IDENTIFYING DEVELOPMENTAL LEVELS AND LEARNING TRAJECTORIES IN STATISTICS FOR GRADE 6-12 STUDENTS

DISSERTATION

Presented to the Graduate Council of
Texas State University-San Marcos
in Partial Fulfillment
Of the Requirements

for the Degree

Doctor of PHILOSOPHY

by

Rini Oktavia, S.Si., M.Si., M.A.

San Marcos, Texas
August 2013

IDENTIFYING DEVELOPMENTAL LEVELS AND LEARNING TRAJECTORIES IN STATISTICS FOR GRADE 6-12 STUDENTS

Committee Members Approved:

_____
M. Alejandra Sorto, Chair

_____
Larry R. Price

_____
Sharon K. Strickland

_____
Qiang Zhao

Approved:

_____
J. Michael Willoughby
Dean of the Graduate College

FAIR USE AND AUTHOR'S PERMISSION STATEMENT

Fair Use

Duplication Permission

DEDICATION

To the loves of my life: my mom, my dad, my husband, my dear sons Hatta & Hamka, and all of my brothers and sisters: I would not have been able to finish this work without your love and support.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my parents Ibun Farida Boerhan and Ayah Ramli Ibrahim for their unconditional love and support for me to complete this challenging journey. I also would like to thank my husband Aidil and my dear sons Hatta and Hamka for loving and supporting me and for sacrificing their happiness for my study. My journey in the United States would not be possible without the love and support showered by my brothers and sisters, Uning, Abang, Nonie, Win, Dek Po, Rully, and Lulu. I also would like to express my appreciation to my mother in law, Mamak Nuraini, and to all family members for their continuous support. I am thankful for having such a loving and supporting big family. I am also thankful for my sisters Rahma Zuhra and Dewi Triarti Hidayati and their families who have supported my journey in the United States since day one.

This dissertation would not be possible without the support from my dissertation committee. I would like to express my sincere gratitude to my teacher and mentor, who is also my dissertation committee chair, Dr. M. Alejandra Sorto, under whose supervision I chose my dissertation topic and began the research. Her insights, ideas, critical comments, punctuality, and professionalism have shaped my research into a complete study. I also would like to thank Dr. Sharon K. Strickland for supporting my study, including getting participants for my pilot and actual surveys. I am thankful for her constructive feedback and encouragement, especially at the time when my confidence hit

its lowest level. I would like to thank Dr. Larry R. Price, for directing me to use the appropriate psychometric analyses for my study. He has helped developing my understanding of psychometric methods. I also would like to express my sincere gratitude to Dr. Qiang Zhao for his constructive comments and feedback. His critical observation on my ordinal regression analysis has strengthened the robustness of this study's results. I thank all my dissertation committee members for everything that they have done for me.

My study would not be possible without the support given by Dr. Vera Ioudina, Alana Rosenwasser, and Texas Mathworks, through its leaders Dr. Max Warshauer and Dr. Hiroko Warshauer, for their assistance in my pilot study. I am also grateful for the support given by Mr. Doug Wozniak, Mr. Luis F. Sosa, Mr. Bradley Gaskill, Mr. Raymond Holland, and all mathematics teachers who helped administering the survey instruments to their students. I thank them all for their incredible assistance for my study. I also would like to express my appreciation for all students who participated in the pilot and actual surveys. I also thank Mr. Carlos A. Mejia Colindres and Mr. Todd Ellertson for proofreading and editing my dissertation.

Printing more than a thousand copies of survey forms would not be possible without the full support from The Department of Mathematics Texas State University. I would like to express my sincere gratitude to Dr. Nathaniel Dean, my teacher and the chair of the department, for his wise advice and generous support for my academic journey. I also would like to thank all staff in the department, LaJuan, Melinda, Bekaye, and Illona. Without their full support and kindness, I would not get to this stage.

I am grateful to the Graduate College for granting a Doctoral Research Stipend Award to support my dissertation research. I am also indebted to The Department of

Mathematics Texas State University, The Syiah Kuala University, The Aceh Government, and The Directorate General of Higher Education of The Republic of Indonesia for funding my study at Texas State. I would like to express my sincere thankfulness to Dr. Joyce Fischer, Dr. Terence McCabe, Dr. M. Alejandra Sorto, Dr. Alexander White, Dr. Hizir Sofyan, Dr. T. Iqbalsyah, Dr. Marlina, and all friends and colleagues at Syiah Kuala University that have helped me in getting the financial support for my study.

I would not be ready to conduct a dissertation study without given a strong theoretical foundation. For this reason, I should thank all my teachers, especially my teachers at Texas State including Dr. Thomas L. Thickstun, Dr. Selina V. Mireles, Dr. Gilbert Cuevas, Dr. Samuel Obara, Dr. Zhonghong Jiang, Dr. Xingde Jia, and Dr. Clarena Larrotta. I am also grateful for all supports that I got in learning, teaching, and everyday life from Dr. Stanley Wayment, Mrs. Sonya Evans, Mrs. Diann McCabe, Dr. Donald Hazlewood, Dr. Maria Acosta, Dr. Ricardo M. Torrejon, Dr. Susan E. Morey, Dr. Gregory Passty, Dr. Jian Shen, and my colleagues doctoral students Yuliya Melnikova, Sarah Hanusch, Jake Hammons, Dr. Aaron Wilson, and all other colleagues graduate students that cannot be mentioned here because of the limited space. I thank them for everything that they have done for me. Finally, I would like to thank everyone who have supported my study that could not be mentioned in this acknowledgement. The truth is, this dissertation is a mutual aid project that involve so many professional, caring, and loving people, and I am thankful for that.

This manuscript was submitted on June 25, 2013.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

ABSTRACT

IDENTIFYING DEVELOPMENTAL LEVELS AND LEARNING TRAJECTORIES IN STATISTICS FOR GRADE 6-12 STUDENTS

by

Rini Oktavia, S.Si., M.Si., M.A.

Texas State University-San Marcos

August 2013

SUPERVISING PROFESSOR: M. ALEJANDRA SORTO

Since the early 80s, reform movements have recommended increasing the content and rigor of statistics in school mathematics curriculum. Two important curriculum documents, the Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A Pre-K-12 Curriculum Framework and the Common Core State Standards in Mathematics (CCSS-M) provide detailed descriptions of what students should know and should be able to do in statistics (Franklin et al., 2007; National Governors Association Center for Best Practices, and Council of Chief State School Officers, 2010). These descriptions are based on the hypotheses of learning trajectories of

statistical concepts. There is a need to empirically understand the learning development and growth of statistical concepts, particularly those that are related to the investigation cycle in statistics (formulating questions, collecting data, analyzing data, and interpreting results). Understanding the learning trajectories of statistics is important for instruction and assessment of statistical concepts. This study aims to understand how students learn statistics by describing the developmental growth of students' understanding of statistical concepts and learning trajectories of several concepts in statistics. To reach this goal, an instrument that measures students' developmental levels in learning statistics and learning trajectories of several statistical concepts was developed. The instrument was administered to 797 high school and middle school students in Central Texas. Three methods of data analyses were conducted: (1) a basic psychometric evaluation of the instrument using classical test theory (CTT), (2) explanatory and confirmatory factor analysis and latent regression analysis by applying structural equation modeling (SEM) approaches, and (3) ordinal regression analysis.

Five structural equation models were developed following the Pre-K-12 GAISE Framework and the CCSS-M that aligned each item in the instrument to the appropriate GAISE developmental level (Level A, Level B, or Level C). The results demonstrate an acceptable fit of the model to the empirical data. The results indicate that the Pre-K-12 GAISE Framework's suggestions that students develop understanding of statistics through three hierarchical levels were supported by the data.

The descriptive analysis results also demonstrate that students who participated in this study performed well on items measuring the statistical process component of formulating questions and collecting data. Students showed lower performance on items

measuring the process component of analyzing data and understanding the nature of variability. The results also indicate that students, who have developed into Level B in the areas of formulating questions, collecting data, analyzing data, and interpreting results, might not necessarily have developed into Level B in understanding variability. For all process components excepting the nature of variability process component, the patterns tend to be similar, where items that measure lower GAISE levels have higher percentage of correct answers (lower difficulty indices) than items that measure higher GAISE levels.

The results of ordinal regression analysis reveal that the more advanced the grade levels and the latest mathematics courses taken, the higher the students' GAISE level. This indicates that students who have better preparation in mathematics tend to have higher GAISE levels. The items were split into two forms - FORM 1 and FORM 2; the ordinal regression result also reveals that FORM 1 is more sensitive in identifying Level A students than Level B and Level C students compared to FORM 2. The results, however, showed that there is no clear relation between students' GAISE Levels and their ages.

CHAPTER I

INTRODUCTION

Statistics is a methodological discipline that exists not for itself, but rather to offer to other fields of study a coherent set of ideas and tools for dealing with data (Cobb & Moore, 1997). Statistics is needed as a discipline because of the omnipresence of variability in our lives. Cobb and Moore (1997) illustrate the omnipresence of variability by pointing out that individuals vary in many aspects. Measuring an individual repeatedly will also give different measurements due to errors that are involved. This is one aspect that makes the teaching and learning of statistics substantially different from the teaching and learning of mathematics (Moore, 1992, 1997; Rossman, Chance, & Medina, 2006).

Even though educational institutions, from K-12 schools to colleges, have included statistics and probability in their curricula, there are still calls for preparing educated citizens, including teachers, to read and understand studies conducted and analyzed by others, published in journals, and reported by the media (Conference Board of the Mathematical Sciences, 2001; Franklin et al., 2007; Utts, 2003, 2010). Responding to these calls, this study aims to understand how students learn statistics by describing the developmental growth of students' understanding of statistical concepts. To accomplish this, an instrument was developed that has the potential to serve as a research and evaluation tool to better understand statistical learning trajectories.

This introduction chapter starts by briefly outlining the inclusion of statistics in school mathematics and its relation to the current reform movement. The discussion is then continued by addressing the significance of this study and the statement of the problem being investigated that led to the research questions. The chapter is concluded by describing the key terms to be used and the delimitations of the study.

Statistics in School Mathematics

In the late 70s, the National Council of Supervisors of Mathematics (NCSM) recognized the important role of data, statistics, and probability in school curriculum. This organization, which consists of leaders of mathematics at district, state, and university levels, defined "basic skills" in mathematics to include not only computation but also estimation, geometry, problem solving, computer literacy, as well as statistics and probability in its *Position Paper on Basic Mathematical Skills* (NCSM, 1977).

In the early 80s, the National Commission for Excellence in Education (NCEE) published A *Nation at Risk* (NCEE, 1983) that recommended high school graduates to be equipped to understand elementary probability and statistics (NCEE, 1983). Later in the decade, a Commission on Standards for School Mathematics established by the Board of Directors of the National Council of Teachers of Mathematics (NCTM) published the *Curriculum and Evaluation Standards for School Mathematics* (NCTM, 1989). This visionary document, for the first time, considered statistics and probability as being equally important as numeration, measurement, algebra, and geometry in K-12 mathematics curricula. A revision of these standards in 2000, *Principles and Standards for School Mathematics* (NCTM, 2000), also suggested that instructional programs should enable all students to: formulate questions that can be addressed with data;

collect, organize, and display relevant data to answer the questions; select and use appropriate statistical methods to analyze data; make inferences and predictions based on data; and understand and apply basic concepts of probability.

The NCTM standards were only guidelines for states to develop their own curricula and assessments. But in 2001, after the No Child Left Behind Act was enacted, state and local districts began working fairly independently to develop student learning expectations to hold the school systems accountable for what students learn. Although the NCTM's documents are the common references, Reys, Dingman, Sutter, & Teuscher (2005) found a wide variety of state-level mathematics curriculum standards, with little consensus on the placement or emphasis of topics within specific grade levels. Reys and Lappan (2007) found that the Grade-level Learning Expectations (GLEs) among states varied. These differences led to the development of a variety of textbooks generated by publishers that include many more topics for each grade level in order to align to as many state standards as possible. Specifically to the statistical content, Sorto (2011) conducted a systematic analysis of state standards and found that there was an uneven distribution of the content and cognitive demand across 49 states and the NCTM standards. Furthermore, when comparing content among the documents at all cognitive levels, Sorto (2011) found that the intersection of content at each cognitive level was almost empty. The only topic in common was the proper use of measures of the center of data.

In 2006, the NCTM published yet another document entitled *Curriculum Focal Points for Kindergarten through Grade 8 Mathematics: A Quest for Coherence* to address concerns with this unfocused and inconsistent implementation of the standards. This document was intended to start a dialogue on mathematical ideas that are important

at each grade level and to be an initial step to develop a more coherent, focused curriculum in the U.S. (NCTM, 2006). In the document, three curriculum focal points are identified and described for each grade level, from prekindergarten through Grade 8. The document recommends that students learn the foundation of data analysis from prekindergarten by using their knowledge in Geometry and Measurement, which are two of the three focal points for prekindergarten (with Number and Operation as the third focal point). In Grade 8 the set of curriculum focal points which combines Data Analysis, Number and Operations, and Algebra (NCTM, 2006) suggests that:

> Students use descriptive statistics, including mean, median, and range, to summarize and compare data sets, and they organize and display data to pose and answer questions. They compare the information provided by the mean and the median and investigate the different effects that changes in data values have on these measures of center. They understand that a measure of center alone does not thoroughly describe a data set because very different data sets can share the same measure of center. Students select the mean or the median as the appropriate measure of center for a given purpose. (p. 20)

Although Data Analysis is considered a focal point only for Grade 8, the document shows that all three focal points of each grade (except for Grade 2) have connections to Data Analysis. Students build their work from previous grades to develop a sound statistical knowledge for dealing with data. This is the first document that suggests a clear learning trajectory of statistical content.

Just a year after the NCTM published the Focal Points, the American Statistical Association (ASA) and NCTM published the *Guidelines for Assessment and Instruction*

in Statistics Education (GAISE) Report: A Pre-K-12 Curriculum Framework (Franklin, et al., 2007). The GAISE Framework was intended to support the NCTM Data Analysis and Probability Content Standard and to include the five process standards (Problem Solving, Reasoning and Proof, Communication, Connections, and Representation). The report also provided K-12 stakeholders (writers of state standards, writers of assessment items, curriculum directors, pre K-12 teachers, and educators at teacher preparation programs) with a developmental framework for the teaching and learning of statistics.

