide supporting evidence towards construct and predictive validity.

v. Those who perceived they were confident in doing statistics received higher Rasch scores with significantly positive correlation, showing convergent validity evidence of SAGS.

4. *Examine the model fit of the SAGS items to 1 PL, 2PL and 3PL IRT models based on simulated data*

   a. Investigate the performance of novel information complexity criteria (ICOMP) over other model selection criteria for selecting the best fitting IRT model

      i. Convectional likelihood ratio test and other information theoretic model selection criteria behaved differently when selecting the best model among the portfolio of IRT models.

      ii. Assuming SAGS items were developed using Rasch model definition there was encouraging evidence on using CAIC and ICOMP for IRT model selections with small sample sizes.

   b. Identify item difficulty, discrimination and guessing parameters

      i. 2 PL model was selected as the best model to describe the observed item responses in first SAGS administration.

      ii. No evidence was found in the simulation study to justify the existence of 3 PL model, thus evidence against significant impact of guessing on SAGS responses.

iii. No overly difficult items (i.e. "b" parameters >3.5) were found but Q2, Q4 and Q18 were flagged as being poorly discriminating (i.e. "a" parameters < 0.40) based on IRT estimates (de Ayala, 2009).

c. *Compare person ability and item location estimates (difficulty, discrimination and guessing) from IRT models to those of CTT indices*

CTT item parameter estimates, difficulty (P) and discrimination Index (R) were highly correlated with corresponding IRT parameter estimates.

SAGS Results – Alignment with Previous Research

Results from the initial SAGS administration aligns with findings from previous research. There were several similarities observed which can be classified in to four major areas; (1) Item difficulty parameters and most and least used statistical procedures, (2) sources for item content validity evidence, (3) item construction elements, and (4) performance of various model selection criteria in selecting the best IRT models.

Item Difficulty Parameters and Most Used and Least Used Statistical Procedures

The body of research on statistical methods used in doctoral dissertations and master's thesis in education and other behavioral sciences for over 40 years has consistently shown ANOVA and mean comparison analyses is one of most used statistical analysis (Aiken et al., 2008; Hsu, 2005; Karadağ, 2010; Keselman et al. 1998; Mubarak, 2011; Onwuegbuzie, 2002; Woehlke, 1988). Goodwin (1985a, b) provides numbers of related statistical techniques used in the Journal of Educational Psychology (JEP) and American education research Journal (AERJ) between 1979 and 1983. The most commonly used statistics in AERJ were ANOVA/ANCOVA (17%), and in JEP were ANOVA/ANCOVA (26%). These results as well as the study that examined literature from 1979 to 1997 by Elmore and Woehlke (1998) indicated that other

popular and mostly used statistical techniques to be descriptive statistics, correlation/regression, t tests, and multivariate techniques respectively. Also, the review of the most commonly taught courses presented in chapter 2 identified that these techniques were taught in introductory and intermediate levels in most graduate programs, thus students are more competent on these analyses. SAGS results showed consistency with these evidence when Repeated Measures ANOVA, Factorial ANOVA, Measures of Variation and Multiple Regression became most easy items. Further, in SAGS all the ANOVA and other mean comparison analyses have difficulty level below the average difficulty level of the items. Interestingly, Keselman et al. (1998) reviewed four hundred and eleven articles in 1994 and 1995 issues of 17 educational and psychological journals, and found that Repeated Measures ANOVA was most frequently used (55%), and it was the item that found to be the least difficult in SAGS.

Canonical Correlation Analysis (CCA) is one of the difficult items in SAGS. According to Shery and Henson (2005) this analysis was underutilized by researchers. CCA is used to address research questions that are multivariate in nature, but most researchers inappropriately used related univariate analyses instead of CCA. Even though this analysis has becoming more popular due to the advancement of statistical software SAGS data shows only few students were able to correctly answer this question. Literature review in chapter 2 recognized that some universities offer categorical data analysis courses to graduate students that cover log-linear analysis. But it extends beyond the compulsory intermediate level courses. Aiken at al. (2008) reported categorical data analysis course as a special course. Also the review of the most used statistical procedures does not show evidence on greater usage of this technique (Elmore and Woehlke, 1998; Godwin, 1985a,b; Hsu, 2005; Karadağ, 2010; Keselman et al. 1998; Mubarak, 2011; Onwuegbuzie, 2002; Woehlke, 1988). Thus, it is evident that only some students had

exposure to log-linear analysis, and this was observed as a higher difficulty item in the SAGS administration. Coefficient of variation (CV) is a popular statistic among few social science disciplines, especially among demographers (Sørensen, 2002), and business and organizational research disciplines (Bedeian & Mossholder, 2000; Powres & Powers, 2009). Even though the mean and standard deviation are heavily reported in education research, CV which is based on both mean and standard deviation is rarely used among education researchers (Powers & powers, 2009; Reed, Lynn, & Meade, 2009). The coefficient variation (CV) measure is proposed to be used along with highly popular Standard Error of Measurement (SEM) reported in reliability studies in behavioral sciences (Atkinson & Nevill, 1998). Further, Sorensen (2002) have noted that the researchers did not have clear idea about correctly using the CV. SAGS show CV as the most difficult item to answer, and provide evidence for lack of popularity and misconceptions through higher difficulty estimate.

Sources for Content Validity Evidence

An expert review approach was used in this study to gather evidence of content validity similar to previous instruments (Garfield, 1998a; Garfield, 2003; Sundre, 2003). Moreover, this study, like similar assessments in statistics and other disciplines, used reviews of commonly used statistical procedures in educations and other behavioral and social sciences during the test item construction (Allen et al/, 2003; Barlow, 2014; Delmas et al, 2004; Horwitz & Switzer, 2009; Ziegler, 2014). Thus, this study has utilized similar procedure as other studies to establish content validity evidence prior to the SAGS administration. Furthermore, Rasch Item difficulty parameters obtained in this study do not indicate extreme difficult items (de Ayala, 2009; Testing Services, 2016). Observed difficulties (most difficult item Q2 had 20% correct, and least difficult item Q12 had 84% correct) fall under the benchmarks (15% correct to 85% correct)

considered as extremes for Rasch sample size determination (Linacre, 2016). Thus the SAGS was not too difficult or too easy for students and shows appropriate content coverage.

Sources for Item Construct Validity Evidence

The unidimensionality assumption in IRT implied that only single construct is influencing the item responses. Zieffler (2014) justifies the construct validity of recently developed Basic Literacy in Statistics (BLIS) assessment for undergraduates by testing for the unidimensionality through confirmatory factor analysis. The same procedure was performed for SAGS data, and similar results were observed providing identical supporting evidence for construct validity. Adding to the discussions made with item difficulties and most used and least used statistical procures it is clear that the top and the bottom levels of items hierarchy (Wright map) are consistent with the available literature (Elmore and Woehlke, 1998; Godwin, 1985a,b; Hsu, 2005; Karadağ, 2010; Keselman et al. 1998; Mubarak, 2011; Onwuegbuzie, 2002; Woehlke, 1988). Aiken et al. (2008) reviewing statistics content taught in graduate programs in psychology identified multivariate methods were to be taught less frequently than ANOVA and multiple regression. More advanced and higher level modeling techniques such as structural equation modeling or multilevel modeling were taught less frequently than multivariate methods. Also, t-tests positions above ANOVA and regressions while chi-square test fall toward the middle of the list of frequently used procedures (i.e. between ANOVA and multivariate methods) (Stallings, West & Carmody, 1983; Godwin, 1985a, b). Further, Elmore and Woehlke (1998) reviews show that multivariate methods position at the top of the intermediate level in their list of most frequently used statistical techniques. Karadağ (2010) identified factor analysis as the most frequently used multivariate technique. Observed SAGS item difficulty in the middle of the

construct map is relatively consistent with these findings, thus provides additional evidence towards construct validity.

Item Construction Elements

Another similarity between the SAGS instrument and similar statistical knowledge assessment instruments is the structure of the assessment itself (Allen et al., 2003; Barlow, 2014; Garfield, 1998a; Garfield, 2003). The SAGS assessment used unique item stems derived from common type research and analysis objectives to construct the items. Each item provided one of these research examples in a one-best-answer format, which is considered as the best approach for writing high quality multiple choice questions (Case & Swanson, 2002).

Performance of Various Model Selection Criteria in Selecting Best IRT Models

Inconsistencies and inaccuracies were found among various model section criteria when selecting the best IRT model form the portfolio of models (Kang & Cohen, 2007). Simulation results by Kang and Cohen show Likelihood Ratio (LR) test, Akaike information Criteria (AIC), and Schwartz's Bayesian Information Criteria (SBC) were more accurate when true models are from 1PL and 2PL. But these criteria were less accurate when the true model is 3PL. Boundary problem of guessing parameter being equal to 0 cause LR test to be inaccurate when selecting between 2 PL 3 PL models (Brown, Templin & Cohen, 2015; Wilks, 1938). Further these studies suggest the impact of the sample size on differential performance of model various selection criteria. As example, Deviance Information Criteria (DIC) does not work well when the data coming from simpler models (1 PL or 2 PL), but it performed better when the sample size become large. Simulation results of this study clearly show the inconsistencies among LR test, AIC and SBC under different models and varying sample sizes supporting the results observed in previous research.

SAGS Results – Expanding Upon Previous Studies

While there have been studies on developing statistics knowledge assessments nearly for two decades, the SAGS assessment has addressed several of the shortcomings these previous instruments possessed. Following section explains three major ways the SAGS assessment has expanded upon and unique than existing instruments. The contributions have been organized by (1) filling the measurement gap (2) expanding the content coverage and improved item quality (3) strengthening the psychometric rigor of assessing statistical cognition and (4) exploring performance of novel model selection criteria.

Filling the Measurement Gap

Looking at the last two decades of statistics education literature attempts were made to develop assessments to measures students' statistics knowledge and skills (Bidgood, Hunt, & Jolliffe, 2010). Almost all these instruments were intended to measure major constructs, statistical reasoning, statistical thinking, and statistical literacy of undergraduate and high school students (Delmas et al., 2002; Garfield, 1998 (a); Garfield et al., 2012; Schield, 2002; Stone, 2006; Sundre, 2003; Ziegler, 2014). Moreover, these instruments are focused on assessing conceptual knowledge rather than assessing skills of applying statistics to find answers to research questions through identifying the correct statistical procedure. Expanding on similar instruments Barlow (2014) developed the Bio Statistics and Clinical Epidemiology Skills assessment (BACES). Some components of BACES test the knowledge of selection of appropriate statistical test for given research situation, but it is targeted for medical residents. There is no validated assessment that is directly targeting graduate students and comprehensively measuring their statistical test/procedure selection ability. Current study addresses this measurement gap by developing the novel SAGS instrument.

Expanded Content Coverage and Improved Item Quality

Most of the previous statistics assessments target the undergraduates and introductory courses, and items cover a limited content area. From these early assessments, SRA covers descriptive statistics, basic probability, and correlations (Garfield, 1998). The CAOS and related ARTIST topic scales add significant testing, mean comparisons, and basic distribution theory beyond SRA (Delmas et al., 2002). The Statistic Concept Inventory (SCI) targets engineering students and mostly test the knowledge of descriptive statistics and basic probability theory (Allen, 2003). Newer assessments such as GOALS and BLIS still test the basic concepts but they are more geared towards the modern day statistics course that utilize more computing technology (Garfield et al., 2012; Ziegler, 2014). Considering somewhat related assessments and specially looking at the novel bio statistics assessment, BASES, there was no indication of testing the knowledge of higher level statistical methods even though it tests the selection abilities of some applied statistics procedures common in graduate medical education (Barlow, 2014). But SAGS items covers a wider range of content area applicable to education and other behavioral social sciences. One of most detailed and frequently used statistics assessment CAOS test has multiple items attached to common vignette (Delmas et al., 2002). Item dependency occurs when the answer to one item directly influences the answer to another item through salient response options or a common example, vignette, etc. (DeMars, 2010). Item dependency was one of the commonly seen writing errors among the other existing instruments including GOALS, SRA, and also the early bio-statistics assessment available in medical literature (Barlow, 2014; Garfield, 1998; Garfield, 2010). But the SAGS assessment was developed to maximize item independency by using a stem with unique research scenario for each item.

Strengthening the Psychometric Accuracy of Assessing Statistics Constructs

One of the important addition to previous instruments was using an Rasch approach to SAGS instrument development, which allows to create more generalizable measurements and customized statistics assessments in the future (Boone et al., 2014; de Ayala, 2009). Although most popular statistical assessments, SRA and CAOS report psychometric properties, majority of others report very limited to no information (Delmas et al., 2002; Garfield, 1998 (a); Garfield et al., 2012; Schield, 2002; Stone, 2006; Sundre, 2003; Ziegler, 2014). Psychometric properties for new assessments developed after 2010 were also not readily available in the literature. Additionally, the utilization of CTT approach for these instrument developments has hindered the use of established psychometric properties beyond the sample which they were originally tested (DeMars, 2010; Hays et al., 2000). The SAGS development, in contrast, fit a Rasch model, the simplest form of IRT model to the item response data. The parameters that were generated from that model can be easily tested in additional samples. Rather than drastically change across administrations, the IRT parameters ought to remain invariant and could be converted using linear transformations (Stage, 1998; DeMars, 2010; de Ayala, 2009, Furr & Bacharach, 2008). This property allows items to be reassembled, or create new test versions without losing their accuracy or consistency in estimating person ability levels. In other words, the SAGS items could be broken up into create shorter version of SAGS instrument, build SAGS item inventory, or create assessments covering selected content area depending on the needs of the examiner or testing program.

Exploring Performance of Novel Model Selection Criteria

The simulation piece of this study provided a new insight towards the use of two additional model selection criteria when selecting the best IRT model from a portfolio of models.

Various model selections indexes, LR, AIC, and SBC tend to select different models as the best model depending on the item properties and sample sizes (Brown et al., 2015; Kang & Cohen, 2007). Further, the literature indicated the presence of correlation among the estimated item difficulty, discrimination, and pseudo-guessing parameters (DeMars, 2010; Hotiu, 2006; Rasiah, & Isaiah, 2006; Sing, 2014; Sushma, 2013). The indexes mentioned before does not take into account the correlation between estimated model parameters and tend to penalize less for higher order models, but ICOMP family criteria adds penalty for over-parametrization and account for dependence among item parameters (Bozdogan, 1990; Kolenikov, 2000). However, these type of criteria have never been explored with the IRT model selections. Results show that the ICOMP family criteria performed differently, especially in the case of small sample sizes. Therefore, this study opens the avenue to explore benefits of using ICOMP type criteria and emphasizes the need of conducting additional research.

Practical Implications

The section on practical implication presents the applicability of this study to graduate students and statistics educators. Also, this section identifies the limitation of current study which leads to next section talks about future research.

Practical Implications of SAGS for Graduate Students

The findings of this study have significant implications to the field of statistics education for a number of reasons. The SAGS instrument developed under the current study serves as a measurement tool for assessing students' judgments of their own ability to choose appropriate statistical test to answer given research questions. Thus, the SAGS offer students the opportunity to self-evaluate their own statistical knowledge. As a learner, assessment of statistical cognition provides a great opportunity to realize how much knowledge currently they possess, how much

knowledge they have gained from their past learning experiences, and how much more they need to gain in order to achieve required level of competencies (Barnes, 2015; Bennett, 2011; Bremner et al., 2014; Delmas et al., 2007; Stiggins, 2005, Wright, 2008). According to Schunk (1995) when students see their progress, it can strengthen their self-efficacy and motivate them to work hard. Hence, assessing statistical abilities through instrument like SAGS could motivate students to further increase their level of statistics knowledge through continuous learning. Self-assessment of student's statistical test selection abilities using multiple-choice questions in SASG allow them to critically evaluate suitability of alternative statistical procedures to solve a research problem, and SAGS answer sheet helps students to recognize the misconceptions they have about various statistical procedures when they got particular items wrong. So students who find lower level of statistical abilities or identify possessing specific misconception can look for extra assistance from faculty and statistics consultant or utilize additional resources to select appropriate statistical test or procedure to solve their research questions. In addition, items of SAGS instrument are closely aligned with the competencies measured at graduate level research methods comprehensive exams, and also with the competencies used by employees when recruiting candidates for employment positions such as statistician, quantitative analyst, and data scientist (Statistics Canada, n.d.; University of Memphis, n.d.; University of Pittsburg, 2013), Thus, individuals can use SAGS as a practice test prior to those exams in order to refresh their applied statistics knowledge.