This aforementioned document addresses student learning objectives in statistics and data analysis and is asserted by its authors to be consistent with the NCTM Principles and Standards (NCTM, 2000). The framework introduces the statistical problem solving process components (Formulating Questions, Collecting Data, Analyzing Data, and Interpreting Results) and focuses on understanding variability. The framework then shows how this process can be presented at each of the three developmental levels (Levels A, B, and C). Although the basic structure of the process is the same at each level, the degree of sophistication in the types of statistical problems addressed over the developmental levels increases (Peck, Kader, & Franklin, 2008). Even though this progression is well described, it is based partially on research and learning theories and has not yet been empirically tested. For example, the possibility exists that students achieve mastery in one process component while not demonstrating mastery in other process components. Therefore, it would be useful to empirically understand how students develop their knowledge and skills across grade levels for each process component.

Do all process components behave similarly across developmental levels? The answer to this question has become more important with the adoption of the Common Core State Standards in Mathematics that is known as the CCSS-M (National Governors Association Center for Best Practices, and Council of Chief State School Officers, 2010), because the CCSS-M have the purpose of ensuring the uniformity of contents and learning expectations in every grade, the same goal projected by the NCTM when it published the Focal Points.

Statement of the Problem and Research Questions

Since the early 80s, reform movements have recommended increasing their content and rigor of statistics in school mathematics curriculum. The Pre-K-12 GAISE Framework (Franklin, et al., 2007) recommends that teachers teach statistics by focusing on statistical problem solving process components and adjust their instructional approaches and assessments based on students' developmental levels in learning statistics that is not necessarily age related. Even though the categorization of students into their developmental levels is well described, it is based partially on experts' opinion, research and learning theories and has not been empirically tested. As previously mentioned, the possibility exists that students who achieve mastery in one process component might not demonstrate mastery in other process components. Therefore, there is a need to empirically understand how students develop their knowledge and skills across grade levels for each process component. Additionally, there is a need to know whether students' understanding of all process components develop similarly across levels. Acknowledging how students develop their understanding of all process components across levels becomes more important with the adoption of the Common Core State

Standards in Mathematics (CCSS-M), because their purpose is to ensure the uniformity of contents and learning expectations at each grade level. These facts lead to the following problem: given the increase in content and rigor of statistics in school mathematics, there is a need to better understand the learning development and growth of statistical concepts, particularly those that are related to the investigation cycle.

This problem leads to the following research questions that will be investigated in this study:

1. How and how much do Middle School and High School students understand statistical concepts that are related to the investigation cycle (formulating questions, collecting data, analyzing data, and interpreting result)?
2. What are the learning trajectories that describe the developmental progression for different concepts and statistical investigation processes?
3. To what extent do students' understandings of statistical concepts develop similarly across developmental levels?
4. Given the structure of the progressions observed in performance of different levels, to what extent can students' developmental level be diagnosed reliably and validly?

Purpose of the Study

This study is intended to describe how students develop their understanding of statistical concepts, especially those that are related to statistical investigation processes: formulating questions, collecting data, analyzing data, and interpreting results. The primary goal of this study is to understand how students learn statistics by describing the developmental growth of statistical concepts. To accomplish this, an instrument was

developed that has the potential to serve as a research and evaluation tool to better understand statistical learning trajectories; the description of students' thinking and learning in a specific mathematical domain, and a related conjectured pathways equipped with instructional tasks to move students through a developmental progression of levels of thinking (Clements & Sarama, 2004). Students' responses to the instrument are expected to describe the developmental growth of statistical concepts as suggested by the Pre-K-12 GAISE Framework (Franklin, et al., 2007) and the CCSS-M (National Governors Association Center for Best Practices, and Council of Chief State School Officers, 2010). In addition to describing how much and how students' understanding of statistical concepts develops across levels, the results are also expected to explain the learning trajectories of statistical concepts. The learning trajectories of statistical concepts in this context are the descriptions of students' thinking and learning in a specific statistical concept, and related conjectured pathways and instructional approaches designed to move students through a developmental progression of levels of thinking (Clements & Sarama, 2004). Given the structure of the patterns observed in the performance of different levels, it is expected that the assessment tool is able to categorize students into developmental levels in a reliable and valid manner.

Significance of the Study

In order to implement the standards mandated by the CCSS-M, teachers, researchers, and curriculum developers currently have the task of understanding how children develop statistical concepts and move from one level to another. The development of curriculum standards motivates mathematics educators to create and modify lessons, instructional approaches, and assessments that align with the curriculum

standards. Evaluation studies are also conducted to investigate the effectiveness of the curriculum implementation. All studies have one common goal: to better understand how to help students achieve significant progress in learning.

Learning trajectories of statistics can explain the order and nature of the steps in the growth of students' statistical understanding and can illuminate the effective instruction that might support students in moving step by step toward the goal of becoming statistically literate high school graduates. One of the direct benefits of knowing where students are in the learning continuum and how students' understanding of statistical concepts develops across levels is to inform the preparation of the statistics curriculum and instructional approaches. Useful information provided by this study includes which concepts should be taught first vs. which concepts should be taught later, and which instructional approaches will be effective.

The expectation is that students' developmental levels in learning statistics for each statistical investigation process component suggested by the Pre-K-12 GAISE Framework can be identified. The results will describe whether there is a clear learning trajectory of the statistical concepts or whether there are different learning trajectories for different statistical contents and processes. Understanding learning trajectories of statistical concepts will enable mathematics and statistics educators to plan and implement effective instructional approaches of statistical concepts.

By providing evidence of valid measures of the developmental growth and learning trajectories of statistical concepts, the instrument developed herein has the potential to be an assessment tool for future research on teaching and learning statistics. The instrument will also be useful as an evaluation tool to assess the effectiveness of

different curricula or pedagogical approaches. Additionally, the instrument can be used as a device to explore the development of statistical literacy, reasoning, and thinking. For example, this instrument will help researchers who conduct intervention studies to measure the impact of particular interventions on students' statistical literacy, reasoning, and thinking. Furthermore, the instrument can also be used by teacher preparation programs for educating future and current teachers on how to diagnose and assess students' understanding of statistical concepts and on how to adjust the instructional approaches based on the assessment results.

Finally, this study also provides baseline information on the development of an instrument to identify students' developmental levels in learning statistics. Researchers can learn from the instrument developed here and the associated development process applied in this study to inform more sophisticated studies in the future. The framework used in this investigation and associated findings gleaned from this study will contribute in developing knowledge and research literature in the area of assessing students' learning of statistics.

Definitions of Terms

- *Learning Trajectory* in mathematics is defined as

  > the descriptions of children's thinking and learning in a specific mathematical domain, and a related conjectured route through a set of instructional tasks designed to engender those mental processes or actions hypothesized to move children through a developmental progression of levels of thinking, created with the intent of supporting children's achievement of specific goals in that mathematical domain. (Clements & Sarama, 2004, p. 83)

All conceptions of trajectories or progressions based on the common facts that students' knowledge and skill in any domain starts out small in the amount and complexity, that becomes much larger over time due to effective instruction, and that the amount of growth clearly varies with experience and instruction but also seems to reflect factors associated with maturation, as well as significant individual differences in abilities, dispositions, and interests (Daro, Mosher, & Corcoran, 2011). Trajectories consists of hypotheses not only about the order and nature of the steps in the growth of students' mathematical understanding, but also about the nature of the effective instruction that might support them in moving step by step toward the goals of school mathematics.

- *Statistical Literacy* is the ability to understand and critically evaluate statistical results that infiltrate our daily lives and to appreciate the contributions that statistical thinking can make in public and private, professional and personal decisions (Wallman, 1993). Gal (2002) identified two components of statistical literacy required by society: (1) the ability to interpret and critically evaluate statistical information, data-related arguments, or stochastic phenomena, which may be encountered in diverse contexts, and when relevant; and (2) the ability to discuss or communicate reactions to such statistical information, such as the understanding of the meaning of the information, the opinions about the implications of this information, or the concerns regarding the acceptability of given conclusions.
- *Statistical Reasoning* could be defined as the way people reason with statistical ideas and make sense of statistical information (Garfield & Gal, 1999). Statistical

reasoning involves interpreting and deducing based on sets of data, graphical representations, and statistical summaries. Combining ideas about data and chance, which leads to making inferences and interpreting statistical results, is the most common statistical reasoning practice. A conceptual understanding of important ideas, such as distribution, center, spread, association, uncertainty, randomness, and sampling stands as the foundation of statistical reasoning.

- *Statistical Thinking* includes "an understanding of why and how statistical investigations are conducted and the 'big ideas' that underlie statistical investigations" (Ben-Zvi & Garfield, 2004, p. 7). Statistical thinkers understand the omnipresent nature of variation and know how to use appropriate methods of data analysis (e.g. numerical summaries and visual displays of data) and when to use them. Understanding the nature of sampling, how we make inferences from samples to populations, and why designed experiments are needed in order to establish causation are elements of statistical thinking. Statistical thinking includes an understanding of how to simulate random phenomena using models, how to estimate probabilities by producing data, and how, when, and why to use existing inferential tools to aid an investigative process. Furthermore, being able to understand and utilize the context of a problem in forming investigations and drawing conclusions, and recognizing and understanding the entire process (from question posing to data collection to choosing analyses to testing assumptions, etc.) are also components of statistical thinking. Finally, statistical thinkers are those who are able to critique and evaluate results of a solved problem or a statistical study.

Delimitations

This study involves Grade 6-12 students in Central Texas. The students are those who have taken statistics and data analysis lessons and those who have never been taught formal statistics and data analysis lessons. The sample of participants was not chosen randomly; instead, a convenience sample was chosen among several groups of students in Central Texas. This convenience sampling method was chosen to minimize the cost of the study by choosing schools that are near Texas State University, the home institution of the researcher. Furthermore, as a preliminary effort to validate and measure the reliability of the developed survey instrument, a convenience sample is considered adequate.

Summary

Since the early 80s, reform movements have recommended increasing their content and rigor of statistics in K-12 school mathematics curriculum. Two important curriculum documents, the Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A Pre-K-12 Curriculum Framework (Franklin, et al., 2007) and the Common Core State Standards in Mathematics (National Governors Association Center for Best Practices, and Council of Chief State School Officers, 2010) actualize the reform by providing detail descriptions of what students should know and should be able to do in statistics. The descriptions follow the hypotheses of learning trajectories of statistical concepts that are based partially on research and learning theories. There is a need to empirically understand the learning development and growth of statistical concepts, particularly those related to the investigation cycle in statistics (formulating questions, collecting data, analyzing data, and interpreting results). Understanding the learning

trajectories of statistics is important for instructions and assessments of statistical concepts.

This study aims to understand how students learn statistics by describing the developmental growth of students' understanding of statistical concepts and learning trajectories of several concepts in statistics. To reach these goals, an instrument to measure students' developmental levels in learning statistics and learning trajectories of several statistical concepts was developed. Besides serving as an assessment tool, the instrument will potentially serve as a research and evaluation tool to better understand statistical learning trajectories. The instrument is expected to help teachers of statistics to diagnose their students' developmental readiness for learning statistical subjects. The instrument is, tentatively, a valid and reliable measure that may help researchers in statistics education to better understand how students develop their statistical literacy, reasoning, and thinking. The results may inform all stakeholders (teachers, parents, administrators, teacher educators, and researchers in statistics education) not only about the order and nature of the steps in the growth of students' mathematical understanding, but also about the nature of the effective instruction that might support them in moving step by step to become statistically literate high school graduates.

In the following chapter a review of literature is presented to build the theoretical framework of this study. A discussion of methodology applied in this study is presented in Chapter 3. The results of this study are discussed in Chapter 4 and the conclusions and implications derived from the results are presented in Chapter 5.

CHAPTER II

LITERATURE REVIEW

In this chapter, the theoretical framework of this study will be discussed thoroughly. In order to make a conjecture of learning trajectories of several statistical concepts, some results from previous studies on how students learn several statistical concepts and what misconceptions that they develop during the learning processes are discussed. The descriptions of how students learn statistical concepts provided by earlier studies combined with the learning trajectories of statistics suggested by the Pre-K-12 GAISE Framework (Franklin, et al., 2007) and the Common Core State Standards in Mathematics (National Governors Association Center for Best Practices, and Council of Chief State School Officers, 2010) were used as guidelines to build hypotheses of concepts that students should understand at certain developmental level, and also a conjecture of learning trajectories of several statistical concepts. These hypotheses were validated by expert panels and then empirically tested. Therefore, in this chapter a review of theories of learning trajectories, an analysis of previous studies' results on how students learn statistical ideas, and an examination on the Pre-K-12 GAISE Framework's and CCSS-M's hypotheses of learning trajectories of several big ideas in statistics are presented. The review consists of three sections described below.

The first section in this chapter is devoted to reviewing the theoretical background of the concept of learning trajectory in mathematics and statistics education. Some results of previous studies on several misconceptions of statistical ideas that students possess along their journey in learning statistics and studies on how students develop their understandings of such statistical concepts are also discussed in this section. In the second section, a summary of expectations and recommendations described in the Pre-K-12 GAISE Framework for instruction and assessment of statistical problem solving process components (formulating question, collecting data, analyzing data, and interpreting result) and of the nature and role of variability is presented. The differentiation of instruction and assessment of the process components for each developmental level (Level A, B, and C) is explained. In the third section, a description of statistical learning trajectories proposed by the CCSS-M is provided. It explains the expectations that are set for students in statistics and probability in the CCSS-M from Grade Six to Grade Twelve. The description also includes a brief explanation of the relation between the content standards in statistics and probability with the content standards for Measurement and Data that are set for students since Kindergarten. The content standards provided by the CCSS-M combined with the Pre-K-12 GAISE Framework and results from studies in learning trajectories and assessment in statistics education were used as the theoretical foundation in building hypotheses of statistical content domain. The content domain is aligned with each developmental level. The CCSS-M, Pre-K-12 GAISE Framework and results from previous studies were also used in building hypotheses about the order and nature of the steps in the growth of students' understanding of the concepts defined in the content domain.

Learning Trajectories of Mathematical & Statistical Concepts

The concept of *learning trajectories* in mathematics is a concept that provides an approach to develop a mastery to define the path that students go through in learning mathematical concepts (Daro, Mosher, & Corcoran, 2011). The concept of learning trajectories that is commonly labeled as *learning progressions* in other educational subjects identifies key waypoints along students' learning path. The key waypoints are several passageways in which students' knowledge and skills are likely to grow and develop in school subjects (Corcoran, Mosher, & Rogat, 2009). They provide empirical-based information about when the best time is to teach a specific concept to a particular type of students. Clements and Sarama (2004) define Learning Trajectory in mathematics as:

> the descriptions of children's thinking and learning in a specific mathematical domain, and a related conjectured route through a set of instructional tasks designed to engender those mental processes or actions hypothesized to move children through a developmental progression of levels of thinking, created with the intent of supporting children's achievement of specific goals in that mathematical domain. (p. 83)

These trajectories involve hypotheses about the order of the steps in the growth of students' mathematical understanding and the nature of each step. The trajectories also involve hypotheses about the nature of instructional experiences that might support the growth of students' mathematical understanding in each step. All conceptions of trajectories or progressions based on the common facts that students' knowledge and skill in any domain start out small in the amount and complexity, then becomes much larger

over time due to effective instruction, and that the amount of growth clearly varies with experience and instruction but also seems to reflect factors associated with maturation, as well as significant individual differences in abilities, dispositions, and interests (Daro, Mosher, & Corcoran, 2011).

Learning trajectories are "based on research of how students' learning actually progresses – as opposed to selecting sequences of topics and learning experiences based only on logical analysis of current disciplinary knowledge and on personal experiences in teaching" (Corcoran, Mosher, & Rogat, 2009, p. 8). Instead of being grounded mostly in the disciplinary logic of mathematics and the conventional wisdom of practice, the hypotheses of learning trajectories are rooted in actual empirical studies of the ways in which students' thinking grows in response to relatively well specified instructional experiences. Learning trajectories focus on identifying significant clusters of concepts and connections in students' thinking that represent key steps forward. As a result of these practices, learning trajectories offer a stronger basis for describing the short-term goals that students should meet in order to reach the common core college and career ready high school standards. Daro, Mosher, and Corcoran (2011) point out that in addition to their stronger basis in describing students' interim goals in learning, trajectories also provide understandable points of reference for designing assessments both for summative and formative purposes that can report where students are in terms of steps in learning growth, rather than reporting only in terms of students' position on performance scale in comparison with their peers.

Daro et al. (2011) review a few investigations in developing assessments that reflect what we know or can hypothesize about students' learning trajectories in

mathematics. According to Daro et al. (2011), one example of the investigations were conducted by Jere Confrey and Alan Maloney at North Carolina State University (NCSU) who were working on assessments that reflect their conception of a learning trajectory for "equipartitioning" as part of the development of rational number reasoning. The author began with an extensive synthesis of the existing literature and supplemented it by conducting cross sectional clinical interviews and design studies to identify key levels of understanding along the trajectory. They then developed a variety of assessment tasks designed to reflect the hypothesized levels. Working with Andre Rupp, a psychometrician at the University of Maryland, they examine students' performances on the tasks using item response theory (IRT) models to see if the item difficulties and the results of alternative item selection procedures produce assessments that behave in the ways that would be predicted if the items in fact reflect the hypothesized trajectory and if that trajectory is a reasonable reflection of the ways students' understanding develops. In this study, a similar method is applied by developing a variety of assessment items to reflect the hypothesized Pre-K-12 GAISE Levels. This study, however, uses structural equation modeling (SEM) models instead of IRT models, since based on the hypothesis, there are three constructs, in this case GAISE Levels, involved in the data. Thus the existence of the three constructs violates the assumption of the unidimensionality of the data, one of the fundamental assumptions with IRT models.

Other researchers and assessment experts in the U.S. are working on the development of similar assessment tools (Daro et al., 2011). A major development of assessments based on more complex conceptions of how students actually learn and produce results that can be more legitimately interpreted in terms of what students

actually know and can do. This leads to much more effort and resources being devoted to solving the problems of developing usable assessments. Researchers and assessment experts seek to develop measures that will report in terms of much more complex conceptions of student learning. They intend to provide students' reports that include not only facts and concrete skills, but also understanding, and ability to use knowledge and to apply it in new situations. The reports that can be used to determine whether students are on the right pathways over the earlier grades to be able to meet the "college-and career-ready" core standards by sometime during their high school years.

According to Daro et al. (2011), Jere Confrey and Alan Maloney from North Carolina University recommend to develop diagnostic assessments that can be used more formally to support and enhance formative assessment practices as another option to having research on learning trajectories directly influence practice through teacher knowledge. In the latter work, the authors seek a means to develop measures and ways of documenting students' trajectories to track students' progress both quantitatively and qualitatively. On the other hand, other scholars including Daro et al. argued that such assessments certainly could be useful, but stressed their belief that effective formative use would still require teachers to understand the research on mathematics learning that supports the conceptions of students' progress. Teachers need to understand research that provides the basis for the assessment designs, and also to know the evidence concerning the kinds of pedagogical responses that would help the students, given what the assessments might indicate about their progress or problems.

Daro et al. (2011) admit that it is extremely important to design a large-scale assessment whose reports would be more informative because they are based on sound

theories about how students' learning progresses. Daro et al. argue that it also will be crucial to continue to focus on developing teachers' clinical understanding of students' learning in a form that enable teachers to interpret and respond to student progress and informs how they implement the curricula they use. Teachers could assess on a daily basis a different grain size of progress from the levels that large-scale assessments used for summative assessments are likely to target. Large scale assessment will tend to reference bigger intervals or significant stages of progress to inform policy and the larger system, as well as to inform more consequential decisions about students, teachers, and schools. Nonetheless, a correspondence between the conceptions of student progress that teachers use in their classrooms and the conceptions that underlie the designs of large-scale assessments need to be initiated. In order for teachers to be able to put their efforts into preparing their students to meet the large assessment expectation, they need to be informed about where their students have been before and where they are heading. The larger picture informing the assessment designs would help the teachers get the information that they need. Furthermore, it should be helpful and reassuring to teachers if the assessments that others use to evaluate their and their students' performance are designed in ways that are consistent with their understanding of students' progress, so that they can have some confidence that there will be agreement between the progress they observe and the progress reported by these external assessments. Also, it would certainly be desirable if those external reports were based on models that provide real assurance of the validity and reliability of the measures used by the external assessments. In this study, this recommendation is addressed by examining previous studies and their relation to learning trajectories.

Discussion on how students develop their statistical thinking and reasoning in statistics and data analysis is provided in the following parts of this section. As the result of reform movements in mathematics education that include more statistics in Pre-K-12 mathematic curricula, there is a strong demand for teachers to put a greater emphasis on helping students develop their statistical thinking and reasoning in statistics and data analysis lessons (NCTM, 1989, 2000; National Governors Association Center for Best Practices, and Council of Chief State School Officers, 2010). Shaughnessy (2007) explains that most of the recent research is in the area of: (1) students' knowledge and reasoning about statistics, (2) teachers' knowledge of statistics, and (3) teaching practices in statistics. He described that research on students' understanding of statistics has focused on particular concepts or big ideas in statistics such as *gathering information from sample*, *centers* (*averages*), *variation or variability*, *comparing data sets*, *and students' understanding of graphs*. Observing from a statistical investigation process perspective, the following discussion begins by *studies on students' understanding of sample and data collection*, the skill that students need in collecting data. The discussion is followed by a discussion on *studies on students' understanding of measures of centers, and studies on students' understanding of graph,* the ideas that students need to understand in order to be able to analyze data. The next section presents a discussion on *students' thinking about association, covariation, and correlation* that is related with students' ability in *comparing data sets,* the skills that students need to develop in order to be able to interpret result. The last discussion regards *studies on students' thinking of variability* that is very important in preparing them to be statistical literate citizens. There

are almost no studies focused on students' understanding of formulating statistical questions.

Studies on students' understanding about sampling and data collection

In investigating Grade 4 and 5 students' understanding of sampling in surveys, Jacobs (1997, 1999) found four categories of children's evaluations of survey methods: *potential for bias, fairness, practical issues, and results.* Some students recognized the potential for bias in certain survey methods, but other students were more concerned with fairness issues. For many students, all possible subgroups in a survey population should have had representatives in the survey samples to assure the fairness of the survey. Jacobs found that students who favored the fair-sample approach would reject any part of randomization. Jacobs also found that some students tended to disagree with results of surveys if they did not match the students' own preconceived notions of what the results should be.

Jacobs' (1997, 1999) tasks included questions that asked students to evaluate three different survey techniques: restricted, self-selection, and random. In one task, Jacobs presented students with six different survey settings in school on gathering students' opinion on whether they were interested in conducting a raffle at a school to raise money. This task was then used by Watson & Callingham (2003) with Grade 3 to 9 students as part of their study to measure students' understanding of sources of variation. Shaughnessy (2003) also used a version of this task with secondary students to see if the students could identify important aspects of sampling. The task is presented in Figure 1 below.

Part 1. *A class wanted to raise money for their school trip to Disney World. They could raise money by selling raffle tickets for a Nintendo Game system. But before they decided to have a raffle they wanted to estimate how many students among the population of the entire school would buy a ticket.*

So they decided to do a survey to find out first. The school has 600 students in grades 7 – 12; 100 students per grade.

How many students would you survey and how would you choose them? Explain why?

Part 2. *Three students in the school suggested different methods to survey the students in the school about buying the raffle tickets.*

a) ***Shannon*** *got the names of all 600 children in the school and put them in a hat, and then pulled out 60 of them. What do you think of Shannon's survey?*

b) ***Raffi*** *surveyed 60 of his friends. What do you think of Raffi's survey?*

c) ***Claire*** *set up a booth outside of the cafeteria. Anyone who wanted to stop and fill out a survey could. She stopped collecting surveys when she got 60 kids to complete them. What do you think of Claire's survey?*

(After each of these sampling methods, students were asked to rate the method, and to give a reason for their rating).

☐ *GOOD* ☐ *BAD* ☐ *NOT SURE*

Why?

d) *Who do you think has the best survey method? Why?*

Figure 1. A version of Jacobs' task on sampling procedures.

Responses from the students who were given the task showed that only about a third of the students indicated that they had a statistically appropriate sampling plan. Many of the students wanted to survey all, or at least most of the students in the entire school. Students heavily favored the self-selection approach outside the cafeteria (64%), meanwhile a third of these students felt that asking friends was a good way to get a sense of the opinion in the school. Students wanted to predetermine the survey results and asking friends was a way to make this happen. Only 12% of the students recognized the potential for bias in the self-selection method outside the cafeteria such as the fact that

some groups of students might not eat lunch in the cafeteria.

Only about a third of the students surveyed preferred the random sampling approach while over half of them preferred the self-selection method outside the cafeteria as the best method, "because everyone has the same chance this way" (Shaughnessy, 2007). These results are consistent with the results found by Jacobs (1997) and Watson & Moritz (2000a, 2000b).

Jacobs (1997) found that the fairness criterion is prevalent among younger students and Shaughnessy (2003) found that this prevalence is still quite robust among older students. In a series of studies to investigate whether students knew what a sample was, whether they would be sensitive to sample size, and whether they would recognize the possibility for bias in real sampling situation from the media, Watson and Moritz (2000a) found that when asked "If you were given a sample, what would you have?", students' thinking ranged from personal examples such as "samples of food" or "blood sample," to the notion of a piece such as "a little bit" or "a small portion," to the idea that a sample should be a representative piece of something larger. The second question that Watson and Moritz (2000a) asked in the same study is whether they would put more faith in a friend's recommendation for a car purchase, in the recommendation of 'Consumer Reports' magazine, or it did not matter one way or the other. The other two questions that Watson and Moritz asked for the students in their longitudinal study are based on articles from a newspaper. One article claimed that over 90% of those who phoned in on a survey were in favor of legalizing marijuana, and another article generalized a claim that "6 of every 10 students from a sample in Chicago could easily bring a handgun to school" to all of the United States. So, the car purchase, legalizing marijuana, and handgun contexts

had the clear potential for bias. Watson and Moritz (2000a) found that there is a progression in students' thinking about samples in which students first (a) do not distinguish between sample and population, then (b) recognize the difference between sample and population but really wanted to sample anyone, and finally (c) realize that samples can be used to represent the population and to estimate population parameters. Watson and Moritz found that 50% of the students improved on the four questions after 2 years, and 75% had improved from their initial responses after 4 years. Students' understanding of the concept of sample seems have been improved following the cumulative school experience and outside world experience. The learning trajectory of sampling and data collection can be described in a path diagram as displayed in Figure 2.

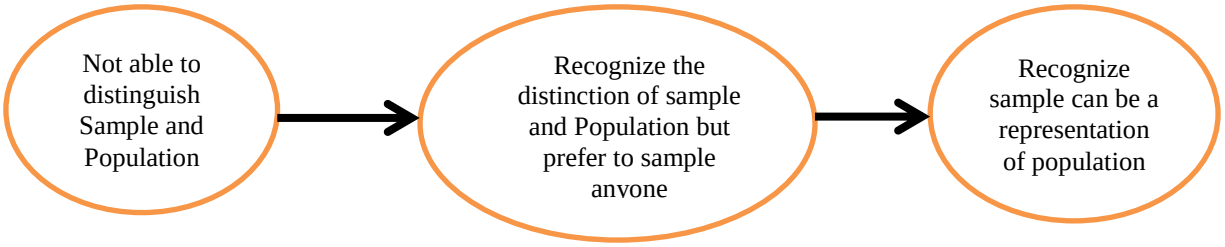


Figure 2. The learning trajectory of sampling and data collection.

Studies on students' understanding of measures of center

Studies focusing on choosing an appropriate measure of center to represent a data set found that school students frequently make poor choices in selecting measures of center to describe data sets (Groth, 2003; Konold & Pollatsek, 2002; Zawojewski & Shaughnessy, 2000). The research studies suggested a need for teachers to be conscious of the difficulties students had in understanding and reasoning about measures of center beyond a computational level. According to Shaughnessy (2007), early studies on

averages were mostly done with college students (e.g., Pollatsek, Lima, & Well, 1981; Mevarech, 1983; Pollatsek, Konold, Well, & Lima, 1984). Pollatsek, Konold, Well, and Lima (1984) found that, given the population mean (400) for SAT scores and a rather extreme data value from a sample of SAT scores, college students did not take the extreme value into account. They merely adjusted their prediction for a sample mean of 400. Sometimes they believed that the mean was the most likely result to occur in a sample, even if the mean itself was not a possible data point. The study also found that college students tended to find the midpoint of two sample means and consider the midpoint as the mean of combined sample, even though the two samples given have unequal sample sizes. This "average of averages" practice is a common mistake done by college students and referred to as "the closure misconception" by Mevarech (1983). These early studies on students' understanding of the mean indicated that many college students considered mean as something to be calculated, a very procedural understanding.