Practical Implications of SAGS for Statistics Educators

Past literature has demonstrated that assessing students' statistics ability related to selecting appropriate statistical test to solve their research questions to be important, and such assessments will ultimately lead to enhance the process of teaching statistics (Alcaci, 2012;

Bidgood et al., 2010; Garfield & Franklin, 2011). Therefore, students' statistical research methodology knowledge assessed through SAGS provide an indication of students' applied statistics performance. In particular SAGS most accurately distinguish students who process below or above average level of statistical research methodology knowledge. Thus, SAGS assessment is ideal screening tool for differentiate intermediate level of statistics knowledge. By administrating SAGS to their students statistics educators will be able to identify strengths and weaknesses of their students and it will be helpful to organize their teaching process by identifying areas needed to be given a higher focus.

Students enter to graduate programs in education and other social and behavioral sciences with different levels of statistics knowledge and skills due to their past statistics education. Since departments offer different levels of statistics classes, recommending an appropriate statistics course to enroll has become an important question for faculty who mentor those students (Dunn et al, 2012; Gelman et al., 2014). At the graduate level, some students may need to begin learning statistics with basic courses as they have no experience, while other students might be able to start with higher level courses. Thus, SAGS can be used to pre-assess students' statistics knowledge, and students can be placed in the most suitable statistics course to advance their knowledge in efficiently. Similarly, SAGS can be used to select top performing students in statistics for more quantitative oriented graduate programs. During the completion of graduate degrees some programs in education require their students to take a research methodology comprehensive exam (University of Memphis, n.d.; University of Pittsburg, 2013). Questions in SAGS closely align with the statistics portion of these exams, thus faculty can use SAGS assessment questions to test a broader range of statistical research methodology competencies.

Moreover, as mentioned in the previous section employers can use SAGS as a screening tool to select candidates for quantitate research oriented positions.

With no statistics assessment measuring the statistics test selection ability of graduate students in education and other behavioral sciences, SAGS instrument could be considered as a new member to the family of validated assessments available in statistics education. Since graduate level content coverage is considerably different from undergraduate level, statistics educators, who are conducting research with the graduate student population about various instructional technologies and different teaching interventions, can use this assessment to measure students' applied statistics skills accurately than using currently available instruments. Importantly, estimated item parameters using Item Response Theory approach in this study allow the ability to customize the SAGS instrument according to various assessment objectives and to develop a SAGS item inventory. Overall, new SAGS instrument provide a unique contribution and fill up a much needed measurement gap in statistics education.

Limitations of Present Study

The study provides positive preliminary evidence for the reliability and validity of the SAGS instrument. It is important to recognize three key limitations of the study design and observed results.

The major limitation of the study was the SAGS administration was in an uncontrolled environment. The test was administered using "Qualtrics" on-line survey software but no time constraint was set for participants to compete the SAGS cognitive questions. Also, because it was an online administration the participants could look at relevant lecture notes or other online materials to help complete the assessment. Since a completely controlled testing situation was not possible, cheating or lack of motivation may be incorporated in SAGS results. The non-

randomized design used in this study may lead to a higher level of possible sampling error, which might have also hindered the generalizability of the results. The present study collected data with the help of instructors majorly from one US University. One instructor each form three other US universities advertised the study to their student. Also, SAGS was advertised in other on-line forums (listserve, websites, social media, and discussion forums) and received almost 50% of the responses. However, detailed compositions of the participants and exact student group who completed the SAGS instrument was relatively unknown, thus to strongly justify the generalizability of the study.

Current study employed Rasch/IRT modeling approach. Even though the observed sample size was appropriate for Rasch modeling, experimenting with higher order IRT models have been conducted under a relatively small sample size. Efforts were made with a simulation study to mitigate the effect of small sample sizes required for IRT analysis. Thus the conclusion observed from the IRT study is under the constraint of simulated data. Further the Rasch analysis estimates were used with self-reported data to establish construct, predictive and convergent validities. Use of self-reported nature of the data without other validated measures of statistical competencies (cognitive or non –cognitive) can be considered as one of the minor limitations attributed to this study.

Future Research

The future goal of this could be considered as an initial step of creating a SAGS item inventory. Through the current study preliminary evidence for content validity, reliability, construct validity, predictive validity, and convergent validity were established, as well as item parameters were estimated. Future research will be directed to confirm them, and research activities can be classified into 3 major areas which falls under; (1) Improving of SAGS items

and expansion, (2) Additional testing for item parameter stability and validity, and (3) IRT methodological advances with Information theoretic model selection criteria.

Improving SAGS Items and Expansion

Future research should be directed initially to modify problematic items (stems and distractors) found though the data analysis during this study. Rasch analysis revealed the need of items with both lower difficulties as well items with higher difficulties, thus such items should be created while covering important concepts that were not covered in the initial SAGS instrument. When creating additional items, it would be beneficial to work towards developing a much larger bank (or inventory) of items that could include multiple items reflecting one statistical test or procedure. Item writing in the future research should be conducted with the collaboration of faculty who teach graduate level applied statistics courses. Future expert review processes could be conducted with the help of much larger expert panels, which could be created through sending open invitations to a larger statistics education community and asking them to participate in an expert panel.

Additional Testing for Item Parameter Stability and Validity

Administering the improved instrument to a larger population should be done to continuously to collect data. A larger sample will give rise to more accurate and stable parameter estimates. Also, a large sample will provide more validity to the confirmatory factor analysis that was conducted to establish the unidimensionality assumption. Analysis of a larger dataset will improve the confidence about the observed results and the respective conclusions made about the quality of items and reliability and validity evidence of the instrument. Also, future research will be designed in a way such that the SAGS administration reflects actual testing conditions. Thus, arrangements could be made with instructors to administer SAGS instrument in their classrooms.

Alternatively, further research could be conducted using Bayesian IRT approach with the available sample data to generate more stable and justifiable parameter estimates. With additional samples more powerful reliability and validity results could be established, and researchers could investigate the validity evidence by conducting Differential Item Functioning (DIF) studies which was not formally conducted in the current study. To address the limitation of using self-reported data, validity testing could be done through experimenting with other validated instruments for measuring statistics anxiety or statistics self-efficacy.

Methodological Advances with Information Theoretic Model Selection Criteria

Preliminary simulations conducted in the current study justifies the differential performance of various model selection criteria. Further, results show encouraging evidence on using ICOMP type model selection criteria for the selection of the best IRT model among a portfolio of models. To further enhance this results future studies will be conducted using randomly generated data from known IRT models. More comprehensive simulation studies will be designed with data generated from different IRT models (Rasch, 1 PL, 2 PL, and 3 PL), with different sample sizes (50, 100, 200, 500, 1000, and 10000). The performance of each criterion will be assessed through the numbers of times each model selection criteria correctly select target distribution in repeated runs (10,000) of the simulation.

Final Summary

The SAGS instrument was designed with the objective of measuring statistical research methodology knowledge of graduate students in education and other social and behavioral sciences. The present study demonstrated acceptable psychometric properties of SAGS, exhibiting convincing evidence for reliability and preliminary evidence for validity. In contrast to most of previous assessments, SAGS is targeted to the graduate-level population, and

specifically measure their ability to choose correct statistical test or procedure to solve applied research problems. The SAGS instrument was developed using an Rasch modeling approach, and its results have opened the avenue for creating customizable yet psychometrically sound assessments for measuring important components of statistical cognition. Therefore, the SAGS instrument has utility for the field of statistics education as an assessment, which could be used by students, faculty, and researchers.

LIST OF REFERENCES

Agre, P. (1997). *Computation and human experience*. Cambridge University Press.

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *American Psychologist*, *63*(1), 32.

Aiken, L. R. (1997). *Psychological testing and assessment.* Allyn & Bacon.

Alacaci, C. (2012). Towards a Pedagogy of Inferential Statistics in Graduate Education Programs: Insights from Cognitive and Educational Research.

Alanazy, S. M. (2011). *Research methods and statistical techniques employed by doctoral dissertations in education*. (Unpublished doctoral dissertation). Wayne State University, Detroit. Retrieved from http://search.proquest.com/docview/886421475.

Allen, K. (2006). The statistics concept inventory: Development and analysis of a cognitive assessment instrument in statistics (Doctoral dissertation). Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2130143.

Allen, K., Reed Rhoads, T., & Terry, R. (2006). Work in progress: Assessing student confidence of introductory statistics concepts. *Proceedings of the 36th ASEE/IEEE Frontiers in Education Conference.* S2E-13 - S2E-14.

Allen, K., Reed-Rhoads, T., Terry, R. A., Murphy, T. J., & Stone, A. D. (2008). Coefficient Alpha: An engineer's interpretation of test reliability. *Journal of Engineering Education, 97*(1), 87-94.

Allen, K., Stone, A., Reed-Rhoads, T., & Murphy, T. J. (2004). The statistics concepts inventory: Developing a valid and reliable instrument. *Proceedings of the American Society for Engineering Education Conference and Exposition.* Session 3230.

Allum, J. (2014). Graduate enrollment and degrees: 2003 to 2013. Washington, DC: Council of Graduate Schools.

Allum, J., & Okahana, H. (2015). Graduate enrollment and degrees: 2004 to 2014. Washington, DC: Council of Graduate Schools.

American Psychological Association. (2010). Ethical principles of psychologist and code of conduct. Retrieved from http://www.apa.org/ethics/code/

American Statistical Association. (n.d.), Curriculum guidelines for undergraduate programs in statistical science. Retrieved from http://www.amstat.org/education/ Curriculumguidelines.cfm.

AMST Newsletter (2015, April). Retrieved from http://magazine.amstat.org/wp-content/uploads/2015an/April_final.pdf

Anastasi, A. & Urbina, S. (2002). *Psychological testing Prentice Hall*: New York.

Astin, A. W. (2012). *Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education*. Rowman & Littlefield Publishers.

Atkinson, G., & Nevill, A. M. (1998). Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports medicine*, *26*(4), 217-238.

Baghaei, P. (2008). The Rasch model as a construct validation tool. *Rasch Measurement Transactions*, *22*(1), 1145-1146.

Baglin, J. (2014). Improving your exploratory factor analysis for ordinal data: a demonstration using FACTOR. *Practical Assessment, Research & Evaluation*,*19*(5), 2.

Baharun, N., & Porter, A. (2010, July). The impact of video-based resources in teaching statistics: A comparative study of undergraduates to postgraduates. In *Proceedings of the Eighth International Conference on Teaching Statistics*.

Baker, F. B. (2001). *The basics of item response theory*. For full text: http://ericae. net/irt/baker.

Baker, F. B., & Kim, S. H. (Eds.). (2004). *Item response theory: Parameter estimation techniques*. CRC Press

Barlow, P. B. (2014). Development of the biostatistics and clinical epidemiology skills assessment for medical residents. Retrieved from http://trace.tennessee.edu/utk_graddiss/ 2676.

Barnes, M. (2015). *Assessment 3.0: Throw out your grade book and inspire learning*. Corwin Press.

Bedeian, A. G., & Mossholder, K. W. (2000). On the use of the coefficient of variation as a measure of diversity. Organizational Research Methods, 3(3), 285-297.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, 18(1), 5-25.

Ben-Zvi, D., & Garfield, J. B. (Eds.). (2004). *The challenge of developing statistical literacy, reasoning and thinking*. Boston, MA: Kluwer Academic Publishers.

Berk, R. A., & Nanda, J. P. (1998). Effects of jocular instructional methods on attitudes, anxiety, and achievement in statistics courses.

Bessant, K. C. (1992). Instructional design and the development of statistical literacy. *Teaching Sociology*, *20*(2), 143-149.

Beurze, S. M., Donders, A. R. T., Zielhuis, G. A., de Vegt, F., & Verbeek, A. L. (2013). Statistics anxiety: a barrier for education in research methodology for medical students?. *Medical Science Educator*, *23*(3), 377-384.

Beyth-Marom, R., Fidler, F., & Cumming, G. (2008). Statistical cognition: Towards evidence-based practice in statistics and statistics education. *Statistics Education Research Journal*, *7*(2), 20-39.

Bidgood, P., Hunt, N., & Jolliffe, F. (2010). *Assessment Methods in Statistical Education*. John Wiley & Sons Inc.

Blalock Jr, H. M. (1987). Some general goals in teaching statistics. Teaching Sociology, 164-172.

Bloom, B. S. (1956). Taxonomy of educational objectives. *New York: David McKay*, *356*, 1998-1999.

Bond, T. G., & Fox, C. M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences*. Psychology Press.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Dordrecht, the Netherlands: Springer.

Bonett, D. G., & Price, R. M. (2005). Inferential methods for the tetrachoric correlation coefficient. Journal of Educational and Behavioral Statistics, 30(2), 213-225.

Bozdogan, H. (1990). On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models. *Communications in Statistics-Theory and Methods*, *19*(1), 221-278.

Bozdogan, H. (1994). Choosing the number of clusters, subset selection of variables, and outlier detection in the standard mixture-model cluster analysis. In *New approaches in classification and data analysis* (pp. 169-177). Springer Berlin Heidelberg.

Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. *Journal of mathematical psychology*, *44*(1), 62-91.

Bradley, K. D., Cunningham, J., Haines, T., Mueller, C. E., Royal, K. D., Sampson, S. O., ... & Weber, J. A. (2010). Constructing and evaluating measures: Applications of the Rasch measurement model.

Brainstrom. (n.d.). In *Wikipedia*. Retrieved October 14, 2009, from https://en.wikipedia.org/

Bremner, M. N., Blake, B. J., Long, J. M., & Yanosky, D. J. (2014). Setting a Benchmark for the Test of Essential Academic Skills (TEAS) V: Striving for First-Semester Success in Nursing School. Journal of Nursing Education, 53(9), 537.

Brown, C., Templin, J., & Cohen, A. (2014). Comparing the two-and three-parameter logistic models via likelihood ratio tests a commonly misunderstood problem. *Applied Psychological Measurement*, 0146621614563326.

Bryce, G. R., Gould, R., Notz, W. I., & Peck, R. L. (2001). Curriculum guidelines for bachelor of science degrees in statistical science. *The American Statistician*, *55*(1).

Bryce, G. R., Gould, R., Notz, W. I., & Peck, R. L. (2001). Curriculum guidelines for Bachelor of Science degrees in statistical science. *The American Statistician*, *55*(1), 7-13. doi:10.1198/000313001300339879

Camilli, G., & Shepard, L.A. (1994). Methods for identifying biased test items (vol. 4). Thousand Oaks, CA: Sage.

Cannon, A., Hartlaub, B., Lock, R., Notz, W., & Parker, M. (2002). Guidelines for undergraduate minors and concentrations in statistical science. *Journal of Statistics Education*, *10*(2). Retrieved from www.amstat.org/publications/jse/v10n2/cannon.html

Capraro, R. M., & Thompson, B. (2008). The educational researcher defined: What will future researchers be trained to do? *The Journal of Educational Research*, *101*(4), 247-253.

Carnegie Melon Eberly Center (n.d.). *Assessing prior knowledge*. Retrieved from http://www.cmu.edu/teaching/designteach/index.html.

Chan, S. W., & Ismail, Z. (2014). Developing Statistical Reasoning Assessment Instrument for High School Students in Descriptive Statistics. *Procedia-Social and Behavioral Sciences*, *116*, 4338-4343.

Chance, B. L. (2000). *Components of statistical thinking and implications for instruction and assessment*. ERIC Clearinghouse.

Chen, H. F., Lin, K. C., Wu, C. Y., & Chen, C. L. (2012). Rasch validation and predictive validity of the action research arm test in patients receiving stroke rehabilitation. *Archives of physical medicine and rehabilitation*, *93*(6), 1039-1045.

Chiesi, F., & Primi, C. (2010). Cognitive and con-cognitive factors related to students statistics achievement. *Educational Research Journal*, 45(3), 573-585. Retrieved from http://iase-web.org/documents/SERJ/SERJ9(1)_Chiesi_Primi.pdf.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Colton, D., & Covert, R. W. (2007). *Designing and constructing instruments for social research and evaluation*. John Wiley & Sons.

Conaghan, P. G., Emerton, M., & Tennant, A. (2007). Internal construct validity of the Oxford Knee Scale: evidence from Rasch measurement. *Arthritis Care & Research*, *57*(8), 1363-1367.

Consortium for Educational Research and Evaluation. (n.d). Retrieved from http://www.coastal.edu/education/cere/

Consortium for the Advancement of Undergraduate Statistics Education (n.d.). Retrieved from https://www.causeweb.org/reserach

Cook, K. F., Kallen, M. A., & Amtmann, D. (2009). Having a fit: impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. *Quality of Life Research*, *18*(4), 447-460.

Costa, J., & Judge, M. (2010). Calculating Geometric Means. Retrieved from http://buzzardsbay .org/geomean.htm

Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Curtis, D. A., & Harwell, M. (1998). Training doctoral students in educational statistics in the United States: A national survey. *Journal of Statistics Education*, *6*(1).

Dani, B. Z., & Joan, G. (2004). Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In *The challenge of developing statistical literacy, reasoning and thinking* (pp. 3-15). Springer Netherlands.

Davis, S. (2004). Statistics anxiety among female African American graduate-level social work students. Journal of Teaching in Social Work, 23(3-4), 143-158.

de Ayala, R.J. (2009). The Theory and Practice of Item Response Theory. New York: Guilford Press.

De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. Medical education, 44(1), 109–17. doi:10.1111/j.1365-2923.2009. 03425.x

Delmas, R. C. (2002). Statistical literacy, reasoning, and learning: A commentary. *Journal of Statistics Education*, *10*(3).

Delmas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, *6*(2), 28-58.

Delmas, R., Zieffler, A., & Garfield, J. (2012). Tertiary students' reasoning about samples and sampling variation in the context of a modeling and simulation approach to inference. *Educational Studies in Mathematics*.

Delucchi, M. (2014). Measuring Student Learning in Social Statistics A Pretest-Posttest Study of Knowledge Gain. *Teaching Sociology*, 0092055X14527909.

DeMars, C. (2010). *Item response theory*. Oxford University Press.

Deng, N., Wells, C., & Hambleton, R. (2008). A confirmatory factor analytic study examining the dimensionality of educational achievement tests.

Devore, J. (2015). *Probability and Statistics for Engineering and the Sciences*. Cengage Learning.

Dillon, K. M. (1982). Statisticophobia. Teaching of Psychology, 9(2), 117-117.

Dowdy, S., Wearden, S., & Chilko, D. (2011). *Statistics for research* (Vol. 512). John Wiley & Sons.

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two parameter logistic model. Applied Psychological Measurement, 13(77). doi:10.1177/014662168901300108

Dubois, J. E., & Gershon, N. (Eds.). (2013). *The information revolution: impact on science and technology*. Springer Science & Business Media.

Duncan, P. W., Bode, R. K., Lai, S. M., Perera, S., & Glycine Antagonist in Neuroprotection Americas Investigators. (2003). Rasch analysis of a new stroke-specific outcome scale: The stroke impact scale. *Archives of physical medicine and rehabilitation*, *84*(7), 950-963.

Dunn, D. S., Smith, R. A., & Beins, B. C. (Eds.). (2012). *Best practices in teaching statistics and research methods in the behavioral sciences*. Routledge.

Earl, L. M. (2012). *Assessment as learning: Using classroom assessment to maximize student learning*. Corwin Press.

Eastern Illinois University. (n.d.). *HST 3800 research methods II*. Retrieved from http://www.eiu.edu/healthst/hst_syllabi/3800.pdf

Ebel, R.L., & Frisbie, D.A. (1986). Essentials of educational measurement. Englewood Cliffs, NJ: Prentice-Hall.

Educational Testing Services. (2003). *Linking classroom assessment with student learning*. Retrieved from https://www.ets.org/Media/Tests/TOEFL_InstitutionalTesting_Program/

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press. ELLM2002.pdf.

Embretson, S. E., & Reise, S. P. (2013). Item response theory. Psychology Press.

Ennis, R. P., Lane, K. L., & Oakes, W. P. (2011). Score reliability and validity of the Student Risk Screening Scale: A psychometrically sound, feasible tool for use in urban elementary schools. *Journal of Emotional and Behavioral Disorders*,1063426611400082.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. Educational and psychological measurement, 58(3), 357-381.

Feinberg, L. B., & Halperin, S. (1978). Affective and cognitive correlates of course performance in introductory statistics. *The Journal of Experimental Education*, *46*(4), 11-18.

Feinberg, L. B., & Halperin, S. (1978). Affective and cognitive correlates of course performance in introductory statistics. *The Journal of Experimental Education*, *46*(4), 11-18.

Finney, S. J., & Schraw, G. (2003). Self-efficacy beliefs in college statistics courses. *Contemporary Educational Psychology*, *28*(2), 161-186.

Finney, S. J., & Schraw, G. (2003). Self-efficacy beliefs in college statistics courses. *Contemporary Educational Psychology*, *28*(2), 161-186.

Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., Klapp, B. F., & Rose, M. (2005). Development of a computer-adaptive test for depression (D-CAT). *Quality of life Research*, *14*(10), 2277-2291.

Fuchs, D., & Fuchs, L. S. (2006). Introduction to response to intervention: What, why, and how valid is it?. *Reading Research Quarterly*, *41*(1), 93-99.

Furr, R. M., & Bacharach, V. R. (2013). *Psychometrics: an introduction*. Sage.

Garcia, K. (2013). C*olorado's educational system of accreditation: characteristics of secondary schools on watch and reforms that are turning schools around.* (Unpublished doctoral dissertation). University of Colorado, Boulder. Retrieved from http://digital.auraria.edu/content/AA/00/00/01/21/00001/AA00000121_00001.pdf

Gardner, P. L., & Hudson, I. (1999). University students' ability to apply statistical procedures. *Journal of Statistics Education*, *7*(1).

Garfield, J. (1995). How students learn statistics. *International Statistical Review/Revue Internationale de Statistique*, 25-34.

Garfield, J. (1998a), *Challenges in Assessing Statistical Reasoning*, AERA Annual Meeting presentation, San Diego.

Garfield, J. (2002). The challenge of developing statistical reasoning. *Journal of Statistics Education*, *10*(3), 58-69.

Garfield, J. B. (1996). Assessing student learning in the context of evaluating a chance course. *Communications in Statistics--Theory and Methods*, *25*(11), 2863-2873.

Garfield, J. B. (1998b). The statistical reasoning assessment: Development and validation of a research tool. In *In the Proceedings of the 5 th International Conference on Teaching Statistics*.

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

Garfield, J. B., & Gal, I. (1999). Assessment and statistics education: Current challenges and directions. *International Statistical Review/Revue Internationale de Statistique*, 1-12.

Garfield, J., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice*. Springer Science & Business Media.

Garfield, J., & Chance, B. (2000). Assessment in statistics education: Issues and challenges. *Mathematical Thinking and Learning*, *2*(1-2), 99-125.

Garfield, J., & DelMas, R. (2010). A Web Site That Provides Resources for Assessing Students' Statistical Literacy, Reasoning and Thinking. *Teaching Statistics*, *32*(1), 2-7.

Garfield, J., & Franklin, C. (2011). Assessment of learning, for learning, and as learning in statistics education. In *Teaching Statistics in School Mathematics-Challenges for Teaching and Teacher Education* (pp. 133-145). Springer Netherlands.

Garfield, J., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM*, *44*(7), 883-898.

Garfield, J., delMas, R., & Chance, B. (2002). ARTIST: Assessment Resource Tools for Improving Statistical Thinking.

Gathercole, S. E., Pickering, S. J., Knight, C., & Stegmann, Z. (2004). Working memory skills and educational attainment: Evidence from national curriculum assessments at 7 and 14 years of age. Applied Cognitive Psychology, 18(1), 1-16.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis* (Vol. 2). London: Chapman & Hall/CRC.

Gibbons, K. S., & MacGillivray, H. (2014, January). Education for a workplace statistician. In Topics from Australian Conferences on Teaching Statistics (pp. 267-293). Springer New York.

Gilmore, J., & Feldon, D. (2010). Measuring Graduate Students' Teaching and Research Skills through Self-Report: Descriptive Findings and Validity Evidence. *Online Submission*.

Gold, A. U., & Harris, S. E. (2013, December). Measuring University students' understanding of the greenhouse effect-a comparison of multiple-choice, short answer and concept sketch assessment tools with respect to students' mental models. In *AGU Fall Meeting Abstracts* (Vol. 1, p. 04).

Gonsalvez, C. J., Bushnell, J., Blackman, R., Deane, F., Bliokas, V., Nicholson-Perry, K., ... & Knight, R. (2013). Assessment of psychology competencies in field placements: Standardized vignettes reduce rater bias. *Training and Education in Professional Psychology*, *7*(2), 99.

Gordon, S. (1995). A theoretical approach to understanding learners of statistics. *Journal of Statistics Education*, *3*(3), 1-21.

Gordon, S. (2004). Understanding students' experiences of statistics in a service course. *Statistics Education Research Journal*, *3*(1), 40-59.

George, D., & Mallery, P. (2003). SPSS for Windows step by step: A simple guide and reference. 11.0 update (4th ed.). Boston: Allyn & Bacon

Griffith, J. D., Adams, L. T., Gu, L. L., Hart, C. L., & Nichols-Whitehead, P. (2012). Students attitudes toward statistics across the disciplines: A mixed-methods approach. *Statistics Education Research Journal*, *11*(2), 45-46.

Groth, R. E., & Bergner, J. A. (2005). Pre-service elementary school teachers' metaphors for the concept of statistical sample. *Statistics education research Journal*, *4*(2), 27-42.

Haapala, A. (2002). How to overcome stumbling blocks in learning applied statistics-the effect of concept mapping.

Hand, D. J. (1984). Expert systems in statistics. *The Knowledge Engineering Review*, *1*(03), 2-10.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications* (Vol. 7). Springer Science & Business Media.

Hannigan, A., Hegarty, A. C., & McGrath, D. (2014). Attitudes towards statistics of graduate entry medical students: the role of prior learning experiences. *BMC medical education*, *14*(1), 70.

Hanushek, E. A., & Jackson, J. E. (2013). *Statistical methods for social scientists*. Academic Press.

Harris, R., & Jarvis, C. (2014). *Statistics for geography and environmental science*. Routledge.

Hasbrouck, J., & Tindal, G. A. (2006). Oral reading fluency norms: A valuable assessment tool for reading teachers. The Reading Teacher, 59(7), 636-644.

Hathaway, R. S., Nagda, B. A., & Gregerman, S. R. (2002). The relationship of undergraduate research participation to graduate and professional education pursuit: an empirical study. *Journal of College Student Development*, *43*(5), 614-631.

Hawkins, A., Jolliffe, F., & Glickman, L. (2014). *Teaching statistical concepts*. Routledge.

Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical care*, *38*(9 Suppl), II28.

Healey, J. (2014). *Statistics: A tool for social research*. Cengage Learning.

Heitman, E., Olsen, C. H., Anestidou, L., & Bulger, R. E. (2007). New graduate students' baseline knowledge of the responsible conduct of research. *Academic Medicine*, *82*(9), 838-845.

Henry, A. D. (2013). The Challenge of Statistics Education in Master of Public Administration Programs: A Review of Two Popular Textbooks. *Journal of Public Administration Research and Theory*, mut045.

Higazi, S. M. (2002). Teaching statistics using technology. *ICOTS6 Proceedings*.

Hirsch, L. S., & O'Donnell, A. M. (2001). Representativeness in statistical reasoning: Identifying and assessing misconceptions. *Journal of Statistics Education*, *9*(2), 61-82.

Holgado–Tello, F. P., Chacón–Moscoso, S., Barbero–García, I., & Vila–Abad, E. (2010). Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality & Quantity*, *44*(1), 153-166.

Horwitz, S., & Hoagwood, K. (2009). Developing questions when the perfect instrument is not available. *Stiffman, A.(2009). The field research survival guide*, 23-37.

Howe, E. D., Bozdogan, H., & Katragadda, S. (2011). Structural equation modeling (SEM) of categorical and mixed-data using the Novel Gifi transformations and information complexity (ICOMP) criterion.

Hsu, T. C. (2005). Research methods and data analysis procedures used by educational researchers. *International Journal of Research & Method in Education*, *28*(2), 109-133.

Hulsizer, M. R., & Woolf, L. M. (2009). *A guide to teaching statistics: Innovations and best practices* (Vol. 10). John Wiley & Sons.

Jackman, S. (2009). *Bayesian analysis for the social sciences* (Vol. 846). John Wiley & Sons.

Jala, L. L., & Reston, E. (2011). Graduate Students' Conceptions of Statistical Inference. Liceo Journal of Higher Education Research, 7(1).

Johnson, R. A., & Wichern, D. V. (2015). Applied multivariate statistical analysis. Statistics, 6215(10), 10.

Jonassen, D. H., & Grabowski, B. L. (2012). Handbook of individual differences, learning, and instruction. Routledge.

Joy, M. (2007). Research methods in education (No. 10). Innovation Way, York Science Park, Heslington, York YO10 5BR: The Higher Education Academy.

Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement*, *31*(4), 331-358.

KaradaÄŸ, E. (2010). An analysis of research methods and statistical techniques used by doctoral dissertation at the education sciences in Turkey. *Current Issues in Education, 13*(4).

Karelia, B. N., Pillai, A., & Vegada, B. N. (2013). The levels of difficulty and discrimination indices and relationship between them in four response type multiple choice questions of pharmacology summative tests of year II MBBS students. *Ie JSME2013, 6*, 41-46.

Keller, G. (2015). *Statistics for Management and Economics, Abbreviated*. Cengage Learning. Retrvied from http://www.cengage.com/search/productOverview.do;jsessionid=5332D2500B8CCFB261E3251166680BF3?N=16&Ntk=P_EPI&Ntt=138779270495543677111016822716339059 18&Ntx=mode%2Bmatchallpartial.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., et al. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, 68, 350-386.

Kirk, R. E. (1991). Statistical consulting in a university: Dealing with people and other challenges. *The American Statistician*, *45*(1), 28-34.

Kline, R. B., & Santor, D. A. (1999). [Principles & Practice of Structural Equation Modelling]. *Canadian Psychology*, *40*(4), 381.

Kolenikov, S. (2000). Icomp: Information complexity measures. Retrieved from http://fmwww.bc.edu/repec/bocode/i/icomp.pdf

Knypstra, S. (2009). Teaching statistics in an activity encouraging format. *Journal of Statistics Education*, *17*(2), n2.