Many students at all levels of education faced the same difficulties in understanding the concept of average (Konold & Higgens, 2003; Groth & Bergner, 2006; Mokros & Russell, 1995; Shaughnessy, 1992, Shaughnessy, 2003; Watson & Moritz, 1999a, 1999b). The students demonstrated a lack of understanding the mean, and could only state how to find it arithmetically (Mathews & Clark, 2003; Clark, Mathews, Kraut, & Wimbish, 1997). Students also demonstrated difficulties in determining the median of data sets, especially when the data sets were presented graphically or in unordered lists (Bright & Friel, 1998; Zawojewski & Shaughnessy, 2000).

One of the first studies investigating younger students' (Age 8 – 14) conceptions of average was conducted by Strauss and Bichler (1988). They identified that students

were quite aware that the mean had to be located between the extreme data values, and that the mean was influenced by particular values in a data set. However, it was extremely difficult for their students to understand that the mean was the value that minimizes deviations, and the zero data value must also have been included and accounted for when calculating the mean. Strauss and Bichler (1988) found that the way children think about the concept of the mean was not the same as the way statistically mature adults do.

In a study to investigate young students' conceptual understanding of averages, Mokros and Russell (1995) found that students who focused on modes in data sets had difficulty in constructing a data set if they were given the mean of the data but they were not allowed to use the mean as a data value. They concluded students focused on modes saw only individual data values and did not see the distribution of the data as an entity. A similar result was reported by Cai (1995) who found that students had great difficulty in filling missing values in the data set when given the mean, even though they could calculate a mean of a data set when given all the data.

Mokros and Russell (1995) also found students had a good sense of average as a midpoint even though they might not have known what a median was. They found that some students worked backwards to construct a distribution by symmetrically choosing values above and below the average. Similar to students who focused on modes, the students also had great difficulty when they were not allowed to use the average itself as a data value. Mokros and Russell (1995) concluded that higher-level conceptions of average for students may be developed by scaffolding the concepts through instructional interactions.

Konold and Pollatsek (2002) postulated four conceptual perspectives for the mean: mean as a *typical value*, mean as *fair share*, mean as a *data reducer*, and mean as a *signal amid noise*. They argued that, statistically speaking, mean as a signal amid noise was the most important and the most useful conception of the mean, because it was the most helpful conception in comparing two data sets. They also argued that other conceptions of mean such as mean as a typical value and mean as a fair share were not as powerful in comparing groups and therefore, should not be emphasized with students. Shaughnessy (2007) summarized that Konold and Pollatsek's (2002) conceptions of mean as a typical value and mean as fair share were more closely tied to a data analysis perspective, while mean as a signal was more closely connected to decision-making in statistics. Shaughnessy (2007) added that data reduction was necessary to locate an informative signal amid the noise of variability that is important as the basis in making decision.

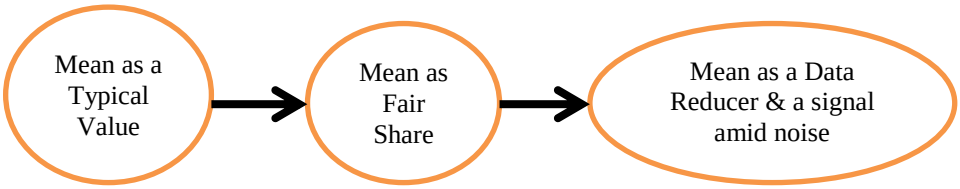


Figure 3.The learning trajectory of the concept of mean (average).

Shaughnessy (2007) argued that from a normative point of view, Konold and Pollatsek's (2002) argument may have been good. However, from Mokros and Russell's (1995) as well as Watson and Moritz's (1999b) research findings, mean as fair share and mean as a typical value were perhaps better introduced first as measures of center, because they built on students' primary intuition. The mean as a data reducer required more sophistication from students and a willingness on their part to let go of some pieces

of information that were considered as noise. Shaughnessy (2007) believed that teachers and students must spend more time focusing on the noise itself before determining both special-cause and common-cause variation in the data. The learning trajectories of the concept of center or averages suggested by the previous studies are depicted in a diagram shown in Figure 3.

A small object was weighed on the same scale separately by nine students in a science class. The weights (in grams) recorded by each student are shown below.
6.3 6.0 6.0 15.3 6.1 6.3 6.2 6.15 6.3

Of the following methods, which would you recommend they use?

A. Use the most common value, which is 6.2.
B. Use the 6.15 since it is the most accurate weighing.
C. Add up the nine numbers and divide by 9.
D. Throw out the 15.3, add up the other 8 numbers and divide by 8.

(Adapted from Statistical Reasoning Assessment (SRA), Garfield 2003)

Figure 4. Example of mean (average) as a signal amid noise item.

An example of an item that assesses students' understanding of the concept of mean (average) is presented in Figure 4. Besides the concept of average, the concept of variation or variability is also important in statistics. A discussion on students' understanding of variability is presented in a later section.

Studies on students' understanding of graphs

Friel, Curcio, and Bright (2001) defined graph comprehensions as "the ability of graph readers to derive meaning from graphs created by others or by themselves" (p. 132). They discussed three factors contributing to the ability of people to make sense of graphs: *visual perception, the characteristics of graph readers, and the experience with statistics*. In an analysis of the students' results on graph item from the 1996 NAEP,

Zawojewski and Shaughnessy (2000) found that students performed well when reading information represented in pictographs and stem-and-leaf plots; however, Grade 12 students' abilities to read and interpret histograms or box plots that required some proportional reasoning were lower than their performance with other graphical representation. Based on the results of 1996, 2000, and 2003 NAEP administrations, students could read graphs fairly well, but they faced difficulty in interpreting graphs, and were not able to make predictions based on graphical information.

Studies focused on how students learn about graphical representation of distributions found that students have an easy time in understanding case-value plots, where a bar or line represents an individual case, however, students revealed confusion in interpreting histograms, where a bar represents multiple cases (delMas, Garfield, & Ooms, 2005). This confusion led students to try describing shape, center, and variation of case-value plots (delMas et al., 2005) or to think that bars in histogram indicated the magnitude of single values (Bright & Friel, 1998). These studies suggested that students should have been given repeated opportunities to compare and reason about multiple representations of the same data set (Bakker & Hoffman, 2005; delMas et al., 2005).

Research studies also found that students tended to use graphs as illustrations rather than as reasoning tools to learn something about a data set or gain new information about a particular problem or context (Wild & Pfankuch, 1999; Konold & Pollatsek, 2002). Current research on student statistical understanding of distribution recommends a shift of instructional focus from drawing various kinds of graphs and learning graphing skills to making sense of the data, detecting and discovering patterns, confirming or generating hypothesis, noticing the unexpected and unlocking the stories in the data

(Ben-Zvi & Amir, 2005; Pfannkuch, 2006; Pfankuch & Reading, 2006; Reading & Reid, 2006).

Friel, Curcio, and Bright (2001) identified six behaviors that students should possess in understanding graph representations that were summarized by Shaughnessy (2007):

1. Recognizing components of graphs (*Reading the data*).
2. Speaking the language of graphs (*Reading the data*).
3. Understanding relationships among tables, graphs, and data (*Reading within the data*).
4. Making sense of a graph, but avoiding personalization and maintaining an objective stance while talking about the graphs (*Reading within the data*).
5. Interpreting information in a graph and answering questions about it (*Reading beyond the data*).
6. Recognizing appropriate graphs for a given data set and its context (*Reading beyond the data*) (p. 991).

Shaughnessy (2007) added two more behaviors that he claimed fell under the level of reading *behind* the data:

7. Looking for possible causes of variation (*Reading behind the data*).
8. Looking for relationships among variables in the data (*Reading behind the data*) (p. 991).

Results from studies on graph sense indicated that students had poor graphical interpretation skills and were often unable to reason beyond graphs (Shaughnessy, 2007). Shaughnessy also noted that reading, reading within, and reading behind the data was

critical to making connections between the context and the data as well as to developing statistical literacy, reasoning, and thinking. Based on the above explanations, a learning trajectory of understanding graphs can be described in a path diagram as displayed in Figure 5 below.

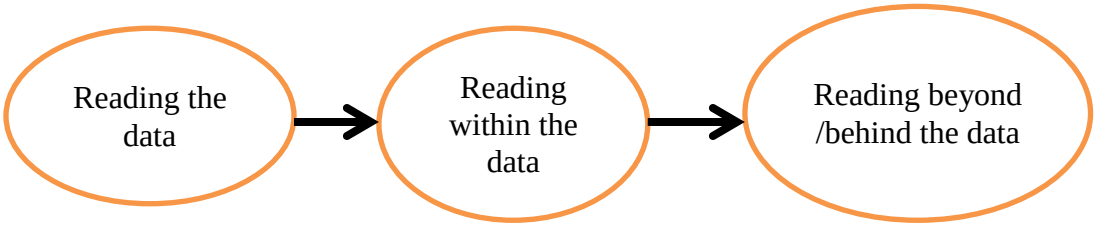
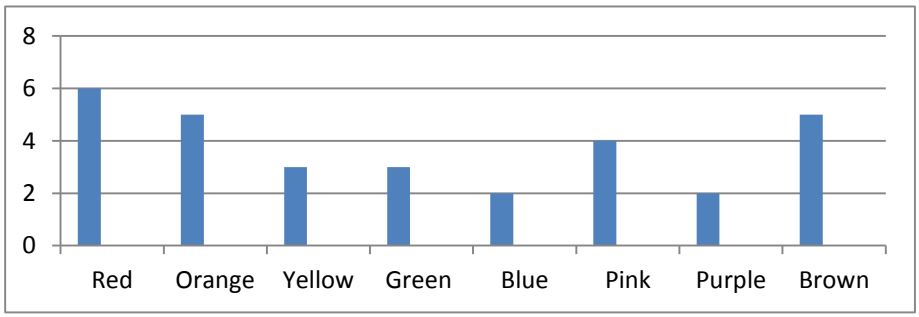


Figure 5. The learning trajectory of graph representations.

An example of items developed to measure students' understanding of graph representations is displayed in Figure 6 below.

Robert's mother lets him pick one candy from a bag. He can't see the candies. The number of candies of each color in the bag is shown in the following graph.



What is the probability that Robert will pick a red candy?

A. *10%*
B. 20%
C. *25%*
D. *50%*

(Adapted from PISA Assessment 2009, OECD, 2009).

Figure 6. An example of items to measure students' understanding of graph representation.

Studies on students' thinking about association, covariation, and correlation

Problems in statistics usually not only involve univariate variability, but also multivariate situations in which association, dependence, and possible causes of variation lurk, all of which could be discovered by a persistent effort to uncover the information hiding in the data. Watson and Moritz (1999a) asked students in Grades 6 and 11 to produce graphical representations of a nearly perfect relationship between an increase in heart attack and an increase in motor vehicle use as described in a newspaper article. They found that most of the students that participated in this study just accepted the claim that there was an association between driving a car and having a heart attack without questioning it. Probably because of their inexperience in critiquing statistical claims about relationships between variables, students too often just accept anything that they read in the newspaper.

Nemirovsky (1996) suggested that covariation might have been best introduced with time as one of variables, because students were interested in trends over time, and this type of data connected naturally to topics that were of interest to them. Shaughnessy (2003) also used graphs over time in interviews and in classroom teaching episodes with middle and secondary level students to investigate students' awareness of potential causes of variability in food consumption over time. Shaughnessy (2003) reported that most students that participated in this study did make some conjectures about the variability over time in such per capita food consumption graphs. If they could provide a rigorous contextual explanation for the humps and dips in the food consumption graphs over time, students would come up with explanations that make more sense to them. They rarely considered that random variation could have been the cause of the jumps and

the dips in the food consumption graphs. They tended to look for “special cause” (Wild & Pfankuch, 1999) variation such as the baby boom, improved production and distribution of food, the Depression, World War II, the war in Vietnam, and some students claimed “it must have been the hippies” as the cause when all else failed. Students did try to make contextual conjectures for why such graphs vary.

Batanero, Estepa, Godino and Green (1996) studied pre-university students' preconceptions about statistical association by analyzing students' strategies in determining the association from a mathematical perspective. They identified three misconceptions of statistical association:

- Determinist conception of association. Students who have deterministic conception of association expect a correspondence that assigns only a single value in the response variable for each value of the explanatory variable. If the data do not show this correspondence, they consider there is no dependency between the variables. They expect that the correspondence between the variables must be, from the mathematical point of view, a function.
- Unidirectional conception of association. Students who have unidirectional conception of association perceive dependence only when the sign is positive (direct association), and they consider an inverse association (negative sign) as independence.
- Local conception of association. Students who have local conception of association form their judgments using only parts of the data provided in the problem, not the whole data. If this partial information confirms a given type of association, the students adopt this type of association as the right conclusion.

- Causal conception of association. Students who have causal conception of association only considered the association between the variables if this could be attributed to a causal relationship between them.

Estepa, Batanero, and Sanchez (1999) investigated students' ability to make associations between two variables of data set represented in two-way tables. Some students used deterministic strategies, like comparing lowest and highest values, comparing ranges, looking at coincidences, or using their personal belief in their arguments. Other students used statistical approaches such as means, totals, percentages, or attempted to compare the whole distributions. Gal (1998) discussed a variety of levels of questions to assess students' understanding of data in two-way tables. Since individual cell frequencies were inadequate to support opinions or to defend claims that are made about data in two-way tables, Gal suggested that percentages were needed to support the explanation. More open-ended, less directive types of questions about two-way tables should have been given to students to promote a higher level of statistical reasoning by students.

Not many studies on students' understanding of association, co-variation, and correlation have been conducted, especially with middle school or high school students. This leads to unclear explanations of how students develop their understanding of association, co-variation, and correlation and also the learning trajectory of those concepts. These concepts, however, are very crucial for interpreting result based on data. Therefore, in developing item to measure students' understanding of association, co-variation, and correlation, Estepa, Batanero, and Sanchez's (1999) investigation of students' ability to make associations between two variables represented in two-way

tables influenced most of the items developed in this study. An example of items to measure students' thinking about association, co-variation, and correlation is presented in Figure 7.

A group of 649 men with lung cancer was identified from a certain population in England. A control group of about the same size was established by matching these patients with other men from the same population who did not have lung cancer. The matching was on background variables such as ethnicity, age, and socioeconomic status. The summary of level of smoking and the number of lung cancer and control cases is given in the following table.

Cigarettes/Day	***Lung Cancer Cases***	***Control***	***Probability of Lung Cancer***
0	*2*	*27*	*2/29 = 0.07*
1 – 14	*283*	*346*	*283/629 = 0.45*
15 - 24	*196*	*190*	*196/ 386 = 0.51*
25 +	*168*	*84*	*168/252 = 0.67*

What is the association between the level of smoking and the number of lung cancer cases that can be inferred by the given data?