Konold, C. (1989). Informal conceptions of probability. *Cognition and instruction*, *6*(1), 59-98.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into practice*, *41*(4), 212-218.

Kuh, G. D., Jankowski, N., Ikenberry, S. O., & Kinzie, J. (2014). Knowing what students know and can do: The current state of student learning outcomes assessment in US colleges and universities. Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment (NILOA).

Larwin, K., & Larwin, D. (2011). A meta-analysis examining the impact of computer-assisted instruction on postsecondary statistics education: 40 years of research. Journal of Research on Technology in Education, 43(3), 253-278.

Learn and Teach Statistics and Operations Research (2013). Teaching service course in statistics. Retrieved from https://learnandteachstatistics.wordpress.com/ 2013/05/06/service-statistics/

Lee-Ellis, S. (2009). The development and validation of a Korean C-Test using Rasch Analysis. *Language Testing*, *26*(2), 245-274.

Lehmann, A. C. (2014). Using Admission Assessments to Predict Final Grades in a College Music Program. *Journal of Research in Music Education*, 0022429414542654.

Linacre, J.M. (2016). Winsteps® (Version 3.92.0) [Computer Software]. Beaverton, Oregon: Winsteps.com. Retrieved January 1, 2016. Available from http://www.winsteps.com/

Linacre, M. (n.d.). *Rasch - Winsteps - Facets online Rasch tutorial PDF*s. Retrieved from http://www.winsteps.com/tutorials.htm.

Linacre, J. M. (1994). Sample Size and Item Calibration Stability “, Rasch Measurement Transactions, 7 (4), 328. Retrieved from *http:// www.rasch.org/rmt/rmt74m.htm*.

Liu, H. J. C. (1998). *A cross-cultural study of sex differences in statistical reasoning for college students in Taiwan and the United States*. (Unpublished doctoral dissertation), University of Minnesota, Minneapolis.

LoBiondo-Wood, G., & Haber, J. (2014). *Nursing research: Methods and critical appraisal for evidence-based practice*. Elsevier Health Sciences.

Lord, F. M. (1953). *On the Statistical Treatment of Football Numbers*. American Psychologist,8, 750-751.

Lord, F. M., & Novick, M.R. (1968). *Statistical theories of metal test scores*. Reading, MA:Addison-Wesley.

Macher, D., Paechter, M., Papousek, I., & Ruggeri, K. (2012). Statistics anxiety, trait anxiety, learning behavior, and academic performance. *European journal of psychology of education*, *27*(4), 483-498.

Marino, M. J. (2014). The use and misuse of statistical methodologies in pharmacology research. *Biochemical pharmacology*, *87*(1), 78-92.

Marusteri, M., & Bacarea, V. (2010). Comparing groups for statistical differences: how to choose the right statistical test?. *Biochemia medica*, *20*(1), 15-32.

Mathews, D., & Clark, J. (1997). Successful students' conceptions of mean, standard deviation, and the Central Limit Theorem. In *Midwest Conference on Teaching Statistics, Oshkosh, WI*. Retrieved from https://www.researchgate.net/profile/David_Mathews4/publication/253438034_Successful_Students%27_Conceptions_of_Mean_Standard_Deviation_and_The_Central_Limit_Theorem/links/54e23d6b0cf2c3e7d2d335af.pdf

McGahee, T. W., & Ball, J. (2009). How to read and really use an item analysis. *Nurse Educator*, *34*(4), 166-171.

Meerah, T. S. M., Osman, K., Zakaria, E., Ikhsan, Z. H., Krish, P., Lian, D. K. C., & Mahmod, D. (2012). Measuring Graduate Students Research Skills. *Procedia-Social and Behavioral Sciences*, *60*, 626-629.

Montgomery, D. C., & Runger, G. C. (2013). *Applied statistics and probability for engineers*. Hoboken, NJ: John Wiley & Sons.

Monthly Labor Review (n.d). Employment and total job openings, by education category, 2010 and projected 2020 and median annual wage. Retrieved from http://www.bls.gov/opub/mlr/2012/01/mlr201201.pdf.

Moore, D. S. (2001). Undergraduate programs and the future of academic statistics. *The American Statistician*, *55*(1), 1-6. doi:10.1198/000313001300339860

Morrow, J.A. (2013). Lecture #2: Measurement and validity [Lecture notes]. Retrieved from University of Tennessee, Knoxville, EDPY 583 Blackboard site.

Morrow, J.A., & Skolits, G. (2015, October). The twelve steps of data cleaning: Strategies for dealing with dirty data. Full-day workshop presented at the annual meeting of the American Evaluation Association, Chicago, IL.

Murdoch, T. B., & Detsky, A. S. (2013). The inevitable application of big data to health care. Jama, 309(13), 1351-1352.

Myers, N. D., Ahn, S., & Jin, Y. (2011). Sample size and power estimates for a confirmatory factor analytic model in exercise and sport: A Monte Carlo approach. Research quarterly for exercise and sport, 82(3), 412-423.49.

Newsom, J. T. (2012). Some clarifications and recommendations on fit indices. Retrieved from http://web.pdx.edu/~newsomj/semclass/ho_fit.pdf

Noether, G. E. (2012). *Introduction to statistics: the nonparametric way*. Springer Science & Business Media.

Onwuegbuzie, A. J. (1998). The dimensions of statistics anxiety: A comparison of prevalence rates among mid-southern university students. *Louisiana Educational Research Journal*, *23*(2), 23-40.

Onwuegbuzie, A. J. (1999). Statistics anxiety among African American graduate students: An affective filter?. Journal of Black Psychology, 25(2), 189-209.

Onwuegbuzie, A. J. (2002a). Common analytical and interpretational errors in educational research: an analysis of the 1998 volume of the British Journal of Educational Psychology. *Educational Research Quarterly*, 26, 11-22.

Onwuegbuzie, A. J. (2002b). Why can't we all get along? Towards a framework for unifying research paradigms. *Education*, *122*(3), 518.

Onwuegbuzie, A. J. (2003). Modeling statistics achievement among graduate students. *Educational and Psychological Measurement*, *63*(6), 1020-1038. doi:10.1177/0013164402250989.

Onwuegbuzie, A. J., & Wilson, V. A. (2003). Statistics Anxiety: Nature, etiology, antecedents, effects, and treatments--a comprehensive review of the literature. *Teaching in Higher Education*, 8(2), 195-209.

Onwuegbuzie, A. J., Da Ros, D., & Ryan, J. M. (1997). The Components of Statistics Anxiety: A Phenomenological Study. *Focus on Learning Problems in Mathematics*, 19(4), 11-35.

Pagano, R. (2006). *Understanding statistics in the behavioral sciences*. Cengage Learning.

Pan, W., & Tang, M. (2004). Examining the effectiveness of innovative instructional methods on reducing statistics anxiety for graduate students in the social sciences. *Journal of Instructional Psychology*, *31*(2), 149.

Pande, S. S., Pande, S. R., Parate, V. R., Nikam, A. P., & Agrekar, S. H. (2013). Correlation between difficulty & discrimination indices of MCQs in formative exam in Physiolog. *South East Asian Journal of Medical Education 2013*, 7, 45-50.

Partchev, I. (2004). A visual guide to item response theory. *Friedrich Schiller Universität Jena*. pdfs/BA-curriculum.pdf

Pathirage, D.P.N.A. (2015). The development and validation of the self-efficacy in statistical practices scale. (Unpublished Doctoral dissertation). University of Tennessee, Knoxville, Knoxville.

Peck, R., Olsen, C., & Devore, J. (2015). *Introduction to statistics and data analysis*. Cengage Learning.

Peiris, S., & Beh, E. J. (2012). Where statistics teaching can go wrong. *International Journal of Innovation in Science and Mathematics Education (formerly CAL-laborate International), 15*(1).

Perepiczka, M., Chandler, N., & Becerra, M. (2011). Relationship between graduate students' statistics self-efficacy, statistics anxiety, attitude toward statistics, and social support. *The Professional Counselor: Research and Practice, 1*(2), 99-108.

Pfannkuch, M., & Wild, C. (2004). Towards an understanding of statistical thinking. In *The challenge of developing statistical literacy, reasoning and thinking* (pp. 17-46). Springer Netherlands.

Powers, R. M., & Powers, T. Y. (2009). A "Coefficient of Variation" for Skewed and Heavy-Tailed Insurance Losses. *Temple University*.

Quilici, J. L., & Mayer, R. E. (1996). Role of examples in how students learn to categorize statistics word problems. *Journal of Educational Psychology*, *88*(1), 144.

Rasiah, S. M. S., & Isaiah, R. (2006). Relationship between item difficulty and discrimination indices in true/false-type multiple choice questions of a para-clinical multidisciplinary paper. *Annals Academy of Medicine Singapore*, *35*(2), 67-71.

Raudenbush, S. W. (2005). Learning from attempts to improve schooling: The contribution of methodological diversity. *Educational Researcher*, *34*(5), 25-31.

Reed, G. F., Lynn, F., & Meade, B. D. (2002). Use of coefficient of variation in assessing variability of quantitative assays. *Clinical and diagnostic laboratory immunology*, *9*(6), 1235-1239.

Reid, A., & Petocz, P. (2002). Students' conceptions of statistics: A phenomenographic study. *Journal of Statistics Education*, *10*(2), 1-18.

Ritter, M. A., Starbuck, R. R., & Hogg, R. V. (2001). Advice from prospective employers on training BS statisticians. *The American Statistician*, *55*(1), 14-18. doi:10.1198/000313001300339888

Roberts, D. M., & Bilderback, E. W. (1980). Reliability and validity of a statistics attitude survey. *Educational and Psychological Measurement*, *40*(1), 235-238.

Rudestam, K. E., & Newton, R. R. (2014). Surviving your dissertation: A comprehensive guide to content and process. Sage Publications.

Rumsey, D. J. (2002). Statistical literacy as a goal for introductory statistics courses. *Journal of Statistics Education*, *10*(3), 6-13.

Sahin, F., (2012). A study for development of statistical literacy scale for undergraduate students. Unpublished master's thesis, Boğaziçi University.

Samuels, M. L., Witmer, J. A., & Schaffner, A. (2012). *Statistics for the life sciences*. Pearson education.

Schau, C., Stevens, J., Dauphinee, T. L., & Del Vecchio, A. (1995). The Development and Validation of the Survey of Antitudes toward Statistics. *Educational and psychological measurement*, *55*(5), 868-875.

Schield, M. (2006). Statistical literacy survey analysis: Reading graphs and tables of rates and percentages. In *Proceedings of the Sixth International Conference on Teaching Statistics*.

Schield, M. (2008). Statistical Literacy Skills Survey. *Project Kaleidoscope and Project Quirk.* Retrieved from http://web.augsburg.edu/~schield/MiloPapers/StatLitKnowledge2r.pdf

Schield, M. (2010). Assessing statistical literacy: Take CARE. *Assessment Methods in Statistical Education*, 133.

Schmidhammer (n.d.). IGSP. Retrieved from http://igsp.bus.utk.edu/

Seaver, W. (2013). Lecture #3: PCA and Factor Analysis [Lecture notes]. Retrieved from University of Tennessee, Knoxville, STAR 579 Blackboard site.

SECU (n.d.). Retrieved from http://www.thesecu.com/

Shaw, F. (1991). Descriptive IRT vs. prescriptive Rasch. *Rasch Measurement*, *5*(1), 131.

Sherry, A., & Henson, R. K. (2005). Conducting and interpreting canonical correlation analysis in personality research: A user-friendly primer. *Journal of personality assessment*, *84*(1), 37-48.

Singh, S. (2014). Test Item Analysis and Relationship Between Difficulty Level and Discrimination Index of Test Items in an Achievement Test in Biology. Indian Journal of Research, 3(6), 56-58.

Society for the Teaching of psychology. (n.d). New teaching resource. Retrieved from http://teachpsych.org/page-1599567/1539254

Sommers, D., & Franklin, J.C. (2012). An overview of the 10-year projections of the U.S. macroeconomy, labor force, industry output and employment, and occupational employment. *Monthly Labor Review*,135(1), 3-20.

Sørensen, J. B. (2002). The use and misuse of the coefficient of variation in organizational demography research. *Sociological methods & research*, *30*(4), 475-491.

Statistics Canada (n.d). Examples of questions for the knowledge and ability test for Mathematical Statisticians. Retrieved from http://www.statcan.gc.ca/eng/employment/

Stage, C. (1998). *A Comparison Between Item Analysis Based on Item Response Theory and on Classical Test Theory: A Study of the SweSAT Subtest ERC*. Department of Educational measurement, Umeå Univ.

Stallings, W. M., West, C. K., & Carmody, C. (1983). The quality of research articles in the Journal of Educational Research, 1970 and 1980. *The Journal of Educational Research*, *77*(2), 70-76.

STATtr@K (2012). The world of applied statistics: where do you fit in. Retrieved from http://stattrak.amstat.org/2012/03/01/applied-statistics/

Stewart-Brown, S., Tennant, A., Tennant, R., Platt, S., Parkinson, J., & Weich, S. (2009). Internal construct validity of the Warwick-Edinburgh mental well-being scale (WEMWBS): a Rasch analysis using data from the Scottish health education population survey. *Health and Quality of Life Outcomes*, *7*(1), 15-22

Stiggins, R. (2005). From formative assessment to assessment for learning: A path to success in standards-based schools. Phi Delta Kappan, 324-328.

Stone, A. (2006). A psychometric analysis of the statistics concept inventory. (Unpublished doctoral Dissertation. University of Oklahoma, Norman, Oklahoma.

Stone, A., Allen, K., Reed Rhoads, T., Murphy, T. J., Shehab, R. L., & Saha, C. (2003). The statistics concept inventory: A pilot study. *Proceedings of the 33rd ASEE/IEEE Frontiers in Education Conference.* T3D-1 - T3D-6.

Stone, A., Allen, K., Reed Rhoads, T., Murphy, T. J., Shehab, R. L., & Saha, C. (2003). The statistics concept inventory: A pilot study. *Proceedings of the 33rd ASEE/IEEE Frontiers in Education Conference.* T3D-1 - T3D-6.

Sundre, D. L. (2003, April). Assessment of quantitative reasoning to enhance educational quality. In *American Educational Research Association Meeting, Chicago, IL*.

Suskie, L. (2009). Using assessment results to inform teaching practice and promote lasting learning. In *Assessment, Learning and Judgement in Higher Education* (pp. 1-20). Springer Netherlands.

Taplin, R. H. (2003). Teaching statistical consulting before statistical methodology. *Australian & New Zealand Journal of Statistics*, *45*(2), 141-152.

Tarpey, T., Acuna, C., Cobb, G., & De Veaux, R. (2000). Curriculum guidelines for bachelor of arts degrees in statistical science. Retrieved from http://www.amstat.org/education/

Tempelaar, D. (2004). Statistical reasoning assessment: An Analysis of the SRA instrument.

Templin, Jonathan. "Basic IRT concepts, models, and assumptions." Lecture, Inter-university Consortium for Political and Social Research, Summer Workshop 2014. Ann Arbor, MI, June 30 18, 2014.

The University of Minnesota. (n.d.). Evaluation and assessment of teaching and learning about statistics (e-ATLAS). Retrieved from http://www.tc.umn.edu/~eatlas/.

Thiese, M. S., Arnold, Z. C., & Walker, S. D. (2015). The misuse and abuse of statistics in biomedical research. *Biochemia medica*, *25*(1), 5-11.

Thorndike, R. M., (2004). *Measurement and Evaluation in Psychology and Education* (7th Ed.). Upper Saddle River, NJ: Prentice Hall.

Tishkovskaya, S., & Lancaster, G. A. (2010, July). Teaching strategies to promote statistical literacy: review and implementation. In *Proceedings of the 8th International Conference on Teaching Statistics, 11-16 July, Ljubljana, Slovenia*.

Trochim, W. M. (2006). Descriptive Statistics. Retrieved from http://www.socialresearchmethods.net/ kb/statdesc.php

Uebersax, J. S. (2006). The tetrachoric and polychoric correlation coefficients.*Statistical methods for rater agreement web site*.

University of Memphis. (n.d.). Research design comprehensive exam questions. Retrieved from http://www.memphis.edu/lead/pdfs/research-design-comprehensive-exam-question.pdf

University of Pittsburgh. (2013). Research methodology research design comprehensive exam study guide. Retrieved from http://www.education.pitt.edu/Portals/0/ Academic%20Departments/PSYED/RM/RM_ResearchDesignStudyGuide.pdf

University of Tennessee, Knoxville. (n.d.). Anthropological statistics II. Retrieved from http://web.utk.edu/~auerbach/ANTH604.pdf

University of Wisconsin (2016). *Testing services: item discrimination I*. Retrieved from http://www.uwosh.edu/testing/faculty-information/test-scoring/score-report-interpretation/item-analysis-1/item-i

University of York (n.d.). Measuring heath and disease. Retrieved from https://www-users.york.ac.uk/~mb55/msc/clinimet/week8/validity.pdf.

Van Abswoude, A. A., van der Ark, L. A., & Sijtsma, K. (2004). A comparative study of test data dimensionality assessment procedures under nonparametric IRT models. *Applied Psychological Measurement*, *28*(1), 3-24.

van der Linden, W. J., & Hambleton, R. K. (Eds.). (2013). *Handbook of modern item response theory*. Springer Science & Business Media

Vance, E. A. (2015). Recent developments and their implications for the future of academic statistical consulting centers. *The American Statistician*, *69*(2), 127-137.

Vanhoof, S., Sotos, A. E. C., Onghena, P., Verschaffel, L., Van Dooren, W., & Van den Noortgate, W. (2006). Attitudes toward statistics and their relationship with short-and long-term exam results. *Journal of Statistics Education*, *14*(3). Retrieved from www.amstat.org/publications/jse/v14n3/vanhoof.html.

Waller, J., Ostini, R., Marlow, L. A., McCaffery, K., & Zimet, G. (2013). Validation of a measure of knowledge about human papillomavirus (HPV) using item response theory and classical test theory. *Preventive medicine*, *56*(1), 35-40.

Welch, P. S., Jacks, M. E., Smiley, L. A., Walden, C. E., Clark, W. D., & Nguyen, C. A. (2015). A Study of Statistics Anxiety Levels of Graduate Dental Hygiene Students. *American Dental Hygienists Association*, *89*(1), 46-54.

Will, M. C. (1986). Educating children with learning problems: A shared responsibility. Exceptional children, 52(5), 411-415.

Williams, A. S. (2010). Statistics anxiety and instructor immediacy. *Journal of Statistics Education*, *18*(2), 1-18.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, *9*(1), 60-62.

Winsteps (n.d.). Retrieved from http://www.winsteps.com/winman/validity.htm

Wise, S. L. (1985). The development and validation of a scale measuring attitudes toward statistics. *Educational and Psychological Measurement*, *45*(2), 401-405.

Wright, B. D. (1992). IRT in the 1990s: Which models work best. *Rasch measurement transactions*, *6*(1), 196-200.

Wright, B. D., & Stone, M. H. (1979). Best Test Design. Rasch Measurement. Chicago: Mesa Press.

Wright, D. B., & London, K. (2009). *Modern regression techniques using R: A practical guide*. Sage.

Wright, R. J. (2008). *Educational assessment: Tests and measurements in the age of accountability*. Sage Publications.

Yen, W. M. (1981). Using Simulation Results to Choose a Latent Trait Model. Applied Psychological Measurement, 5(2), 245–262. doi:10.1177/014662168100500212

Yilmaz, M. R. (1996). The challenge of teaching statistics to non-specialists. *Journal of statistics education*, *4*(1), 1-9.

Yu, C. H. (2003). Resampling methods: concepts, applications, and justification. *Practical Assessment, Research & Evaluation*, *8*(19), 1-23.

Yu, C. H., Popp, S. O., DiGangi, S., & Jannasch-Pennell, A. (2007). Assessing unidimensionality: A comparison of Rasch modeling, parallel analysis, and TETRAD. *Practical Assessment, Research & Evaluation*, *12*(14), 1-18.

Zanakis, S. H., & Valenzi, E. R. (1997). Student anxiety and attitudes in business statistics. *Journal of Education for Business*, *73*(1), 10-16.

Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. Psychometrika, 64(2), 213-249.

Zeidner, M. (1991). Statistics and mathematics anxiety in social science students: Some interesting parallels. *British Journal of Educational Psychology*, *61*(3), 319-328.

Zeph, C. P. (1991). Graduate study as professional development. *New Directions for Adult and Continuing Education*, *1991*(51), 79-88.

Zieffler, A., Garfield, J., Delmas, R., & Reading, C. (2008). A framework to support research on informal inferential reasoning. *Statistics Education Research Journal*, *7*(2), 40-58.

Ziegler, L. A. (2014). *Reconceptualizing statistical literacy: developing an assessment for the modern introductory statistics course* (Unpublished doctoral dissertation). University of Minnesota, Minnesota. Retrieved from http://conservancy.umn.edu/bitstream/handle/11299/165153/Ziegler_umn_0130E_15130.pdf?sequence=1&isAllowed=y.

APPENDICES

Appendix A
Statistics Course Descriptions – South Eastern Conference

Table A.1. Course Descriptions.

University/ College	***Courses and Descriptions***
University of Georgia College of Education Educational Research and Measurement	**ERSH 6300 Applied Statistical Methods in Education** Techniques for describing and summarizing data for educational research studies. Applications of the standard normal distribution and the use and interpretation of standard scores. Inferential statistics for one and two population studies including means, proportions, and correlations **ERSH 8310 Applied Analysis of Variance Methods in Education** Experimental design and the analysis of data from experiments, including orthogonal analysis of variance for single and multifactor designs, randomized block, repeated measures, and mixed models. Computer applications and reporting results using APA style. **ERSH 8320 Applied Correlation and Regression Methods in Education** Non-experimental and quasi-experimental research studies, including simple and multiple regression techniques, non-orthogonal analysis of variances, correlation techniques, and analysis of covariance. Other Courses: **ERSH 8350 Multivariate Methods in Education** **ERSH 8360 Categorical Data Analysis in Education**
University of Florida College of Education Education: Found. & Policy	**EDF 7403 Quantitative Foundations of Educational Research** Examination of appropriate methods in applied educational contexts. Consideration of analysis strategies for educational data, emphasis on identification and interpretation of findings. **EDF 7405 Quantitative Methods II** Correlation, regression, path analysis, and structural equation modeling in educational studies. Use of path analysis and structural equation modeling to test theory

Table A.1. *(Continued)*

University/ College	***Courses and Descriptions***
University of Florida College of Education Education: Found. & Policy	**EDF 7406 Multivariate Statistics in Education** Statistical methods that simultaneously analyze multiple measurements on an individual or object under investigation
University of South Carolina College of Education Educational Research Methods	**EDRM 710 Educational Statistics I** Introductory course in statistics for graduate students in education and the other social sciences. Central tendency and variability, normal distribution, simple correlation and regression, z and t tests for one and two samples, and the chi-square test. Use of statistical software. **EDRM 711 Educational Statistics II** Continuation of Educational Statistics I. Inference for one and two samples, factorial designs, repeated measures designs, and multiple regressions. Use of statistical software.
University of Tennessee, Knoxville Education, Health and Human Sciences Educational Psychology	**EDPY 577 Statistics in Applied Fields I** Applications of descriptive and inferential statistics to problems in applied fields. Use of internet sites and computer programs to analyze data. **EDPY 677 Statistics in Applied Fields II** Application of intermediate statistical procedures (e.g., factorial analysis of variance, analysis of covariance, multiple regression, multivariate analysis of variance) via statistical package. **EDPY 678 Statistics in Applied Fields III** Techniques in advanced multivariate statistics will be reviewed. Reviewing literature on topics such as logistic regression, multilevel modeling, structural equation modeling, and factor analysis, as well as learning how to conduct these analyses using statistical software will be covered. Other Courses: **EDPY 550 Applied Statistical Concepts**

Table A.1. *(Continued)*

University/ College	***Courses and Descriptions***
University of Kentucky College of Education Educational & Counseling Psychology	**EDP 557 Gathering, Analyzing, And Using Educational Data** The course covers applications of statistical and graphical methods for educational and evaluation data. Basic descriptive statistics, correlation, normal distributions and hypothesis testing will be covered. An emphasis is placed on exploratory data analysis and interpretation of results within the broad contexts of education and evaluation **EDP 660 Research Design and Analysis in Education** This is a statistics-oriented course that focuses on various aspects of regression analysis. Topics to be covered include, but are not limi ted to, simple correlation and regression, multiple regression (with or without interaction terms), regression diagnostics, logistic regression, etc. The course aims to familiarize students with cleaning data for regression analysis, building regression models, selecting the optimal regression model for the data in hand, gain requisite foundation of knowledge necessary to learn more complex statistical tests and procedures, and become more critical of statistical presentations in academic journals and the mass media **EDP 707 Multivariate Analysis in Educational Research** Multivariate statistics will prepare student to understand multivariate statistical methods and draw the link between statistics previously learned. Students will be able to conduct, interpret, and critique procedures such as factorial ANOVA, multiple regression, MANOVA, ANCOVA, MANCOVA, PCA, EFA, discriminant function analysis, logistic regression, canonical correlation, hierarchal linear regression, and multivariate analysis of change. Become familiar with statistical software for implementing multivariate procedures. Develop an understanding of the concepts, terms, and symbols used in multivariate statistics (e.g., Matrix Algebra, effect sizes). Gain an appreciation of the role of multivariate procedures in the research process. Gain requisite knowledge necessary to learn more complex statistical procedures. Other Courses: **EDP 558 Gathering, Analyzing, and Using Educational Data II**

Table A.1. *(Continued)*

University/ College	***Courses and Descriptions***
University of Missouri College of Education Educational, School and Counseling Psychology	**ESC_PS 8830: Quantitative Analysis in Educational Research I** This is the first course in the sequence of statistical analysis methods. Topics covered in this course include statistical inference, simple regression, multiple regression, regression assumptions, regression with categorical predictors, model selection methods polynomial regression, and model validation. **ESC_PS 8830: Quantitative Analysis in Educational Research I** This is the first course in the sequence of statistical analysis methods. Topics covered in this course include statistical inference, simple regression, multiple regression, regression assumptions, regression with categorical predictors, model selection methods polynomial regression, and model validation. **ESC_PS 9650: Application of Multivariate Analysis in Educational Research** The focus of this course will be on applications of multivariate analysis in educational research. Other Courses: **ESC_PS 7170 Introduction to Applied Statistics** **ESC_PS 9710 Structural Equation Modeling** **ESC_PS 9720 Hierarchical Linear Modeling**
University Of Arkansas College of Education and Health Professions Educational Statistics and Research Methods	**ESRM 6403: Educational Statistics and Data Processing** Theory and application of frequency distributions, graphical methods, central tendency, variability, simple regression and correlation indexes, chi-square, sampling, and parameter estimation, and hypothesis testing. Use of the computer for the organization, reduction, and analysis of data

Table A.1. *(Continued)*

University/ College	***Courses and Descriptions***
University Of Arkansas College of Education and Health Professions Educational Statistics and Research Methods	**ESRM 6423: Multiple Regression Techniques for Education** Introduction to multiple regression procedures for analyzing data as applied in educational settings, including multicollearity, dummy variables, analysis of covariance, curvi-linear regression, and path analysis **ESRM 6453. Applied Multivariate Statistics** Multivariate statistical procedures as applied to educational research settings including discriminant analysis, principal components analysis, factor analysis, canonical correlation, and cluster analysis. Emphasis on use of existing computer statistical packages.

Appendix B
Commonly Statistical Procedures/ Tests in Educational Research

Table B.1. Commonly Used Statistical Procedures.

Source	***Statistical Techniques***	
Mubarak (2011) Time Period: 2008 – 2010 (Dissertations : 110)	Descriptive Statistics ANOVA Multiple Regression	Bi-variate correlations T-test
Hsu (2005) Time Period: 1971-1998 In Articles: AERJ: American Educational Research Journal (713), JEE: Journal of Experimental Education (638), JER: Journal of Educational Research (875)	Descriptive ANOVA/ANCOVA(MANOVA) Correlation (canonical correlation) MANOVA	Regression (Log-linear logistics Modeling, structural equations) t-test non-parametric (chi-square)
Karadağ (2010) Time Period: 2003-2007 Dissertations: 2011 Un-published doctoral dissertation research in Turkey	Descriptive statistics ANOVA Bivariate correlation Kruskal Wallis-H Test Chi-Score Test Multiple regression Kormogrov Smirnov Path analysis	t-test Factor analysis Many Whitney-U Test ANCOVA Wilcoxon Linear regression MANOVA
Onwuegbuzie (2002) Time period: 1971-1998 BJEP: British Journal of Education Psychology	ANOVA MANOVA Factor analysis	Correlation Regression chi-sq.

Table B.1. *(Continued)*

Source	***Statistical Techniques***	
Curtis & Harwell (1998) Time Period: 1995-1996 National Survey of 30 top ranking education schools. Only included topics that coves by more than two third of schools	**ANOVA** (Covariance analysis, Repeated measures designs, Power/sample size calculations, Mixed-effects models, Random-effects models, Non-orthogonal designs, Thorough coverage of multiple comparison procedures, Cell means models, Complex designs) **Traditional Multivariate Procedures** (Canonical correlation, MANOVA, Discriminant analysis, Factor analysis MANCOVA, Principal components analysis)	**Multiple Regressions** (Ordinary least squares estimation, weighted least squares estimation, Nonlinear-in-the-predictors models, Logistic regression, Nonlinear-in-the-parameters models) **Other Topics and Procedures** (Matrix algebra, Meta-analysis, Structural equation models)

Appendix C
Statistics Assessment of Graduate Students (SAGS) Instrument

Applied Statistics Knowledge Self-Assessment

Today you are being asked to participate in an ongoing research project conducted by Dammika Lakmal Walpitage, a Ph.D candidate in Evaluation, Statistics, & Measurement at the University of Tennessee, Knoxville.

For this study, you will be asked to answer several multiple-choice questions about your knowledge of statistical applications. Your participation in this project is completely voluntary, and you may stop participating at any point. Please take the next few minutes to answer the following questions. If you are uncertain about the answer to a particular question, just give your best guess. Although more than one response option may be plausible, each question has a *single best answer based on the information provided.*

Your responses and total scores on this assessment will remain confidential, and your participation in today's study will not affect your standing with your institution in any way. After you have completed this self-assessment, you will receive a copy of the answer booklet, so that you may use it as a study guide in the future.

Thank you for your participation!

1. A researcher is interested in examining the college students' level of physical activity. Using a survey, which asked individuals to report the amount of time (in minutes) doing physical activity, the researcher collected data from several individuals. Looking at the data he noticed that there were a few students who had spent considerably more time doing physical activity than others.

Which summary statistic should be used?

(a) Mean
(b) Median
(c) Mode
(d) Kurtosis

2. A graduate student wanted to compare the instability of scores of two standardized math tests. He administered both math tests to a simple random sample of 5th grade students from a local elementary school. One test is scored in the range of 200 through 500 while the other is scored in the range of 300 through 800.