A. *A decrease in the lung cancer rate is associated with an increase in cigarette smoking.*
B. ***An increase in the lung cancer rate is associated with an increase in cigarette smoking.***
C. *An increase in the lung cancer rate is associated with a decrease in cigarette smoking.*
D. *There is no association between the level of smoking and the number of lung cancer cases.*

Figure 7. An example of items to measure students' understanding of association, co-variation, and correlation.

Studies on students' understanding of variability

Variability is the fundamental component of statistical thinking (Pfannkuch, 1997; Pfannkuch & Wild, 2004; Shaughnessy, 1997). Variability is what makes decisions in the face of uncertainty so difficult. Statistics becomes so challenging and interesting because of the presence of variability. Variability allows us to interpret, model and make predictions from data (Gould, 2004). Moore (1992) suggested that variability should be integrated, revisited, and highlighted in statistics curriculum and instruction.

Garfield and Ben-Zvi (2007) presented seven areas of knowledge of variability including the key ideas of each area as below:

(1) Developing intuitive ideas of variability that includes recognizing that variability arises everywhere; in data (qualitative or quantitative variables), in samples, and in distributions. Individuals have varied characteristics, and repeated measurements on the same characteristics could have variation. Shaughnessy (2007) argued that since variability occurs within many levels of statistical objects, students needed to develop their intuition for what is a reasonable or an unreasonable amount of variability in these objects. There is little variation in certain things, and there is a lot of variation of other things. Variation occurs within samples and distributions and also across samples and distributions. Data should be considered as an entity, rather than as individual points or as a combination of center and extreme values.

(2) *Describing and representing variability.* Using graphs of data to show variation in data may reveal patterns to help us focus on global features of distributions and identify the signal in the noise. It is important to study more than a single graph of a data set, because different graphs may reveal different aspects of variability in the data set. We

can use one number to represent a global feature (such as variability) of the distribution. Different numerical summaries tell us different things about the variability of a data set. For instance, while the Range informs us of the overall variability from highest to lowest value, the Standard Deviation (SD) informs us of the typical deviation from the mean. The Interquartile Range (IQR) informs us of the variation of the middle half of a distribution. While the IQR and SD tell us about variability of data, they are most useful for interpreting variability when we also know the related measure of center (mean for SD, median for IQR) as well as the general shape of the distribution. Measures of variability and center (as long as we consider them together) are more or less informative for different types of distribution. For example, the mean and SD tell us useful information about symmetric distributions, in particular, the normal distribution. For skewed distributions, the median and IQR are more useful summaries.

(3) *Using variability to make comparisons*. In comparing two or more data sets, it is helpful to examine their graphs on the same scale, as this allows us to compare the variability and speculate on why there are differences in the data sets. Using global summaries of variation and center when comparing groups is more helpful rather than comparing individual data points or 'slices' of the graphs. Examining both the variability within a group and the variability between groups and distinguishing these two types of variability are important.

(4) *Recognizing variability in special types of distributions*. In a normal distribution, the mean and SD provide useful and specific information about variability. There is variability in a bivariate data distribution, and we need to consider the variability of both variables as well as the variability for y values given individual values of x. The

variability of a bivariate data set (covariation) may reveal a relationship between the variables and whether we might be able to predict values of one variable (*y*) for values of the other (*x*).

(5) *Identifying patterns of variability in fitting models*. In fitting models and judging the fit of models (e.g., fitting the normal curve to a distribution of data, or fitting a straight line to a scatterplot of bivariate data), there is variability involved. The variability of the deviations from the model (residuals) can tell us about how well the model fits the data. Data may sometimes be reorganized and transformed to better reveal patterns or fit a model.

(6) *Using variability to predict random samples or outcomes*. Samples vary in some predictable ways, based on sample size and the population from which they are drawn and how they are drawn. If we have random samples the variability can be more readily explained and described. Larger samples have more variability than smaller samples, when randomly drawn from the same population. However, sample statistics from the larger samples vary less than statistics from smaller samples. There is variability in outcomes of chance events. We can predict and describe the variability for random variables. In some situations we can link the variability in samples to variability in outcomes, making predictions or statistical inference.

(7) *Considering variability as part of statistical thinking*. In statistical investigations, we always need to begin by examining and discussing the variability of data. Data production is designed with variation in mind. Aware of sources of uncontrolled variation, we avoid self-selected samples, insist on comparison in experimental studies, and introduce planned variation into data production by the use of randomization (Moore,

1990). In statistical analysis we try to explain variation by seeking the systematic effects behind the random variability of individuals and measurements (Moore, 1990). The ideas listed above are all part of statistical thinking, and come into play when exploring data and solving statistical problems (Wild & Pfannkuch, 1999).

The seven areas presented above will become the domain of statistical ideas on variability that will be assessed in the instrument developed for this study. Several components of variability were also identified by Wild and Pfannkuch (1999) in their model of statistical thinking that are consistent with those suggested by Garfield & Ben-Zvi (2007): *acknowledging*, *measuring*, *explaining*, and *controlling* variation. Reading and Shaughnessy (2004) added two other aspects of variation: *describing*, and *representing* variability, meanwhile Canada (2004) provided a detailed framework for analyzing students' thinking about *noticing*, *describing*, and *attributing causes of* variation. Reading & Shaughnessy (2004) reported a description hierarchy and a causation hierarchy for variability based on students' response on several tasks on variability. They gave an example that in the description hierarchy, lower level responses might be concerned only with either outliers or middles (*uni-structural*), while a higher level response might mention both middles and extremes (*multi-structural*). At an even higher level, a student response might discuss the deviations of data from some fixed value such as mean or median, and making connections between the concepts of center and variability (*relational*). Based on the studies described above, a learning trajectory model for the concept of variability is described in the following paragraph.

Using Reading and Shaughnessy's (2004) terms, the lowest level in understanding variability is called the uni-structural level. In this level students have developed the

intuitive ideas of variability and are able to describe and represent variability using graphs and might be concerned only with either outliers or middles. The middle level is called the multi-structural level. In this level students have developed the ability to use variability to make comparison, recognize variability in special types of distribution, and identify patterns of variability in the data by mentioning both middles and extremes. The highest level is called the relational level. In this level students have developed the ability to use variability to predict random samples or outcomes, and consider variability as part of statistical thinking. Students in this level are also able to discuss the deviations of data from some fixed value such as mean or median, and make connections between the concepts of center and variability (relational). This learning trajectory of variability described above can be presented as a path diagram as displayed in Figure 8.

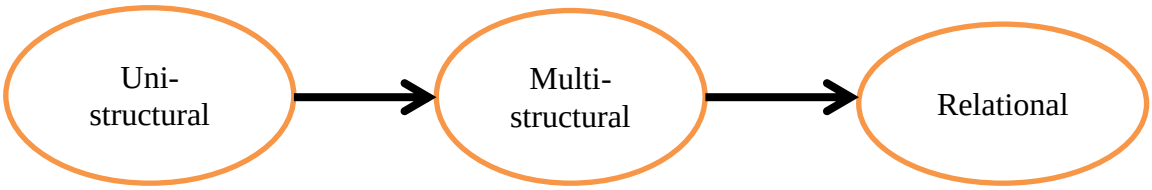


Figure 8.The learning trajectory of the concept of variability.

After understanding the learning trajectory of the concept of variability, it is also necessary to understand how to measure students' thinking of variability. In the following paragraphs, a discussion on how to measure students' understanding of variability is presented. In order to explore students' thinking about the variability of the data in a sampling situation, several problems, called Lollie problems, were administered in United States, Australia, and New Zealand by Shaughnessy, Watson, Moritz, & Reading as cited by Shaughnessy (2007). Three hundred students in grades 4-6, 9, and 12 were

given three different versions of the Lollie task (List, Choice, and Range). Students were asked to give the reasons behind their answers. This task has subsequently been administered to thousands of students in grades 3-12, primarily in Australia and the United States (Reading & Shaughnessy, 2004; Torok & Watson, 2000). The problems are given in Figure 9 below.

A bowl has 100 wrapped lollies in it. 20 are yellow, 50 are red, and 30 are blue. They are well mixed up in the bowl. Jenny pulls out a handful of 10 lollies, counts the number of reds, and records it on the board. Then Jenny puts the lollies back into the bowl, and mixes them all up again.

Four of Jenny's classmates, Jack, Julie, Jason, and Jerry do the same thing. One at a time they pull ten lollies, count the reds, and write down the number of reds, and put the lollies back in the bowl and mix them up again.

What do you think? (List Version)
I think the number of reds the students pulled were

_____ _____ _____ _____ _____

I think this because: _____

I think the list for the number of reds is most likely to be (circle one)
8, 9, 7, 10, 9
3, 7, 5, 8, 5
5, 5, 5, 5, 5
2, 4, 3, 4, 3
3, 0, 9, 2, 8 (Choice Version)

I think this because: _____

I think the number of reds went from (a low of) _____ *to a high of* _____

I think this because: _____ (Range Version)

(Shaughnessy, 2007, p. 975).

Figure 9. A *version of the Lollie Task.*

Shaughnessy (2007) explained that there are five types of responses of Lollie Tasks that indicated differences among students on how they acknowledge variability in samples: *high*, *low*, *wide*, *narrow*, and *reasonable*. Some students predicted all high numbers of reds, like 6, 7, 5, 8, 9, mostly numbers above the expected value of 5 and the students reasoned that there were "a lot of red in there, so it (the red ones) will happen a lot." Some students, mostly in Grade 4, predicted all low numbers (all numbers ≤ 5) and reasoned there were a lot of "non-reds" in the mixture that would prevent the reds from being pulled very often. Other students predicted a wide list of outcomes, for example, 1, 5, 7, 9, 2 (range ≥ 8) because "any result could occur, you never know," suggesting that they may have been using "outcome approach" reasoning (Konold, 1989) or an equi-probability conception (Lecoutre, 1992).

Some other students, frequently among Grade 12 students, predicted a very narrow list for the numbers of reds, for example, 5, 5, 5, 5, 5, or 5, 6, 5, 5, 6 (range ≤ 1). They reasoned that "5 is the most likely outcome" or "5 is what you are supposed to get." Reading & Shaughnessy (2000) found that the "narrow" predictors were reticent to change their answers, even after they acknowledged that it was unlikely to have identical repeated samples. The reticence might be caused by the student's understanding of the theoretical probability of getting reds (Reading & Shaughnessy, 2000). The last type of responses of the Lollie tasks were those that were "reasonable," where the list of number of reds were distributed in a more normative way around 5, such as 3, 7, 5, 6, 5, centered around the expected value within a reasonable range.

All studies discussed previously have identified how students develop their understanding of several big ideas in statistics and the learning trajectories of the ideas.

These identifications then were articulated in a formal document that provided a conceptual structure for statistics education by providing learning objectives for solving statistical problems through three developmental levels. This document is known as the Pre-K-12 Guidelines for Assessment and Instruction in Statistics Education (GAISE) Framework (Franklin et al., 2007). A discussion about this document is presented below.

Pre-K-12 Guidelines for Assessments and Instructions in Statistics Education (GAISE) Framework

The Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A curriculum Framework known as the Pre-K-12 GAISE Framework, has been influential in focusing attention on statistics and data analysis components in mathematics curriculum in the United States. This document addresses student learning objectives in statistics and data analysis and is asserted by its authors to be consistent with other curriculum such as the National Council of Teachers of mathematics (NCTM) standards (NCTM, 2000), the Mathematics Education for Teachers (MET) Report (Conference Board of the Mathematical Sciences, 2001), and the K-12 Common Core State Standards in Mathematics (National Governors Association Center for Best Practices, and Council of Chief State School Officers, 2010). The main objective of the GAISE Report is to provide teachers and teacher educators with a developmental framework for instruction and assessment of statistical concepts (Franklin et al., 2007). This document is intended by its authors to supplement the mathematics curriculum standards, not to replace them.

The Pre-K-12 GAISE Framework comprehensively addresses student learning objectives and gives detail guidance for instruction and assessments in the areas of

Table 1

Pre-K-12 GAISE Framework (*Franklin et al., 2007)*

Process Component	Level A	Level B	Level C
Formulate Question	Beginning awareness of the *statistics question distinction* Teachers pose questions of interest Questions restricted to classroom	Increase awareness of the *statistics question distinction* Students begin to pose their own questions of interest Question not restricted to classroom	Students can make the *statistics question distinction* Students pose their own questions of interest Questions seek generalization
Collect Data	Do not yet design for differences Census of classroom Simple experiment	Beginning awareness of design for differences Sample surveys; begin to use random selection Comparative experiment; begin to use random allocation	Students make design for differences Sampling designs with random selection Experimental designs with randomization
Analyze Data	Use particular properties of distributions in the context of a specific example Display variability within a group Compare individual to individual Compare individual to group Beginning awareness of group to group Observe association between two variables	Learn to use particular properties of distributions as tools of analysis Quantify variability within a group Compare group to group in displays Acknowledge sampling error Some quantification of association; simple models for association	Understand and use distributions in analysis as a global concept Measure variability within a group; measure variability between groups Compare group to group using displays and measures of variability Describe and quantify sampling error Quantification of association; fitting of models for association

Table 1 continued

Process Component	Level A	Level B	Level C
Interpret Results	Students do not look beyond the data No generalization beyond the classroom Note difference between two individuals with different conditions Observe association in displays	Students acknowledge that looking beyond the data is feasible Acknowledge that a sample may or may not be representative of the larger population Note the difference between two groups with different conditions Aware of distinction between observational study and experiment Note differences in strength of association Basic interpretation of models for association Aware of the distinction between association and cause and effect	Students are able to look beyond the data in some contexts Generalize from sample to population Aware of the effect of randomization on the results of experiments Understand the difference between observational studies and experiments Interpret measures of strength of association Interpret models of association Distinguish between conclusions from association studies and experiments
Nature of Variability	Measurement variability Natural variability Induced variability	Sampling variability	Chance variability
Focus on variability	Variability within a group	Variability within a group and variability between groups Covariability	Variability in model fitting

Note. The Pre-K-12 GAISE Framework. Reprinted from "*Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A Pre-K-12 Curriculum Framework*," by C. Franklin, G. Kader, D. Mewborn, J. Moreno, R. Peck, M. Perry, & R. Scheaffer, 2007, Copyright 2005 by the Joint American Statistics Association/ National Council of Teachers of Mathematics. Adapted with permission.

statistics and probability. Franklin et al. (2007) claimed that the Pre-K-12 GAISE Framework "provides a conceptual structure for statistics education that gives a coherent picture of the overall curriculum" (p. 5). This document is intended to complement the NCTM's *Principles and Standards* (2000).