Which summary statistic should be used?

(a) Range
(b) Standard deviation
(c) Coefficient of variation
(d) Interquartile range

3. A teacher wanted to compare a particular student's math and reading ability relative to class peers. The teacher used scores from the last math and the last reading assessment of the semester to conduct the analysis. Different numbers of students have taken the math and reading assessments.

Which summary statistic should be used?

(a) Student's rank on the math and reading tests
(b) Student's scores on the math and reading tests
(c) Class means on the math and reading tests
(d) Student's percentile ranks on the math and reading tests

4. A Dean of the College of Engineering wanted to determine the association between students' ACT scores and their GPAs after their first semester of college. Using the data retrieved through the college's student information system, the Dean made a graphical summary and found that the association is non-linear. The Dean was then interested in identifying the direction and magnitude of the association between the students' ACT scores and GPAs.
Which correlation coefficient should be used?

(a) Pearson product-moment correlation coefficient
(b) Spearman rank correlation coefficient
(c) Kendall tau rank correlation coefficient
(d) Phi correlation coefficient

5. A college administrator claims that the incoming freshmen at a public university are better critical thinkers, on average, than incoming freshmen at the national level. The national average score for incoming college freshman on the California Critical Thinking Skills Test (CCTST) is 65 out of 100. The college administrator selects a random sample of 50 students from the incoming class of 4,700 and administers the CCTST. The average score for the 50 students is 67 with a standard deviation of 10. The college administrator wants to statistically support the claim.

Which statistical test should be used?

(a) One sample t-test
(b) Independent samples t-test
(c) Dependent samples t-test
(d) Independent samples z-test

6. An after school program administrator wants to evaluate the effectiveness of a new eating disorder prevention program. Students in the program were given a survey that asked about their eating patterns prior to beginning the program and again after 6 months of participation. The administrator wanted to assess whether the students, on average, reported healthier eating patterns after participation in the program, using health eating scores calculated at the beginning of the program and at the end of the program.

Which statistical test should be used?

(a) Independent samples t-test
(b) Dependent samples t-test
(c) ANCOVA
(d) Two-way between subjects ANOVA

7. A school psychologist wanted to compare minority and non-minority students' social studies scores on a state-wide standardized test using a simple random sample of 200 middle school students from a county in Tennessee. The school psychologist decided to include gender in the analysis in case any difference between minority and non-minority students' social studies scores could be dependent on the gender of the students.

Which statistical test should be used?

(a) One-sample t-test
(b) Chi-square goodness of fit test
(c) One-way repeated measures ANOVA
(d) Factorial between subjects ANOVA

8. A researcher conducted a study to test the effectiveness of a new classroom instructional intervention using 20 volunteer college students. The researcher randomly assigned each participant to one of the two treatment groups. The experimental group (N=10) received the new instruction and the control group (N=10) received traditional instruction. All participants were given a pre-test and post-test. The researcher also wanted to analyze the effectiveness of the intervention while controlling for the differences in cognition, as measured by the pre-test scores of students in the two groups.

Which statistical test should be used?

(a) ANCOVA using pre-test scores as a covariate
(b) Two-way between subjects ANOVA with post-hoc tests
(c) Independent samples t-test comparing post-test scores
(d) Dependent samples t-test

9. A statistics professor wanted to compare the students' scores of introductory probability taught by two different teaching methods: 1) "Using simulation software" and 2) "Traditional lecture." The professor randomly assigns 40 students to one of two teaching methods. One group (N=20) will use simulation software, while the other group (N=20) will receive a traditional lecture. The professor will administer an end of course assessment (scored 0-100) to both groups and compare students' mean post-test scores.

Which statistical test should be used?

(a) Chi-square test for association
(b) Two-way within subjects ANOVA
(c) ANCOVA
(d) t-test for independent samples

10. A Tennessee state health official is interested in determining the association between adolescent obesity and region of the state. The health official randomly selects 250 middle school students attending a summer program at the University of Tennessee for the study. The students are classified into one of three obesity categories: 1) "Obese," 2) "Normal Weight," and 3) "Underweight" and into one of three regions of the state: 1) "East," 2) "Middle," and 3) "West." Health official intends uses counts of adolescents belonging to each of these group classifications for the analysis.

Which statistical analysis should be used?

(a) Two-way between subjects ANOVA
(b) Chi-square test of independence
(c) Multiple linear regression
(d) Pearson correlations

11. A product developer at a popular golf company is interested in determining which one of three golf drivers in the company's product line-up (Driver A, Driver B, and Driver C) produces the longest golf drives, on average, for professional golfers. The product developer finds 60 golf professionals who volunteer to participate in the study. The golfers are randomly assigned to one of the three golf driver groups and each golfer hits the driver 10 times. The outcome measure is the average of the golfers 5 best drives.

Which statistical analysis should be used?

(a) Chi-square test for independence with odds ratios
(b) Series of dependent sample t-tests with multiple comparisons
(c) One-way between subjects ANOVA with post-hoc comparisons
(d) Two-way between subjects ANOVA without post-hoc comparisons

12. A regional assessment development company introduced a new testing system for K-5 classrooms in a school district. Each student in the district was given 9 monthly progress monitoring math tests. An analyst wanted to examine whether, on average, the third grade students in this school district had improved their math knowledge over the ninth month period.

Which statistical test should be used?

(a) One-way between subjects ANOVA
(b) Independent samples t-test
(c) One-way repeated measures ANOVA
(d) Factorial between subjects ANOVA

13. A doctoral student is interested in predicting the Body Mass Index (BMI) of college females using both of the following variables: 1) "Daily Caloric Intake and 2) "Hours of Sedentary Behavior." A random sample of 100 female students from a large university was selected to participate. The participants tracked daily caloric intake and activity data for one month, resulting in a measure of daily caloric intake and sedentary behavior.

Which statistical analysis should be used?

(a) Multiple linear regression
(b) Multiple comparisons
(c) Simple linear regression
(d) Pearson product-moment correlation

14. A statistical report indicated that household income, number of members in the household, and the poverty index of the residential district determines the household government welfare benefits: 1) "Maximum benefits," 2) "Medium Benefits," and 3) "Minimum Benefits" that a household can receive. Using a simple random sample of households from a large metropolitan area and controlling for the effects of the aforementioned variables, a research analyst wanted to determine the change in welfare benefits that a household would receive if the monthly household income increased by $100.

Which statistical analysis should be used?

(a) Simple linear regression
(b) Multiple linear regression
(c) Binary logistic regression
(d) Multinomial logistic regression

15. A researcher at a school with many non-traditional students wanted to see if students from three different age groups: 1) “Less than 25 years,” 2) “25- 35 years,” and 3) “More than 35 years” would evaluate an on-line class activity differently than a hands-on class activity used in introductory biology courses. The researcher randomly selects 20 students from each of the age groups from the population of freshman biology students. Half of the students from each age group were randomly assigned to an on-line activity, while the other half received hands-on activity. The researcher also decided to explore the difference of on-line and hands-on methods evaluations could be dependent on the age of the students.

Which statistical test should be used?

(a) One-way ANOVA with Interaction
(b) Two-way between subjects ANOVA with Interaction
(c) Two-way between subjects ANOVA without Interaction
(d) Three-Way between subjects ANOVA without Interaction

16. A researcher randomly assigns 30 nurses into three different treatment groups. Group 1 receives team development training from an online website. Group 2 receives the same training through direct classroom instruction. Group 3 receives the information from a series of webinars by the same instructor. The participating nurses are asked to provide scores (0-100) for three different aspects of the instruction: 1) “Recognizing the Importance of New Strategies,” 2) “Understanding Training Material,” and 3) “Motivation to Use New Strategies in the Field after Training.” The researcher found these three ratings variables to be moderately correlated and represent a construct called “Training Team Development.” Researcher wanted to compare the effectiveness of the 3 modes of instruction on “Training Team Development.”

Which statistical test should be used?

(a) Two-Way between subjects MANOVA
(b) One-way between subjects MANOVA
(c) Two-way between subjects ANOVA
(d) Multiple one-way between subjects ANOVA

17. A personality researcher at a university collected data from students coming from three different cultures: 1) “Western,” 2) “Asian,” and 3) “Middle Eastern Arabic.” The researcher also obtained interval/ratio scale data from each student on their 1) “Social responsiveness,” 2) “Respect for women,” and 3) “Religious Attachment.” The researcher is interested in classifying future students into their cultural origin groups based on the students’ scores on the three response variables.

Which statistical analysis should be used?

(a) Multiple linear regression
(b) Multivariate multiple linear regression
(c) Factor analysis
(d) Discriminant analysis

18. A social work researcher collected data on three family health variables and four socio-economic variables, all measured on a continuous scale. The social worker is interested in how the set of three family health variables relates to the four socio-economic variables. In particular, the researcher wants to explore the nature of the latent constructs which are necessary to understand the association between the two sets of variables.

Which statistical analysis should be used?

(a) Canonical correlation
(b) Pearson product-moment correlation
(c) Factor analysis
(d) Latent class analysis

19. A researcher was interested in developing an instrument to measure international students' sense of belonging within the university community. Based on theoretical models of sense of belonging the researcher constructed and administered a preliminary 50-item survey to a random sample of international students enrolled in US universities. The researcher was interested in identifying interpretable "Sense of Belonging" constructs from the set of items.

Which statistical analysis should be used?

(a) Confirmatory factor analysis
(b) Exploratory factor analysis
(c) Latent class analysis
(d) Discriminant function analysis

20. A doctoral student in an Educational Psychology Department developed a scale to measure undergraduate students' statistical reasoning ability. The doctoral student carefully developed items such that all of the items were meant to measure one construct "Statistical Reasoning Ability," but not any other latent constructs. After collecting the data, the student wanted to make sure that the observed responses to the items justified the claim that items measured only statistical reasoning ability.

Which statistical analysis should be used?

(a) Multi-dimensional scaling
(b) Canonical correlations
(c) Confirmatory factor analysis
(d) Multivariate multiple regression

21. A math education researcher administered a validated assessment to a college Algebra class to investigate students' conceptual knowledge of factorization. The assessment provides scores for eight areas of common misconceptions. The researcher wanted to use the data to generate homogenous groups of students with the objective of giving different tutoring sessions to each group to correct their misconceptions.

Which statistical analysis should be used?

(a) Discriminant analysis
(b) MANOVA
(c) Cluster Analysis
(d) ANOVA

22. A review of literature reveals both causal and non-causal associations among 4 different financial constructs: 1) "Financial attitudes," 2)" Financial motivation," 3)" Financial Confidence," and 4) "Financial Behaviors" of college students. Further investigation into literature revealed that "Financial Education" has direct impact on "Financial behaviors." Additionally, higher levels of "Financial Motivation" and "Financial Attitudes" have caused students to engage in financial education. A group of doctoral students are interested in examining the associations among all these constructs and how these established associations might affect financial behaviors when considered simultaneously.

Which statistical analysis should be used?

(a) Structural equation modeling
(b) Multilevel modeling or Hierarchical linear models
(c) Multiple linear regression
(d) Factor analysis

23. A researcher wanted to model the impact of a new 5-year state-wide education program on 5th grade student's standardized test score improvement. The researcher received a large dataset, containing data from a random sample of 20 elementary schools, chosen from 15 school districts in the state of Tennessee. The dataset included standardized test scores of all 5th graded in selected schools, both prior to the program and for each of the five program years, demographic data (about districts and schools), and variables related to the implementation of the program at the school level. Reading several formative evaluation reports, the evaluator found that the score change resulted from 5-yr education program is homogenous for students in a particular school.

Which statistical analysis should be used?

(a) Multi-dimensional scaling
(b) Multivariate multiple regression
(c) Structural Equation modeling: Growth curve analysis
(d) Multilevel modeling or Hierarchical linear models

24. A researcher is interested in determining if there are memory task performance differences for college students coming from three different ethnic backgrounds. Memory task performance consists of three components: 1) “Recognition Score,” 2) “Free recall score,” and 3) “Cued Recall score.” All three components are moderately correlated and represent the construct of “Memory Task Performance”. The researcher wants to compare memory task performance of the different ethnic groups while controlling for the students’ academic ability as measured by their ACT score.

Which statistical test should be used?

(a) Two-way between subjects MANOVA
(b) MANCOVA
(c) Two-Way between subjects ANOVA
(d) ANCOVA

25. The director of a national vaccination program is interested in the associations among the following variables of pregnant mothers: 1) “Education” (1- high school or above, 2 –below high school), 2) “Family Income” (1- above average, 2- equal or below average), and 3) “Vaccination Status” (1- completed, 2- uncompleted). The researcher collects data from a simple random sample of 1,000 expectant mothers and created an aggregated table with counts for each combination of the above variables to be used in the analysis. He wanted to examine the association between “Vaccination Status” and “Education”, controlling for the effect from “Family Income”.

Which statistical analysis should be used?

(a) Chi-square test of Independence
(b) ANCOVA
(c) Log-linear model
(d) Multi-level modeling or Hierarchical linear models

Demographic Questionnaire

Demographic and Background Questions

Directions: Please use this page to answer a few background questions by selecting the single answer that best describes you or writing down the appropriate statement.

1. What level of graduate program are you currently enrolled in?

☐ Masters ☐ Doctoral ☐ Prefer not to answer

2. What is your major area of study? ……

3. What is your study concentration (If applicable)? ……

4. What year of the program are you currently completing?

☐ First Year ☐ Second Year ☐ Other, please specify ……

☐ Third Year ☐ Fourth Year ☐ Prefer not to answer

5. What is your gender? ☐ Male ☐ Female ☐ Prefer not to answer

6. Are you currently taking Statistics class/es? ☐ No ☐ Yes

Have you ever completed any Statistics class? No Yes (if Yes, circle how many classes)

6. Graduate level ☐ ☐ # Classes: 1 2 3 or more

7. Undergraduate level ☐ ☐ # Classes: 1 2 3 or more

9. Other ☐ ☐ # Classes: 1 2 3 or more

10. How long has it been since you completed a statistics class?

☐ 1 semester ago ☐ Within 1 year

☐ Within 1 to 2 years ☐ More than 2 years

11. Have you ever taken any research methodology (covers research design and data analysis) class/es? ☐ No ☐ Yes

12. How often are you involved with conducting statistical analysis for research *other than classes*?

☐ Never ☐ Rarely ☐ Occasionally ☐ Often

13. Please rate your level of confidence in your ability to conduct statistics related tasks?

☐ Not Confident at all ☐ Slightly confident ☐ Somewhat confident ☐ Quite confident

---------- **Thank you** ----------

Answer Booklet

Question	Answer	Question	Answer	Question	Answer
1	B	11	C	21	C
2	C	12	C	22	A
3	D	13	A	23	D
4	B	14	D	24	B
5	A	15	B	25	C
6	B	16	B		
7	D	17	D		
8	A	18	A		
9	D	19	B		
10	B	20	C		

Appendix D
Letter to Instructors – Focus Group

Dear Instructor,

I am writing to request your assistance in helping me to develop a new measure of applied statistics knowledge. I am developing a new instrument, the Statistics Assessment of Graduate Students (SAGS) for assessing graduate students' statistics knowledge required to successfully and efficiently complete applied research in education and other social and behavioral sciences. This instrument is being developed for my dissertation research at the University of Tennessee, where I am a doctoral candidate in the Department of Educational Psychology and Counseling with a concentration in Evaluation, Statistics, and Measurement. For developing the scale, I am working with my faculty advisor, Dr. Gary J. Skolits.

During the first phase of the study, initial items for the SAGS instrument have been developed. Next, it is expected to conduct a focus group with graduate students to further review and modify these initial SAGS items/instrument. I am now in the process of recruiting graduate students who have completed a higher level statistics course(s) to be participate in this focus group.

I kindly request you to support this study by forwarding the announcement of this focus group (see attached letter to Students) to graduate students who are currently enrolled in as well as students who have recently completed your (***Course title here)*** class.

Thank you for your time and consideration.

Sincerely,
Dammika Lakmal Walpitage
Doctoral Candidate: Evaluation, Statistics and Measurement
Department of Educational Psychology and Counseling
University of Tennessee
503 Jane and David Bailey Education Complex
1122 Volunteer Boulevard
Knoxville, TN 37996-3452
Phone: (865) 599 - 9813
Email:dwalpita@vols.utk.edu

Supervisor:
Gray Skolits, Ed.D.
Associate Professor: Evaluation, Statistics and Measurement
Director: Institute for Assessment and Evaluation
503 Jane and David Bailey Education Complex
1122 Volunteer Boulevard
Knoxville, TN 37996-3452
Email: gskolits@vutk.edu
Phone: (865) 974-2777
Fax: (865) 974-0135

Appendix E
Invitation Letter to Students – Focus Group

Dear Fellow Graduate Student,

I am a doctoral candidate in the Department of Educational Psychology and Counseling in Educational Psychology and Research with a concentration in Evaluation, Statistics, and Measurement. I am writing to request your help to develop a new instrument as a measure of statistical research methodology knowledge. The new instrument, the Statistics Assessment of Graduate Students (SAGS), assesses graduate students' statistical knowledge required to successfully and efficiently complete applied research in education and other social and behavioral sciences. This instrument is being developed for my dissertation research at the University of Tennessee, Knoxville. For developing the instrument, I am working with my faculty advisor, Dr. Gary J. Skolits.

Specifically, I would like to invite you to participate in a focus group (5-10 graduate students) with the intention of improving the initially developed SAGS instrument which includes a cognitive test and demographic questionnaire. During the focus group, students will be asked to review SAGS cognitive items. After each item question stem, the correct answer as well as any distractors will be discussed with the group. Moreover, the demographic items will be reviewed and the group will be asked to identify additional demographic items needed or which items to be deleted or modified. Further, quality of the instructions to participants will be discussed as a group.

Your participation in the focus group is completely voluntary and your participation will remain confidential. A paper copy of the informed consent will be given prior to the focus group meeting and you may decline to participate/withdrawal in the study at any time. Please be assured that any information you provide during the focus group will be kept strictly confidential. Your focus group data will not be made available to any person other than principal investigator, Walpitage and his faculty advisor, Dr. Gary Skolits during the study, following the study, and when reporting research results. The focus group will not, at any time, be audio or video recorded. Focus group should last between 60 and 75 minutes. You will not have any direct benefit from participating in this focus group, however your participation will be valuable in developing a validated instrument that can be considered an important contribution to statistics education literature.

If you are willing to participate in this study, please confirm your participation with me by sending an e-mail (dwalpita@vols.utk.edu) message. I will then make necessary arrangements to finalize the focus group date and time based on the schedules of the students who agreed to participate in this focus group. If you have any questions or need more information don't hesitate to contact myself or Dr. Gray Skolits (gskolits@utk.edu).

Thank you for your time and consideration.

Sincerely,

Dammika Lakmal Walpitage
Doctoral Candidate: Evaluation, Statistics and Measurement
Department of Educational Psychology and Counseling
University of Tennessee
503 Jane and David Bailey Education Complex
1122 Volunteer Boulevard
Knoxville, TN 37996-3452
Phone: (865) 599 - 9813
Email:dwalpita@vols.utk.edu

Supervisor:
Gray Skolits, Ed.D.
Associate Professor: Evaluation, Statistics and Measurement
Director: Institute for Assessment and Evaluation
503 Jane and David Bailey Education Complex
1122 Volunteer Boulevard
Knoxville, TN 37996-3452
Email: gskolits@utk.edu
Phone: (865) 974-2777
Fax: (865) 974-0135

Appendix F

Informed Consent Statement

Development and Validation of Statistics Assessment of Graduate Students (SAGS) Instrument

INTRODUCTION

You are invited to participate in a research project to develop a new instrument as a measure of statistical research methodology knowledge. We are developing this new instrument, the Statistics Assessment of Graduate Students (SAGS) for assessing graduate students' statistics knowledge required to successfully and efficiently complete applied research in education and other social and behavioral sciences. You must be at least 18 years of age to participate in this study.

INFORMATION ABOUT PARTICIPANTS' INVOLVEMENT IN THE STUDY

Your participation in this study asks you to participate in a 60-75-minute focus group with other graduate students in University of Tennessee, Knoxville. At the conclusion of the focus group, your involvement is completed. You will be asked to review the items in the SAGS instrument individually and SAGS instrument as a whole, and provide feedback to improve the quality of the instrument. The focus group will not be audio or video recorded. You may make notes and edits in the provided hard copy of the instrument. You may provide your notes at the end of the focus group to the principal investigator. However, you are not required to take notes nor submit notes to the researcher at the end of the focus group.

RISKS

There are no foreseeable risks other than those encountered in everyday life. We will make every effort to protect the confidentiality of participants' data obtained during this study.

BENEFITS

A benefit to you may be increased applied statistics knowledge with regards to selecting appropriate statistics tests/procedures to address the given research problems.

This research study will contribute to statistics education literature as a new measure to assess graduate students' knowledge in statistics for conducting applied research.

CONFIDENTIALITY

When conducting a focus group, researchers cannot guarantee the confidentiality of subjects, as researchers cannot control what subjects might share outside of the research environment. We will ask you to please not share our conversation outside our group. Only the researchers will have access to your information and focus group participants' feedback. All data related focus group will be stored securely and will be made available only to persons conducting the study

_____ (initials here)

unless participants specifically give permission in writing to do otherwise. The focus group will not be audio or video recorded. All notes made by researchers and focus group participants will be destroyed after the notes has been used to modify the instrument and sanitized for any identifying information. No reference will be made in oral or written reports which could link participants to the study.

CONTACT INFORMATION

If you have questions at any time about the study or the procedures, (or you experience adverse effects as a result of participating in this study,) you may contact the researcher at the University of Tennessee, Dammika Lakmal Walpitage, at dwalpita@vols.utk.edu or his advisor, Dr. Gary J Skolits at gskolits@utk.edu.

If you have questions about your rights as a participant, you may contact the University of Tennessee IRB Compliance Officer at utkirb@utk.edu or (865) 974-7697.

PARTICIPATION

Your participation in this study is completely voluntary; you may decline to participate without penalty. If you decide to participate, you may withdraw from the study at any time without penalty and without loss of benefits to which you are otherwise entitled. If you withdraw from the study before data collection is completed your data will be returned to you or destroyed.

CONSENT

I have read the above information. I have received a copy of this form. I agree to participate in this study.

Participant's Name (printed) _____

Participant's Signature _____ Date _____

Appendix G
Graduate Student Focus Group Protocol
Development and Validation of Statistics Assessment of Graduate Students (SAGS) Instrument

Instructions and Introduction

My name is Dammika Lakmal Walpitage and I'm here today to discuss the Statistics Assessment of Graduate Students (SAGS) instrument. I drafted the instrument and I am in the process of improving the quality of these items and instrument as a whole. Today, I'll be talking with you approximately one hour about SAGS instrument and will ask for your comments suggestions to make improvement.

I appreciate your honest feedback. I want to hear from each and every one of you; your opinions and comments are important to me. Your feedback will be used to modify the initially developed SAGS instrument. I thank you in advance for your feedback.

Even though you have come here today, you still have the option of declining participation. This is a purely voluntary activity. If you do not wish to participate, please do not. You can opt out at any point during the group even if you start the process. Please understand that the information you provide us today is kept confidential, therefore I will not be able to connect your responses with your true identity.

I will be not recording this focus group. I will be taking my notes. You can make notes and edits on the provided hard copy of the instrument. Please give me your notes at the end of the focus group. But you are not required to take notes or give me your notes to me at the end of the focus group. All notes made by me and the focus group participants will be destroyed after the notes have been used to modify the instrument and sanitized for any identifying information. Do you have any questions?

Before we get started, does anyone have any questions? Okay, let's get started.

Questions

Please look at the copy of the initial SAGS instrument. First read the introduction and answer the first item. (give 4- 5 minutes to complete this task)

We will talk about the question in detail.

Is there any unclear or confusing content for question stem? Can you explain this a little bit?

Could you identify any grammatical mistakes that you think should be addressed?

Could you identify any part of the item stem that you think could be improved?

Anticipated answer for this item is C. What do you think about this anticipated answer? How effective were the distractors?

Now, we will move to next 5 questions. ……..

Please look at the demographic questionnaire. We will now review the items in this second part of the instrument

Could you identify any item/s that you think need to be modified? What changes you would think most appropriate?

What comments or suggestions do you have for the ***instructions to complete the instrument?***

What type of new items (statistical test/procedures) do you think should be included in this instrument? Which items do you think should be deleted?

Is there anything else you would like me to know about the SAGS items or about the SAGS instrument?

Conclusion

That is the end of our time together. Thank you for taking the time out of your day to participate in this focus group. If you don't have anything else to share, that concludes the focus group. I appreciate your time!

Appendix H
Invitation Letter to Expert Reviewers

Dear Dr. *Type Name Here*,

I am a doctoral candidate in the Department of Educational Psychology and Counseling in Educational Psychology and Research with a concentration in Evaluation, Statistics, and Measurement. I am writing to request your help in developing a new instrument as a measure of statistical research methodology knowledge. The new instrument, the Statistics Assessment of Graduate Students (SAGS), assesses graduate students' statistical knowledge required to successfully and efficiently complete applied research in education and other social and behavioral sciences. This instrument is being developed for my dissertation research at the University of Tennessee, Knoxville. For developing the instrument, I am working with my faculty advisor, Dr. Gary J. Skolits.

Specifically, I would like to invite you to participate in the study as an expert reviewer to improve the quality and establish validity evidence of the initially developed SAGS instrument which includes a cognitive test and demographic questionnaire. Expert review should take approximately one hour. As an independent expert reviewer you will be asked to review the SAGS items individually and provide suggestions regarding the face and content validity. Further, you will be asked to provide feedback on the item stem and the quality of the item distractors. You will also be requested to provide information on any potential new items that should be added, existing items that should be deleted or revised, issues regarding spelling and/or grammar, and appropriateness of the item ordering. Moreover, you will be asked to evaluate the clarity of instructions and provide an estimate on the approximate time that students from the target population would take to complete the SAGS instrument.

Your participation in the expert review panel is completely voluntary and your participation will remain confidential. You may decline to participate in the study at any time. Please be assured that any information you provide during the exert review will be kept strictly confidential. Your expert review comments will not be made available to any person other than Principal Investigator, Walpitage and his faculty advisor during the study, following the study, and when reporting research results. You will not have any direct benefit from participating in this expert review however, your participation will be valuable in developing a validated instrument that can be considered an important contribution to statistics education literature.

If you are willing to participate in this study, please confirm your participation with me by sending an e-mail (dwalpita@vols.utk.edu) message. Then I will make necessary arrangement to provide you with a paper copy of the informed consent, review protocols, and a copy of the SAGS instrument prior to the expert review. If you have any questions or need more information don't hesitate to contact me or Dr. Gray Skolits (gskolits@utk.edu).

Thank you for your time and consideration.

.

Sincerely,

Dammika Lakmal Walpitage
Doctoral Candidate: Evaluation, Statistics and Measurement
Department of Educational Psychology and Counseling
University of Tennessee
503 Jane and David Bailey Education Complex
1122 Volunteer Boulevard
Knoxville, TN 37996-3452
Phone: (865) 599 – 9813
Email:dwalpita@vols.utk.edu

Supervisor:
Gray Skolits, Ed.D.
Associate Professor: Evaluation, Statistics and Measurement
Director: Institute for Assessment and Evaluation
503 Jane and David Bailey Education Complex
1122 Volunteer Boulevard
Knoxville, TN 37996-3452
Email: gskolits@utk.edu
Phone: (865) 974-2777
Fax: (865) 974-0135

Appendix I

Informed Consent Statement

Development and Validation of Statistics Assessment of Graduate Students (SAGS) Instrument

INTRODUCTION

You are invited to participate in this research project to develop a new instrument as a measure of statistical research methodology knowledge. We are developing this new instrument, the Statistics Assessment of Graduate Students (SAGS) for assessing graduate students' statistics knowledge required to successfully and efficiently complete applied research in education and other social and behavioral sciences.

INFORMATION ABOUT PARTICIPANTS' INVOLVEMENT IN THE STUDY

Your participation in this study asks you to participate in the study as an expert reviewer to improve the quality and establish validity evidence of the initially developed SAGS instrument. You will be asked to review the items in the SAGS instrument individually and SAGS instrument as a whole using the rubric provided by the researchers. At the end of review, you could return the expert review documents to the researcher. Expert review will take approximately one hour.

RISKS

There are no foreseeable risks other than those encountered in everyday life. We will make every effort to protect the confidentiality of participants' data obtained during this study.

BENEFITS

This research will contribute to statistics education literature as a new measure to assess graduate students' knowledge in statistics for conducting applied research.

CONFIDENTIALITY

The feedback that you will be providing in the expert review will be kept confidential. Only the researcher will have access to your information and expert review comments. All data related to expert review will be stored securely and will be made available only to persons conducting the study, unless participants specifically give permission in writing to do otherwise. All comments or notes made by expert reviewers will be destroyed after their feedback has been used to modify the instrument and sanitized for any identifying information. No reference will be made in oral or written reports which could link expert reviewers to the study.

_____ (Initial here)

CONTACT INFORMATION

If you have questions at any time about the study or the procedures, (or you experience adverse effects as a result of participating in this study,) you may contact the researcher at the University of Tennessee, Dammika Lakmal Walpitage, at dwalpita@vols.utk.edu or his advisor, Dr. Gary J Skolits at gskolits@utk.edu.

If you have questions about your rights as a participant, you may contact the University of Tennessee IRB Compliance Officer at utkirb@utk.edu or (865) 974-7697.

PARTICIPATION

Your participation in this study is completely voluntary; you may decline to participate without penalty. If you decide to participate, you may withdraw from the study at any time without penalty and without loss of benefits to which you are otherwise entitled. If you withdraw from the study before data collection is completed your data will be returned to you or destroyed.

CONSENT

I have read the above information. I have received a copy of this form. I agree to participate in this study.

Participant's Name (printed) _____

Participant's Signature _____ Date _____

Appendix J
Expert Review Rubric for Face and Content Validity Evidence
SAGS INSTRUMENT ITEM REVIEW RUBRIC

Thank you again for being willing to participate in the expert review process of my dissertation research developing and validating the Statistics Assessment of Graduate Students (SAGS) instrument. SAGS will be used to assessing graduate students' statistics knowledge required to successfully and efficiently complete applied research in education and other social and behavioral sciences (SBE). This document contains a rubric to guide you through the review process for each item as well as well as score sheet to provide your feedback.

SAGS INSTRUMENT ITEM REVIEW RUBRICINSTRUCTIONS: Next page (Page 2) asks broad questions about the overall format of the assessment ***including the instructions, length, and item order*** (i.e. "flow). After these first questions, please use the rubric on Page 3 to review the components of each item (the stem, response options, and content), and rate them on the worksheet I have created on Page 5. Finally, Page 6 contains an area for you to provide any additional comments or suggestions you may have for improving the instrument before it is tested.

Item review can be greatly simplified by reading from the top row where the numbers show the general rating for each component with 1 = "Heavy revisions necessary" and 4 = "Keep as is, no revisions necessary." The descriptions in each cell of the rubric are simply to assist you in the review process should you be confused or need more clarification. It may also be helpful to print the rubric out and have it at your side while reading through the items rather than flipping back-and-forth.