The framework displayed in Table 1 introduces the statistical problem solving process and then shows how this process can be presented at each of three developmental levels (Levels A, B, and C). Although the basic structure of the process is the same at each level, the degree of sophistication in the types of problems addressed over the developmental levels increases (Peck, Kader, & Franklin, 2008).

The framework presents a two-dimensional model of a conceptual structure for learning statistics. The first dimension consists of components of the statistical problem-solving process that involves four components: question formulation, data collection design and implementation, data analysis, and interpretation (Franklin et al., 2007). The second dimension includes the three developmental levels of statistical education; Level A, B, and C. The framework also provides an additional emphasis on understanding the role of variability in the problem solving process: (1) anticipating variability in question formulation process by making the statistics question distinction, (2) acknowledging variability in data collection process by designing for difference, (3) accounting of variability in data analysis process by using distributions and (4) allowing for variability in interpretation process by looking beyond the data (Franklin, et al., 2007).

Acknowledging and understanding the role of variability in the statistical problem solving process requires maturation in statistical literacy, reasoning, and thinking.

Franklin et al. (2007) describe this maturation over the three developmental levels (A, B, and C) as follows:

> Although these three levels may parallel grade levels, they are based on development in statistical literacy, not age. Thus, a middle school student who has had no prior experience with statistics will need to begin with Level A concepts and activities before moving to Level B. This holds true for a secondary student as well. If a student hasn't had Level A and B experiences prior to high school, then it is not appropriate for that student to jump into Level C expectation. The learning is more teacher-driven in Level A, but becomes student driven at Levels B and C. (p. 13)

This suggests that students in the same mathematics classroom could very well have different levels of statistical literacy and thinking. Of course, this heterogeneity of students' knowledge and skill levels is not applicable only for statistics. A similar situation happens in most classrooms, regardless of the subjects being taught. Therefore, as with other subjects, it is necessary that a teacher identifies the developmental level of each student in his/her class before teaching statistics lessons. By identifying students' developmental level, the teacher can develop appropriate instructional approaches at the identified level for each student or group of students.

Peck, Kader, and Franklin (2008) illustrated the problem solving process in each level that is also supported by the learning trajectories suggested by previous studies as below. At Level A, questions are formulated limited to subjects in the classroom and data are collected by taking a census of the classroom. Data are analyzed using simple picture graphs, tallies, frequency tables and bar graphs, or dot plots (line plots). The mode is

introduced as the category having the highest frequency in a categorical data. Mean is developed as the fair share value and median is introduced as the center for numerical data. The range is introduced as a basic measure of variation. Interpretation is focused on comparing individual-to-individual variability and individual-to-group variability within the context of the question posed. Generalizations beyond the classroom are not expected.

In Level B, the questions posed are not just restricted within the classroom, but usually broaden beyond the classroom. Students are introduced to the concept of random selection. Level B students are those who are able to use multiplicative and proportional reasoning that enable them to summarize numerical data into pictographs, circle graphs (pie charts), relative frequency tables and bar graphs. Mean is interpreted as the balance point of the data distribution, and variation in data is measured using the mean absolute deviation (MAD) as a transitional measure to the standard deviation developed later at Level C. Interpretation of the data is focused on comparing both within group variability and between group variability. Students also compare the variability within group and between groups using relative frequency tables and bar graphs, conditional percentages, and segmented bar graphs for categorical data, and using dot plots and box plots for numerical data. Students are introduced to the inter quartile range (IQR) as a measure of variation using box plots. At level B, students begin to understand that the ability to generalize conclusions depends on how the data are collected.

In Level C the questions posed now require generalization from a sample to a larger group. Data are collected using random selection. While Levels A and B focus on interpreting variability through descriptive statistics, Level C students begin to think about sampling variability, the role of probability in statistical problem solving, and their

impact on conclusions and generalizations using simulation. Table 1.1 shows a summary of the Pre-K-12 GAISE framework with guidelines of expected knowledge or skills that students should accomplish for each process component and level.

Since the Pre-K-12 GAISE Framework focuses only on guiding the instruction and assessment for statistics, the implementation of this framework cannot be detached from the school curriculum that includes other subjects that are considered important for students to learn. Fortunately, a new national curriculum standards document that has been adopted by 45 states in the U.S. shows the same spirit of realizing the importance of statistics as one subject that should be mastered by students in order to face the challenges in the real world today. In particular, the standards for mathematics in this document include a large proportion of statistical contents commence at Grade 6. The document, known as the K–12 Common Core State Standards in Mathematics (National Governors Association Center for Best Practices, and Council of Chief State School Officers, 2010) also provides the learning trajectories of statistical concepts that students are expected to master. A thorough discussion of the CCSS-M is presented below.

The Common Core State Standards in Mathematics

In the K-12 Common Core State Standards in Mathematics (CCSS-M), the content standards for statistics and probability start at Grade 6. The K–12 CCSS-M has included the content standards for Measurement and Data since Kindergarten. In the standards for Measurement and Data, students are expected to be able to collect, handle, and analyze data - a critical ability in statistics. At Grade 6, in the CCSS-M students are expected to develop understanding of statistical variability and summarize and describe distributions. At Grade 7, students are expected to use random sampling to draw

inferences about a population and draw informal comparison of two populations. At Grade 8, students are expected to investigate patterns of association in bivariate data. A detailed description of expectations of Grade 6, 7, and 8 can be seen in the learning trajectory display of the common core state standards for statistics (Confrey, Maloney, & Nguyen, 2010) in Appendix A. In high school, students are expected to be able to interpret categorical and quantitative data, make inferences and justify conclusions, understand conditional probability and the rules of probability, and use probability to make decisions.

Groth and Bargagliotti (2012) suggest that the learning expectation of the CCSS-M for statistics naturally fall under the Pre-K-12 GAISE Framework, thus allowing practitioners to use the Pre-K-12 GAISE Framework as a roadmap to help implement the CCSS-M. In addition, Pre-K-12 GAISE Framework is a compelling supplement to the CCSS-M because it offers the following ideas that are not contained in the CCSS-M: (1) pedagogical approaches for statistics; (2) meaningful statistical connection; (3) developmental trajectories for students' statistical learning; and (4) enhancement of the curriculum prescribed by the CCSS-M (Groth & Bargagliotti, 2012).

Summary

The concept of *learning trajectories* in mathematics is a concept that provides an approach to develop the knowledge needed to define the pathways that students go through in learning mathematical concepts (Daro et al., 2011). The hypotheses of learning trajectories are rooted in actual empirical studies of the ways in which students' thinking grows in response to relatively well specified instructional experiences. Specific

to statistics, several studies on how students develop their understanding of several big ideas have been conducted.

Investigating how students develop their understanding for collecting data and sampling, Watson and Moritz (2000a) found that there is a progression in students' thinking about samples in which students first (a) do not distinguish between sample and population, then (b) recognize the difference between sample and population but really wanted to sample anyone, and finally (c) realize that samples can be used to represent the population and to estimate population parameters. For understanding graph representations, Shaughnessy (2007) summarized the learning trajectory as (1) *Reading Data*; (2) *Reading within the data*, (3) *Reading within the data*, (4) Reading *beyond the data*, and (5) Reading *beyond the data*. Shaughnessy (2007) added one more level, called reading *behind* the data.

There are not many studies investigating how students develop their understanding about association, co-variation, and correlation, especially at the middle and high school levels. Among the few studies, Batanero, Estepa, Godino and Green (1996) were able to identify four misconceptions of statistical association: (1) deterministic conception of association, (2) unidirectional conception of association, (3) local conception of association, and (4) causal conception of association. In investigating how students develop their understanding about variability, it is found that, using Reading & Shaughnessy's (2004) terms, the lowest level in understanding variability is called uni-structural level, the middle level is called multi-structural level, and the highest level is called relational level.

The identifications on how students develop their understanding of several big ideas in statistics and the learning trajectories of the ideas were articulated in the Pre-K-12 Guidelines for Assessment and Instruction in Statistics Education (GAISE) Framework (Franklin et al., 2007). The framework presents a two-dimensional model of a conceptual structure for learning statistics. The first dimension consists of components of the statistical problem-solving process that involves four components: question formulation, data collection design and implementation, data analysis, and interpretation (Franklin et al., 2007). The second dimension includes the three developmental levels of statistical education; Level A, B, and C. The framework also provides an additional emphasis on understanding the role of variability in the problem solving process.

The Pre-K-12 GAISE Framework's spirit on realizing the importance of statistics as one subject that should be mastered by students is also showed by a new national curriculum standards document, known as the K–12 Common Core State Standards in Mathematics (National Governors Association Center for Best Practices, and Council of Chief State School Officers, 2010). The CCSS-M also provides the learning trajectories of statistical concepts that students are expected to master. In the K-12 CCSS-M, the content standards for statistics and probability start at Grade 6.

Groth and Bargagliotti (2012) suggest that the learning expectation of the CCSS-M for statistics naturally fall under the Pre-K-12 GAISE Framework, thus allowing practitioners to use the Pre-K-12 GAISE Framework as a roadmap to help implement the CCSS-M. In addition, Pre-K-12 GAISE Framework is a compelling supplement to the CCSS-M because it offers some ideas that are not contained in the CCSS-M.

CHAPTER III

METHODOLOGY

In this chapter, the methodological approach applied in this study is presented. The primary goal of the methodology is to answer the research questions investigated in this study:

1. How and how much do middle and high school students understand statistical concepts that are related to the investigation cycle (formulating questions, collecting data, analyzing data, and interpreting result)?
2. What are the learning trajectories that describe the developmental progression for different concepts and statistical investigation processes?
3. To what extent do students' understandings of statistical concepts develop similarly across developmental levels?
4. Given the structure of the progressions observed in performance of different levels, to what extent can students' developmental level be diagnosed reliably and validly?

To answer these questions, an instrument to identify developmental levels and trajectories in learning statistics for several statistical ideas has been developed. Some items of the instrument were developed by creating original items that began in September 2011.

Others were adopted or adapted from instruments created in earlier studies on assessing students' understanding of statistics and probability. To ensure the validity of the items, an expert survey was conducted in November 2011 that required two statistician/statistics educators to validate the contents assessed by the items and the alignment of instrument items with the Pre-K-12 GAISE Level. After revising the instrument based on experts' opinions, the instrument was piloted with 19 middle school and high school students participating in Junior Summer Math Camp organized by Texas Mathworks in summer 2012. Another pilot administration of the instrument was conducted in summer 2012 that involved 66 undergraduate students who were expected to have the same level of statistical understanding as high school students. The instrument was then administered to 797 students in Grades 6 to 12 in fall 2012. An expert panel convened again in spring 2013 with one statistician and one statistics educator serving as the experts to validate the content and the alignment of items with the GAISE levels. Students' responses were collected and several psychometric and statistical analyses were conducted to answer the research questions.

The first research question was answered by evaluating the results of a descriptive statistical analysis of student response data. The difficulty levels of all items under investigation were compared. The second research question was answered by conducting a structural equation modeling (SEM)-based confirmatory factor analysis (CFA) and classical test theory (CTT) - based analysis using Mplus Version 7 ((Muthen & Muthen, 2012) and SPSS AMOS 21 (IBM, 2012) programs. The hypothesis that students develop their understanding of statistical problem solving process components through three developmental levels suggested by the Pre-K-12 GAISE Framework was tested. The third

research question was answered by comparing students' performances in each process component across GAISE levels. Finally, the fourth research question was answered by developing a model to assign a score for each developmental level. Students were assigned their Level A, Level B, and Level C scores based on their responses to all items in the instrument. A criterion was defined to diagnose students' developmental level in learning statistics based on their scores in all levels. The detailed descriptions of these methodologies are presented in the following sections.

Instrument Development

There are two instruments that have been developed in this study: the expert survey instrument and the student survey instrument. The expert survey asked experts to align the students' survey items with the Pre-K-12 GAISE Framework and determine whether the items were appropriate for measuring several statistical concepts. Several items were developed to measure students' developmental levels and learning trajectory in statistics. The items were adapted from the Statistical Reasoning Assessment (Garfield, 1991, 2003), the Statistical Literacy Assessment (Callingham & Watson, 2005; Watson, 1997; Watson & Callingham, 2003), the assessment item database of the Assessment Research Tool for Improving Statistical Thinking (ARTIST), and released items of the Program for International Student Assessment (PISA) (OECD, 2009). New items were developed based on Pre-K-12 GAISE Framework and Common Core State Standards guidelines.

The first expert survey was conducted during fall 2011 that involved two statistician/statistics educators. In the first survey, 40 items in multiple choice formats were included. The experts were asked to judge the alignment of the items with statistics

problem solving process components and GAISE levels suggested by Pre-K-12 GAISE Framework. They were also asked to give their opinion on the clarity of the items. Two experts provided feedback and based on this feedback, the items were revised and classified into GAISE Levels. Some items were considered inappropriate due to wording or content. Some items were revised and other items were replaced. Based on the experts' opinion, the items were classified to measure the GAISE levels (Level A, B, and C).

The forty items were divided into three forms to ensure that participants only needed less than one hour to respond to all items in each Form. ITEMS 01-13 were in Form 1, ITEMS 14-26 in Form 2, and ITEMS 27-40 in Form 3. The items were piloted by administering them to middle and high school students participating in Junior Summer Math Camp that was organized by Texas Mathworks during summer 2012. The pilot survey was also conducted with undergraduate students that same semester. Students participating in the pilot study were asked to respond to the items and give their comments about the items. Several students mentioned that they did not understand some sophisticated words such as "simultaneously" and "standard deviation." Some words were revised to be consistent with middle school students' reading levels. Some words, however, were kept in their original forms, especially for Level C Items. Students who have developed into Level C are expected to understand several concepts such as mean, median, and standard deviation. Pilot study data were analyzed using CTT-based analyses. Four items were excluded from the instrument due to their high difficulty indices or their low point-biserial indices. Items that were too easy (difficulty level < .2) or too difficult (difficulty level > .8) were discarded. Items with point-biserial less than .2

were also discarded because the items could not discriminate students based on their ability as expected.

Before the pilot study was conducted, several possible threats were identified. The first threat was the possibility of an absent or unclear conceptual match between the instrument and the intended results. This threat can be avoided by choosing expert panelists who agree to align items with the Pre-K-12 GAISE Framework and to suggest some revision to the items due to unclear wording or content. The threat can also be anticipated by providing an adequate number of items to be analyzed, in this case 40, which made it appropriate for experts to assist; considering time consumed in reviewing items were reasonable. The threat had actually been resolved by optimizing the number of items that were large enough for a valid CTT-based item analysis and SEM-based confirmatory factor analysis but also small enough to guarantee that the time needed by the experts to validate the contents of the items were not too long.

The second threat to the validity of the study was the possibility that bias might result from survey administration where certain groups of students only took one form of the survey. For instance, if Grade 8 students in Algebra classes only took Form 1 and Grade 8 students in Geometry class only took Form 2, then the results would not represent the sensitivity of the instrument to identify Grade 8 students' developmental level in learning statistics. A plan to avoid this threat included randomly assigning both forms to the participants. Due to a technical fallacy in printing the survey forms, however, this plan did not execute well. The proportion of Form 1 and Form 2 that were administered was not balanced. At the end of the survey, only 140 students took Form 1 and 657 students took Form 2.

The third possible threat to the validity was the possibility that students did not put their full efforts into answering the survey. This threat was resolved by requesting full supports from the teachers to administer the survey to their students. Under their own teachers' supervision, students tended to show genuine effort to answer the items, even though they were informed that the survey was not related to the courses they took at the time of administration of the instrument.

The fourth threat was the possibility that we could not get a large enough sample size, particularly for the purposes of item analysis which needs a large sample size. This threat was anticipated by contacting several schools that had similar characteristics. The threat, however, did not become a reality as the principal of one middle school and two teachers of one high school decided that they were willing to allow their students to participate in this study. As mentioned before, we finally got a large sample of 797 students.

The opinions of experts and students' comments on the items during pilot study were used as *face validity* evidence of the instrument. Only a few items were not considered obvious for the participants and those items were then revised. Opinions of experts during pilot study were also used as *content validity* evidence of the instrument. Students' responses in pilot study were analyzed using classical test theory (CTT) approach. The difficulty index of each item that measures how difficult the items were and the point-biserial of each item that measure the correlation of students' response to the item with their total scores were computed. These indices are important in CTT as criteria to judge the quality of items.

Among the 40 items developed in the pilot study, four items were considered inappropriate to be included in the instrument because of difficulty indices that were too low or too high, or because their point-biserial was too low. After revision, the 36 appropriate items were then distributed into two forms. ITEMS 01–18 were in Form 1 and ITEMS 19-36 were in Form 2. Items were distributed into forms based on their intended levels suggested by the experts in the pilot study. Each form was designed to include a similar proportion of items from each level.

Six big ideas in probability and statistics are included in the instrument. The six ideas include the concepts of (1) awareness of statistical question distinction, (2) sampling and data collection method, (3) measure of centers (averages), (4) variability, (5) graph representation and interpretations, and (6) association, covariation, and correlation. The last five ideas have been studied in earlier research on students' learning of statistics as discussed in the previous chapter. The concept of awareness of statistical question distinction is a new concept that is suggested by the Pre-K-12 GAISE Framework that has not been discussed in the literature. In this study, two items that assess students' understanding of this idea were included in the instrument.

The second expert survey was conducted in spring 2013. The experts were, again, asked to align the 36 items administered to the students. Experts' opinions in the second expert survey were then used to develop the structural equation models that later were tested for construct validity and scaling purposes.

The six contents are organized by the four process components in statistical investigation suggested by Pre-K-12 GAISE Framework: formulating question, collecting data, analyzing data, and interpreting result. Some of the items also assess students'

understanding of the nature of variability and the role of variability. With the increased attention to the development of students' statistical literacy, reasoning, and thinking at all levels, the six contents were also projected to measure statistical literacy, reasoning and thinking.

Statistical literacy, reasoning, and thinking are terms to describe certain cognitive skills that are expected to be developed in learning statistics. Ben-Zvi and Garfield (2004) define the terms as follows: Statistical *Literacy* is the ability to understand and critically evaluate statistical information. Two components of statistical literacy required by society are: (1) the ability to interpret and critically evaluate statistical information and (2) the ability to discuss or communicate reactions to such statistical information (Gal, 2002). For example, students who have developed their statistical literacy are able to understand statistical information reported by the media and critically analyze the report. *Statistical Reasoning* could be defined as the way people reason with statistical ideas and make sense of statistical information (Garfield & Gal, 1999). Statistical reasoning involves interpreting and deducing based on sets of data, graphical representations, and statistical summaries. For instance, students who have developed their statistical reasoning should be able to compare and make inferences regarding the comparison of two groups, based on the data of the two groups represented by box plots. *Statistical Thinking* includes "an understanding of why and how statistical investigations are conducted and the 'big ideas' that underlie statistical investigations" (Ben-Zvi & Garfield, 2004, p. 7). Statistical thinkers understand the omnipresent nature of variation and know how to use appropriate methods of data analysis (e.g. numerical summaries and visual displays of data) and when to use them. For example, students who have developed

their statistical thinking should be able to determine which graphs should be used to represent data in order to answer the research question. Based on this cognitive skill hierarchy and the statistical investigation process components, the items were developed by following the blue print presented in Table 2. Due to their different theoretical frameworks that are outside the scope of this study, analyses on how students perform related to their statistical literacy, reasoning, and thinking were not conducted. In the process of developing the instrument, the focus of assuring validity and reliability of the measurement is necessary. In the following section, a discussion of efforts to assure the validity and reliability of the measurement is presented.

Table 2

Instrument Blue Print

Process Component	Statistical Literacy	Statistical Reasoning	Statistical Thinking
Formulating Questions	Item # 5, 8		
Collecting Data	Item # 3, 27, 29	Item # 3,13, 25, 27, 29, 31	Item # 3, 25, 29, 31
Analyzing Data	Item # 1, 4, 9, 11, 16, 19, 21, 24, 26, 30, 32, 33	Item #, 11, 16, 24	Item # 9, 11, 24,
Interpret Results	Item # 15, 17, 18, 34, 35, 36	Item # 6,7, 14, 15, 17, 18, 34, 35, 36	Item # 17, 34, 35, 36
Nature of Variability	Item # 2, 20, 22, 23, 28	Item # 2, 10, 12, 20, 22, 28	Item # 2, 12, 28,

As can be seen in the table, this instrument has only two items that assess the Formulating Questions process component. This lack of items that assess students' understanding on formulating question is due to the fact that in Levels B and C, students are expected to increase their awareness of the distinctions of statistics questions and then

be able to formulate their own questions based on available data or based on problems that they want to investigate. These type of skills are difficult, if not impossible, to be assessed by multiple choice items.

About one third of the items in the instrument are items that measure the analyzing data process component. This is due to the fact that there are more aspects of analyzing data that should be assessed compared to other process components. For instance, in the analyzing data process component, students should know how and be able to use measures of center and measures of variability, understand graph and interpret information presented on graphs, organize data, etc. On the other hand, in the collecting data process component, students are expected to understand sampling methods and the importance of randomization which do not involve as many statistics contents as the analyze data process component. This instrument does not have any item to assess Focus of Variability. Among the 36 items, 11 items (Item # 2, 4, 6, 7, 10, 14, 20, 21, 22, 23, and 33) address probability problems and the others address statistics problems.

Validity and Reliability

Validity is possibly the most important criterion for the quality of an instrument. The term *validity* denotes whether an instrument measures what it is claimed to measure. The validity of an instrument specifies how well the instrument meets the standards by which it is judged. An instrument that exhibits poor validity consists of items that do not measure what the instrument purports to measure. In this study, several ways to estimate the validity of the measurement were conducted, including face validity, content validity, and construct validity (Bornstein, 2003; Vogt, 2007). Descriptions of these types of validity are presented below.

Face validity

Bornstein (2003) describes face validity as "an estimate of the degree to which a measure is clearly and unambiguously tapping the construct it purports to assess." In other words, face validity refers to the "obviousness" of a test—the degree to which the purpose of the test is apparent to those taking it. Tests are said to have high face validity if the purpose is clear, even to naïve respondents (Nevo, 1985). The concept of face validity is similar to item subtlety, but there are important differences as well. Whereas face validity describes the transparency of an entire test, item subtlety describes the transparency of individual test items (Bornstein, Rossner, Hill, & Stepanian, 1994). Examples of face validity evidence could include comments from people who read the instrument items and whether the items are clear and understandable to them.

Content validity

Content validity primarily rests upon an appeal to the propriety of content and the way that it is presented (Nunnally & Bernstein, 1994). In this study, content validity is a logical process where connections between items and contents in probability and statistics for middle school and high school curriculum are established. This was accomplished by aligning the context with the Common Core State Standards and by experts' review. These experts were given the list of content areas specified in the test blueprint and also the Pre-K-12 GAISE Framework, along with the test items intended to be based on each content area. The experts were then asked to indicate whether or not they agreed that each item was appropriately matched to the content area indicated. The experts were also asked to indicate the appropriate level suggested by Pre-K-12 GAISE Framework for each item. Any items that the experts identified as being inadequately

matched to the contents listed in the blueprint, or flawed in any other way, were either revised or dropped from the instrument.

Construct validity

Construct validity was measured by applying confirmatory factor analysis (CFA) to verify whether the instrument actually measured three constructs, GAISE Level A, B, and C and whether the constructs displayed associations among each other as suggested by the Pre-K-12 GAISE Framework. A structural equation modeling (SEM) approach for CFA was conducted using SPSS AMOS 21 (IBM, 2012) and Mplus version 7.0 (Muthen & Muthen, 2012) programs.

Test reliability

Test reliability is a central concept in educational measurement and CTT. Reliability refers to the consistency of the scored acquired from a test or other instrument. Basically, the goal of reliability estimation is to measure the extent to which an experiment, test, or any measuring procedure yields the same results on repeated trials (Carmines & Zeller, 1979). Reliability tells us whether a test is likely to yield similar results if administered to the same group of test-takers multiple times. In other words, the test items should behave the same way with different populations of test-takers, by which it is generally meant that the items should have approximately the same ranking when sorted by their item difficulty indices. Reliability can be estimated in a number of different ways such as: (1) Test-Retest reliability, where the same test is administered twice for the same participants after certain periods of time; (2) Inter-Rater reliability, where two or more independent judges score the test and the scores are then compared to determine the consistency of the raters' estimates; (3) Parallel-Forms reliability, where

two different tests are created using the same content and then are administered to the same subjects at the same time; and (4) Internal Consistency Reliability that is used to judge the consistency of results across items on the same test. Considering the limitation of this study that only involves one survey administration, the reliability of the survey instrument in this study will be estimated using internal consistency reliability.

Two common measures of internal consistency reliability under CTT are the split-halves and Cronbach's α reliability coefficient (Shu, CTB/McGraw-Hill, & Schwarz, 2010). The split-halves coefficient assumes that the two forms conform to the classically parallel model (i.e., their test scores have the same mean and variance). Shu and Schwarz (2010) explained that the most restrictive definition of reliability is represented by classically parallel measurement which specifies that common skills are measured and equal means, true score variance, observed score variance and error variance exist.

The mathematical expression of the classical test theory definition of internal consistency reliability is known as the Kuder-Richardson Formula 20 or KR20 (Kuder & Richardson, 1937). This definition expresses test reliability as the ratio of true score variance (that is not known) to observed score variance (test performance); it is generally expressed symbolically as the following:

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}$$

where the reliability, $\rho_{XX'}$ of test *X* is the ratio between the true score variance, σ_T^2 , and observed score variance, σ_X^2. Observed score variance is defined as the combination of true score variance and error variance, σ_E^2. As error variance is reduced, reliability increases (that is, students' observed scores are more reflective of students' true scores or

actual proficiencies. The estimation of this reliability can be mathematically represented by

$$KR20 = \left[\frac{k}{k-1}\right]\left[\frac{\sigma_X^2 - \sum_{i=1}^{k} p_i\,(1 - p_i)}{\sigma_X^2}\right],$$

where KR20 is a lower-bound estimate of the true reliability, k is the number of items in test *X*, σ_X^2 is the observed score variance of test *X*, and p_i is the proportion of students who got item *i* correct (in psychometric p_i is also known the p-value of item *i*). This formula is used when test items are scored dichotomously.

Coefficient alpha, also known as Crohnbach's alpha (Crohnbach, 1951), is an extension of KR20 to case where items are scored polytomously and are computed as follows:

$$\alpha = \left[\frac{k}{k-1}\right]\left[1 - \frac{\sum_{i=1}^{k} \sigma_i^2}{\sigma_X^2}\right],$$

where α is a lower-bound estimate of the true reliability, *k* is the number of items in test *X*, σ_X^2 is the observed score variance of test *X*, and σ_i^2 is the sample variance of scores for item *i*. If the data is dichotomous, the value of α is exactly the same as KR20, because by definition for dichotomous item, the sample standard deviation of item *i* is defined as

$$\sigma_i^2 = p_i\,(1 - p_i).$$

For the rest of this chapter, the coefficient α is used to represent the internal consistency reliability index of each survey form.

In this study, the reliability KR-20 or Crohnbach's α reliability coefficient (Crohnbach, 1951) was applied to estimate the internal consistency reliability of

participants' responses to the instrument items. Since there are three different constructs (Level A, Level B, and Level C) involved in the instrument, high internal consistency reliability is not expected. Measuring reliability for each GAISE level might give a better reliability coefficient; however, since there is only a small number of items per level, a high reliability coefficient of each level is less likely to exist.

Population and Samples

The target population for this study is students in Central Texas. This population was chosen for efficiency reasons considering the result of this study is not expected to be generalized to a larger population. Furthermore, as a preliminary effort to validate and measure the reliability of the developed survey instrument, a convenient sample is considered appropriate. One of the reasons is that the schools' locations are close to the base of this study, Texas State, which enabled the researcher to save the cost of transportation and accommodation. According to Texas School Directory (Texas Education Agency, 2011), there are 1654 middle school students in San Marcos CISD that are distributed into two schools: Doris Miller (724) and Goodnight (930) Middle School. The number of student in San Marcos High School is 2111. During the two months of administration period, 649 students from Doris Miller Middle School and 148 students from San Marcos High School participated in this study.

Data Analysis

By focusing the data analysis to answer the research questions, it is necessary that we review again the research questions investigated this study:

1. How and how much do middle School and high School students understand statistical concepts that are related to the investigation cycle (formulating questions, collecting data, analyzing data, and interpreting result)?
2. What are the learning trajectories that describe the developmental progression for different concepts and statistical investigation processes?
3. To what extent do students' understandings of statistical concepts develop similarly across developmental levels?
4. Given the structure of the progressions observed in performance of different levels, to what extent can students' developmental level be diagnosed reliably and validly?