Thank you again for your willingness to lend your expertise, and happy reviewing!

OVERALL REVIEW

What comments or suggestions do you have for the ***instructions***? Will the participant know what is expected of them when they are given the instrument?
Is the ***length*** of the instrument appropriate? Are there enough items to sufficiently address applied statistics knowledge for education and other human sciences graduate students?
What comments or suggestions do you have for the way in which the items are ***ordered***?

INDIVIDUAL ITEM REVIEW MATRIX

Component	1 (Heavy Revisions Necessary)	2 (Some Revisions Necessary)	3 (Minimal Revisions Necessary)	4 (Keep as is, no revisions necessary)
Item Stem (the narrative part of the question)	Item stem is unclear; stem is missing important information that is necessary for answering the question. There is irrelevant or "trick" information that would prevent a student who knows the concept correctly answering the question. There are clues or hints that would help a student with no knowledge of the concept to answer the question correctly.	Stem does not clearly provide the information necessary to answer the question, and contains at least one of the following: •some irrelevant or "trick" information •clue or hint to the correct answer •grammatical or content-related errors	The item stem needs minor clarification. However, provides the information necessary to answer the question, but contains one of the following: •some irrelevant or "trick" information •clue or hint to the correct answer •grammatical or content-related errors	Stem clearly provides the information necessary to answer the question. The length for the stem is appropriate. There is no irrelevant or "trick" information, clues or hints to the correct answer, or grammatical errors.

Response Options (A – D)	The response options are not clearly written, the keyed answer is factually inaccurate, or there are significant errors in any of the following: • No clear "Order" of correctness • Grammatical connections to stem and/or vignette • Uneven length • Response option links to other items in the instrument.	Some response options are clearly written, and the keyed answer is the correct option, but there are errors in least two of the following: • No clear "Order" of correctness • Grammatical connections to stem and/or vignette • Uneven length • Response option links to other items in the instrument.	All response options are clearly written, but contain minor errors in any of the following: • No clear "Order" of correctness • Grammatical connections to stem and/or vignette • Uneven length • Response option links to other items in the instrument.	All response options are clearly written, and can be arranged in order of "correctness" with the keyed answer as the single best option. There are no grammatical links from the response set to either the vignette or stem. All options are of similar length, and do not link themselves to other items in the instrument.
Content (the statistical concepts /procedures terminology used in the item)	The content chosen for the question is not relevant to applied statistics. The question is not appropriate for the SBE graduate student population. (OR) There is significant lack of contextual relevance to applied research, or plausibility that could affect responses.	The content chosen for the question is relevant to applied statistics, but may not be appropriate SBE graduate student population. (OR) There is some lack of contextual relevance to applied research, or plausibility that could affect responses.	The content chosen for the question is relevant to applied statistics, but may not be appropriate SBE graduate student population. (OR) There is minor lack of contextual relevance to applied research, or plausibility that could affect responses.	The content chosen for the question is relevant to applied statistics, but may not be appropriate SBE graduate student population. (OR) There is appropriate contextual relevance to applied research, or plausibility that could affect responses.

ITEM REVIEW WORKSHEET

---

1. Please use the rubric on the previous page to rate each item in the table below (each is rated between 1 = "Heavy revisions necessary" and 4 = "Keep as is, no revisions necessary").
2. After you have rated each component, please rate the overall quality of the item from 1 = "Very poor" to 5 = "Excellent."

Item #	Item Stem	Response Options	Content	Overall
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				
21				
22				
23				
24				
25				

ADDITIONAL COMMENTS

Please use this page to write any additional comments have about specific items or components of the SAGS instrument (and example is given). Thank you!

Example Q6: "Rearrange the order of response options to make it more logical to the reader".

Appendix K
Letter to Instructors – Main Assessment

Dear Instructor,

I am writing to request your assistance in helping me to gather data for developing a new measure of statistical research methodology knowledge. I am developing a new instrument, the Statistics Assessment of Graduate Students (SAGS) for assessing graduate students' statistics knowledge required to successfully and efficiently complete applied research in education and other social and behavioral sciences. This instrument is being developed for my dissertation research at the University of Tennessee, where I am a doctoral candidate in the Department of Educational Psychology and Counseling with a concentration in Evaluation, Statistics, and Measurement. For developing the instrument, I am working with my dissertation chair, Dr. Gary J. Skolits.

For the purpose of this instrument, applied statistics ability is defined as individuals' capability to identify and select best statistical test/procedure among a list of similar procedures to analyze given research situation/question. The data gathered will be used to determine the underlying structure of applied statistics ability as well as to evaluate the validity and reliability of the *SAGS* instrument.

I kindly request you to forward the announcement of this study (see attached flyer) with the link to the online instrument to students enrolled in your graduate level courses. Students will be able to complete this instrument (takes approximately take 40 minutes to complete) online at their convenience.

Thank you for your time and consideration.

Sincerely,

Dammika Lakmal Walpitage
Doctoral Candidate: Evaluation, Statistics and Measurement
Department of Educational Psychology and Counseling
University of Tennessee
503 Jane and David Bailey Education Complex
1122 Volunteer Boulevard
Knoxville, TN 37996-3452
Phone: (865) 599 - 9813
Email:dwalpita@vols.utk.edu

Supervisor:
Gray Skolits, Ed.D.
Associate Professor: Evaluation, Statistics and Measurement
Director: Institute for Assessment and Evaluation
503 Jane and David Bailey Education Complex
1122 Volunteer Boulevard
Knoxville, TN 37996-3452
Email: gskolits@utk.edu
Phone: (865) 974-2777
Fax: (865) 974-0135

Appendix L
Study Flyer

Statistics Assessment of Graduate Students

Are you graduate student (doctoral or master's) who has exposure to applied Statistics?

This is a great opportunity to self-evaluate your knowledge in selecting the appropriate statistical test/procedure.

Do you want to help fellow doctoral student complete his dissertation?

Please take 40 minutes to complete this on-line instrument.

Click on the URL or scan the QR code below to read more:

https://tiny.utk.edu/Applied_Statistics_Tests

Students who complete the instrument have a chance to receive complete answer booklet with explanations.

If you have any questions, contact:

D. Lakmal Walpitage
Ph.D. Candidate - University of Tennessee
503 Jane and David Bailey Education Complex
1122 Volunteer Boulevard
Knoxville, TN 37996-3452
(865) 599 – 9813
dwalpita@vols.utk.edu.

Faculty Advisor: Dr. Gary J Skolits
gskolits@utk.edu

Appendix M
Invitation Letter to List-serve and Social Media Page Administrators

Dear *Type Name Here*,

I am writing to request your assistance in helping me to gather data for developing a new measure of Statistical Research Methodology knowledge. I am developing a new instrument, the Statistics Assessment of Graduate Students (SAGS) for assessing graduate students' statistics knowledge required to successfully and efficiently complete applied research in education and other social and behavioral sciences. This instrument is being developed for my dissertation research at the University of Tennessee, where I am a doctoral candidate in the Department of Educational Psychology and Counseling with a concentration in Evaluation, Statistics, and Measurement. For developing the instrument, I am working with my dissertation chair, Dr. Gary J. Skolits.

For the purpose of this instrument, applied statistics ability is defined as individuals' capability to identify and select best statistical test/procedure among a list of similar procedures to analyze given research situation/question. The data gathered will be used to determine the underlying structure of applied statistics ability as well as to evaluate the validity and reliability of the *SAGS* instrument.

I kindly request you to post an announcement of this study in your organization's list-serve/social media page (see attached flyer/Social media post) with the link to the online instrument to members of your list-serve/ social media friends. Members/friends will be able to complete this instrument (takes approximately take 40 minutes to complete) online at their convenience.

Thank you for your time and consideration.

Sincerely,

Dammika Lakmal Walpitage
Doctoral Candidate: Evaluation, Statistics and Measurement
Department of Educational Psychology and Counseling
University of Tennessee
503 Jane and David Bailey Education Complex
1122 Volunteer Boulevard
Knoxville, TN 37996-3452
Phone: (865) 599 - 9813
Email:dwalpita@vols.utk.edu

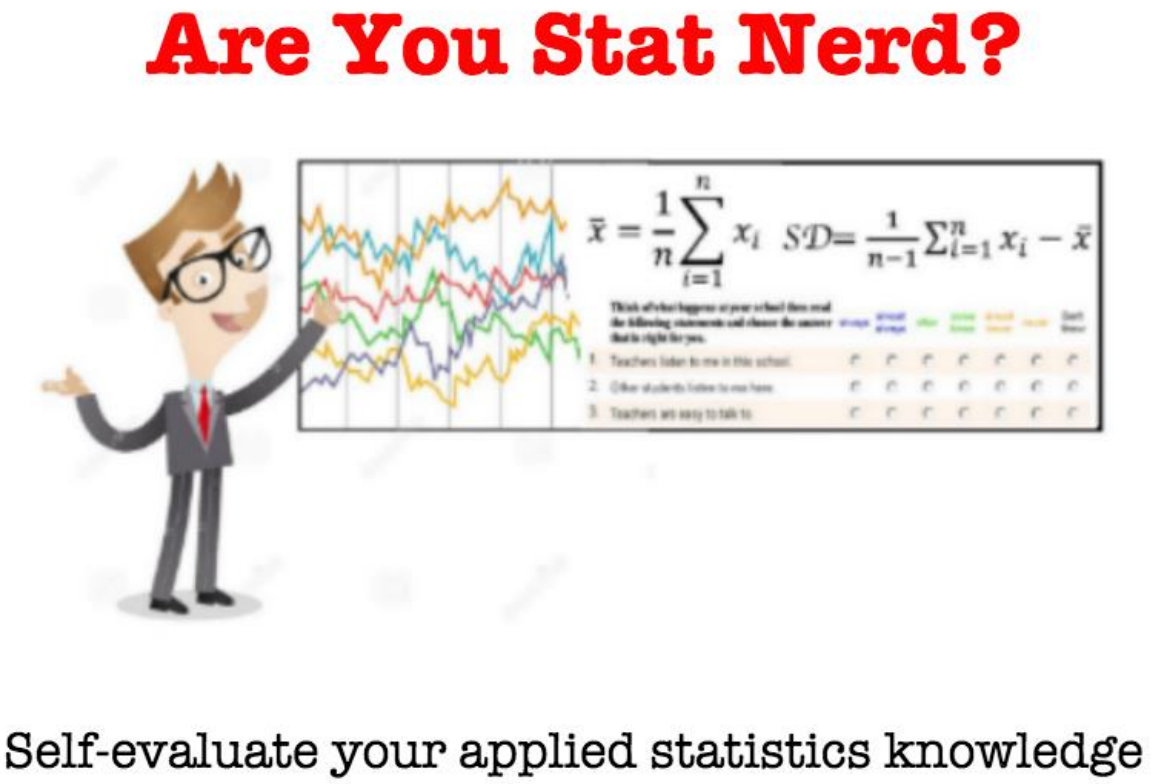
Supervisor:
Gray Skolits, Ed.D.
Associate Professor: Evaluation, Statistics and Measurement
Director: Institute for Assessment and Evaluation
503 Jane and David Bailey Education Complex
1122 Volunteer Boulevard
Knoxville, TN 37996-3452
Email: gskolits@utk.edu
Phone: (865) 974-2777
Fax: (865) 974-0135

Appendix N
Social Media Post

Development and Validation of Statistics Assessment of Graduate Students

Are you graduate student (doctoral or master's) who has exposure to Statistics? Do you want to self-evaluate your knowledge in selecting the appropriate statistical test/procedure? Do you want to help fellow doctoral student complete his dissertation?

See more at: https://tiny.utk.edu/Applied_Stat_Test



Appendix O

Informed Consent Statement

Development and Validation of the Statistics Assessment of Graduate Students (SAGS) Instrument

INTRODUCTION

You are invited to participate in this research project to develop a new instrument as a measure of statistical research methodology knowledge. We are developing this new instrument, the Statistics Assessment of Graduate Students (SAGS) for assessing graduate students' statistics knowledge required to successfully and efficiently complete applied research in education and other social and behavioral sciences. As a student, you can provide us with valuable information regarding your knowledge in selecting appropriate statistical procedures for a given research situation which will be used to test the reliability and validity of the instrument. You must be at least 18 years of age to participate in this study.

INFORMATION ABOUT PARTICIPANTS' INVOLVEMENT IN THE STUDY

Your involvement in the study is to complete an online instrument that consists of a short test and demographic questionnaire. Confidentiality will be protected to the extent that is allowed by law. We will make every effort to protect the confidentiality of participants' data obtained during this study.

RISKS

There are no foreseeable risks other than those encountered in everyday life. We will make every effort to protect the confidentiality of participants' data obtained during this study.

BENEFITS

This research will contribute to statistics education literature as new measure to assess graduate students' knowledge in statistics for conducting applied research

CONFIDENTIALITY

The information that you will be entering in the on-line instrument will be kept confidential. Only the researchers will have access to your information and responses to online instrument. All data will be stored securely and will be made available only to persons conducting the study unless participants specifically give permission in writing to do otherwise. No reference will be made in oral or written reports which could link participants to the study.

CONTACT INFORMATION

If you have questions at any time about the study or the procedures, (or you experience adverse effects as a result of participating in this study,) you may contact the researcher at the University

of Tennessee, Dammika Lakmal Walpitage, at dwalpita@vols.utk.edu or his advisor, Dr. Gary J Skolits at gskolits@utk.edu.

If you have questions about your rights as a participant, you may contact the University of Tennessee IRB Compliance Officer at utkirb@utk.edu or (865) 974-7697.

PARTICIPATION

Your participation in this study is completely voluntary; you may decline to participate without penalty. If you decide to participate, you may withdraw from the study at any time without penalty and without loss of benefits to which you are otherwise entitled. If you withdraw from the study before data collection is completed your data will be returned to you or destroyed.

CONSENT

I have read and understood the above information. Please print or save a copy of this information for your records. If you agree to participate in this study, please click the "Next" button to complete the instrument and indicate your consent. If you do not wish to participate in this study, then simply close the web browser window.

VITA

Dammika Lakmal Walpitage earned a BSc (Hons) degree in Statistics from the University of Colombo, Sri Lanka, where he built a solid foundation in mathematical statistics. Then he worked in industry for brief period and later joined Sri Lankan university system as junior faculty and worked in various statistics teaching positions. In 2014, Lakmal completed his MS degree in Statistics offered by the Statistics and Business Analytics Department, at the University of Tennessee, Knoxville. During his five years at the University of Tennessee, he received three competitive awards including the second place of an international statistics competition. He also completed the training programs in Certificate in Grant Writing and Proposal Development. Moreover, he has been actively contributing to disseminating knowledge through manuscripts published in peer-reviewed journals, various research, evaluation, and grant reports, and presentations at professional conferences. For his graduate research assistantships at the University of Tennessee, Knoxville, he worked at the National Institute of Mathematical and Biological Synthesis *(NIMBioS)* during his last year. At *NIMBioS* he supported the STEM evaluation projects and assessment development projects. Previously he worked as a Statistical consultant attached to Research Computing Support (RCS) group of Office of Information Technology. As a consultant he provided support on research design and statistical data analysis for faculty, staff and students. He was able to gain teaching experience by volunteering as a Graduate Teaching Assistant and a Co-instructor. Lakmal's research interests include Statistics education and statistical modeling. He is also interested in teaching statistics and mathematics. Dammika Lakmal Walpitage graduated from the University of Tennessee, Knoxville in December 2016 with a Ph.D. in Educational Psychology and Research with a concentration in Evaluation, Statistics, and Measurement.