In order to answer the first question, descriptive statistics of students' difficulty indices across all process components were analyzed. Data were represented in boxplots and comparisons among *process components* were discussed. The boxplots of difficulty indices *for all levels* were also compared and displayed. These analyses were conducted to answer the second and the third research question, i.e. to identify the learning trajectories of the statistical investigation process and also the learning trajectory of several statistical concepts involved in the instrument and to determine to what extent students' understandings of statistical concepts develop similarly across developmental levels. The results described whether the developmental levels suggested by the Pre-K-12 GAISE Framework remain true for all statistical investigation process components.

From the literature and experts' review, four models that describe the alignment between items and levels and the relation among levels were developed. For example, an initial model hypothesized that ITEM 02, 07, 11, and 14 assessed Level A mastery;

ITEM 03, 05, 06, 08, 09, 10, and 12 assessed Level B mastery; and ITEM 04, 13, 15, 16, 17, and 18 assessed Level C mastery. The models also delineated that Level C mastery impacted Level B mastery and Level B mastery impacted Level A mastery, as suggested by the Pre-K- 12 GAISE Framework.

A confirmatory factor analysis using structural equation modeling (SEM) approach was conducted where model fit testing was applied to examine to what extent the four developed models fit the empirical data. Several criteria of the goodness of fit indices of the models were compared. The Mplus Version 7 and SPSS AMOS 21 programs were used in this analysis. The results are used to support or reject the hypotheses that the developmental progression of statistical investigation process components is explained by the models. The results are also used to describe learning trajectories of several statistical concepts included in the instrument. The best model for each form was chosen and used as the guideline for aligning items with their appropriate levels. Descriptive statistics of students' performances in each level were then analyzed.

To answer the fourth research question, several psychometric and statistical analyses were conducted. First, classical test theory (CTT) analysis was conducted in order to examine the quality of items in the instrument. The quality of items determined one aspect of the validity of the measurement. The CTT was also used to analyze the internal consistency reliability of the measurement. All analyses conducted in this study had a primary goal: to show to what extent the instrument developed in this study validly measured what it was supposed to measure, in this case, the developmental level of learning statistics. Therefore, the SEM analysis results conducted previously, is also useful as construct validity evidence by showing that the instrument indeed measures

three different factors, in this case GAISE Level A, B, and C. The analysis showed to what extent this hypothesis is supported by the data. In the following discussion, a brief description of CTT and SEM analyses is provided.

Classical test theory (CTT)-based item analysis

There are three major types of item analysis: *classical test theory* (CTT), *item response theory* (IRT), and *Rasch measurement.* Among these three types of item analysis, only CTT does not assume unidimensionality of the data. Since, theoretically, our data has three constructs or dimensions, only CTT item analysis is appropriate for this study. CTT has been the foundation of measurement theory for more than eight decades. The foundation for all types of CTT rests on aspects of a total test score an examinee gets from responding to multiple items. Most classical approaches assume that the raw of an individual score (X) is obtained by adding the true score (T) of the individual and a random error (E) as shown below:

$$X = T + E.$$

The true score of a person (T) is a hypothetical score that is found by taking the mean score a person would get on the same test if s/he assumedly had an infinite number of testing sessions.

Embretson and Reise (2000) reviewed several rules of CTT. The first rule was that the standard error of measurement was generated by the large numbers of individuals who took the test and was applied to all scores in a particular population and did not differ from individual test taker. The second rule of CTT was that longer tests are more reliable than shorter tests. Therefore, in CTT, a larger number of items better sample the universe of items and resulting statistics generated by them (such as mean test scores) are

more stable if they are based on more items. Another ramification of CTT is that the important statistics about test items such as their difficulty depend on the sample of respondents being representative of the population.

CTT concentrates on two main statistics: item facility and item discrimination. Item facility is calculated as below:

$$Fac\ (X) = \frac{\bar{X}}{X_{max}}$$

where *Fac* (*X*) = the facility value of question *X*.

$\bar{X}$ = the mean score obtained by all examinees attempting item X.

X_{max} = the maximum score on the item (McAlpine, 2002).

It is clear that for dichotomous responses, item facility of question X represents the percentage of correct responses to the question. On items that have a high proportion of correct answers, it is desirable for the facility value to be close to 0.5, to promote maximal differentiation. It is clear that for dichotomous type of responses, item facility of an item is determined by percentage correct responses of the item.

There are several pieces of information that can be used to determine if an item is useful and/or it performs in relation to the other items on the test. The mean and standard deviations of items will inform which items will be useful and which will not. For example, if the variance of an item is low, this means that the item might not be useful since there is a little variability in the item responses. If the mean response to a dichotomous item is 0.90, then the item is negatively skewed and may not provide the kind of information needed. Generally, the higher the variability of the item and the more the mean of the item is at the center point of distribution, the better the item will perform.

The mean of a dichotomous item is equal to the proportion of individuals who respond to the item correctly (denoted *p*). The variance of a dichotomous item is calculated by multiplying $p \times q$ (where q is the proportion of individuals who respond wrongly to the item). The standard deviation is the square root of $p \times q$. So, for example, if 1000 individuals respond to an item and 400 respond correctly, then the *p* value for that item is 400/1000, or 0.40. The q is 0.60 (1.0 – 0.40 = 0.60). The variance of the item is 0.24 (0.40 × 0.60 = 0.24) and the standard deviation is the square root of 0.24, or 0.49. Discrimination index of an item can be assessed using several methods. For dichotomous items, the Pearson point-biserial or Pearson biserial correlation coefficients are available in SPSS. For both statistics, the relationships between how individuals responded to each item are correlated with the *corrected* total score on the test. The corrected total score is the total score excluding the response to the item in question. Since the total score including the response to the item is highly correlated to the item in question, then this correction is an appropriate correction.

As mentioned above, another statistic that is important in CTT is the point-biserial correlation of items. The formula for the point-biserial correlation coefficient is

$$r_{pbis} = [\frac{\bar{Y}_1 - \bar{Y}}{\sigma_Y}] \times \sqrt{p_x/q_x}$$

where $\bar{Y}_1$ = the mean of the total test scores for those whose dichotomous response was 1, $\bar{Y}$ = the mean of the total test scores for the whole sample, σ_Y = the standard deviation of all scores on the total test, p_x = the proportion of individuals whose dichotomous response was 1, and q_x = the proportion of individuals whose dichotomous response was 0 (Kline, 2005).

The survey instrument for this study is divided into two forms; Form 1 and Form 2. Among the 797 participants, 140 participants took Form 1 and 657 students took Form 2. Due to too many missing data, ITEM 19 was not included in the data analysis. Due to its zero variance, ITEM 01 was also excluded from the CTT item analysis. CTT item analysis results provide information about the item facility and the point-biserial correlation index of the other 34 items that are presented in the next chapter. The statistics produced by the CTT analysis give the information about the quality of items in the instrument. The results could be considered as evidence of validity in the instrument developed for this study.

Other evidence of validity, especially construct validity, can be provided by showing that participants' responses fulfill the underlying theoretical framework that the instrument measure three constructs, in this case, GAISE Level A, Level B, and Level C. The evidence could be established by conducting factor analysis on the students' responses. There are several choices of factor analysis methods that can be applied that are categorized into two classes: classical factor analysis and structural equation modeling (SEM).

Classical approaches of factor analysis usually have purposes of determining groups and clusters of variables, such as which variables belong to which group and how strongly they belong, how many dimensions are needed to explain the relations among the variables, a frame of reference (coordinate axes) to describe the relations among the variables more conveniently, and scores of individuals on such groupings (Nunnally & Bernstein, 1994). One key assumption underlying classic factor analysis is that the variables are continuous, which is not the case for the measurement used in this study.

Variables involved in this study consist of item responses which are discrete (0 is the response is wrong and 1 if the response is correct).

Classical approaches of factor analysis, however, are incapable of either assessing or correcting for measurement error. On the other hand, SEM is able to provide explicit estimates of the error variance parameter. Classical approaches of factor analysis assume that error(s) in the explanatory variables vanish (es). In general, this assumption will lead to serious inaccuracies. Furthermore, data analysis using SEM procedures can incorporate both unobserved (i.e. latent) and observed variables which are different with previous data analysis methods that are based on observed measurements only. SEM methodology also provides widely and easily applied alternative methods for modeling multivariate relations, or for estimating point and/or interval indirect effect that former methods cannot provide. In this study, a SEM–based confirmatory factor analysis was conducted using SPSS AMOS 21 and Mplus Version 7 programs. A brief description about structural equation modeling is presented below.

Structural equation modeling

Structural equation modeling (SEM) is "a statistical methodology that takes a confirmatory (i.e. hypothesis testing) approach to the analysis of a structural theory bearing on some phenomenon" (Bryne, 2010). Two important aspects of procedures in SEM are: (a) a series of regression equations that represent the "causal" relationships among variables; and (b) these structural relations can be modeled pictorially to enable a clearer conceptualization of the theory under study. SEM is different from the older generation of multivariate procedures in several aspects. First, SEM uses a confirmatory approach in data analysis rather than an exploratory approach. SEM enables researchers

to do hypothesis testing that is difficult to do in older multivariate procedures. Second, SEM provides explicit estimates of error variance parameters whereas traditional multivariate procedures are incapable of either assessing or correcting for measurement error. Alternative methods assume that error(s) in the explanatory variables vanish (es) which can lead to serious inaccuracies. Third, data analysis using SEM procedures can incorporate both unobserved (i.e. latent) and observed variables which is different from former data analysis methods that are based on observed measurements only. Fourth, SEM methodology provides widely and easily applied alternative methods for modeling multivariate relations, or for estimating point and/or interval indirect effect that former methods cannot provide. In this study, data analyses are conducted using SEM approach by applying SPSS AMOS 21 (IBM Corp. Released , 2012) and Mplus Version 7 (Muthen & Muthen, 2012) programs. In the following paragraph, a discussion on estimating SEM parameters is presented.

In reviewing structural equation model parameter estimates, three criteria are of interest: (a) the feasibility of the parameter estimates, (b) the appropriateness of the standard errors, and (c) the statistical significance of the parameter estimates (Bryne, 2010). Each criterion will be discussed next.

Feasibility of parameter estimates

All individual parameters in a structural equation model are expected to fit with the data. In confirmatory factor analysis (CFA) the parameter estimates should be viable, in the sense that the parameter estimates should possess correct sign and size, and be consistent with the underlying theory. Any estimates should fall inside the admissible range. If these criteria are not satisfied, there is an indication that either the model is

wrong or the input matrix lacks sufficient information (Bryne, 2010). For example, it is expected that each correlation < 1.00, each variance is positive, and the covariance or correlation matrices are positive definite. In all models that are developed in this study we found several factor loadings (regression weight estimates) that fell outside the admissible range (.20 – 1.00). Analyses of all models will be presented in the following chapter. All variance are positive, and all covariance and correlation matrices are positive definite. Therefore, in the following part of this section only analysis of regression weight estimates will be discussed.

Appropriateness of standard errors

Bryne (2010) advises that small values of standard errors are expected as they indicate accurate estimations. However, standard errors that are extremely small are not favored since they indicate poor model fit. Cited Bentler (2005), Bryne gave an example that the related parameter of a test statistic cannot be defined if the standard error influenced by the unit approaches zero. Cited Jöreskog & Sörbom (1993), Bryne explained that standard errors that are extremely large indicate parameters that cannot be determined. Because standard errors are influenced by the units of measurement in observed and/or latent variables, as well as the magnitude of the parameter estimate itself, no definitive criterion of "small and large" has been established (Bryne, 2010). All models in this study have calculated standard errors that fall between .050 - .666. Since our parameter estimates fall between -1 and 1, the range of standard errors is adequate.

Statistical significance of parameter estimates

The test statistic to determine the statistical significance of parameter estimates is the critical ratio (C.R.). The C.R is the value of the parameter estimate divided by its

standard error. Therefore the C.R. acts as a z-statistic in testing that the estimate is statistically different from zero. Based on a probability level of .05, then, the C. R. needs to be > ±1.96 to reject the hypothesis that the estimate equals 0.0. Non-significant parameters, except error variances, can be considered unimportant to the model, and hence, they should be deleted from the model (Bryne, 2010). All models in this study showed several estimates have critical ratios that are less than ±1.96. In the next chapter, variables that are considered unimportant to the models are identified and deleted.

Assessment of normality

In general, a critically important assumption in the conduct of SEM analyses is that the data are multivariate normal. When the data are not multivariate normal, interpretations based on the usual Maximum Likelihood (ML) estimation in AMOS may be problematic so that an alternative method of estimation is likely more appropriate (Bryne, 2010). Bryne, however, informed that Chou, Bentler, and Satorra (1991) have argued that it may be more appropriate to correct the test statistics rather than use a different mode of estimation. Bryne (2010) also explained that Satorra and Bentler (1988, 1994) developed a statistic that incorporates a scaling correction for the χ^2 statistic when distributional assumptions are violated; the statistic is called S-Bχ^2. S-Bχ^2 has been shown to be the most reliable test statistic for evaluating mean and covariance structure models under various distributions and sample size (Curran, West, & Finch, 1996).

Unfortunately, this robust method is not available in SPSS AMOS 21, the program that is initially used in this study. Byrne (2006, 2010) compared the usual ML estimation in the AMOS program and the S-B Robust ML estimation in EQS program and found that, although the standard error underwent correction to take nonnormality

into account resulting critical ratios that differed across the AMOS and EQS programs, the final conclusion regarding the statistical significance of the estimated parameters remained the same. Byrne (2010) warned that it should be noted that the uncorrected ML approach tended to overestimate the degree to which the estimates were statistically significant. Byrne (2010) was confident that although she was unable to directly address the issue of nonnormality in the data for technical reasons, and despite the tendency of the uncorrected ML estimator to overestimate the statistical significance of these estimates, overall conclusions were consistent across both approaches. Therefore, we are also confident that overal conclusions of our data analysis results using SPSS AMOS 21 are quite informative and valid.

According to Nessom (2012), there seems to be growing consensus that the best approach to analysis of categorical variables (with few categories) is the continuous/categorical variable methodology (CVM) approach implemented in Mplus. This approach is usually referred to as a robust weighted least squares (WLS) approach in the literature (estimator = WLSMV or WLSM in Mplus). Citing Muthén, du Toit, and Spisic (1997) and Flora & Curran (2004), Nessom (2012) explained that the WLSMV approach seemed to work well if sample size was 200 or better. After conducting SEM analysis using SPSS AMOS 21 program, another analysis was also applied using the Mplus Version 7 program. It found that the results were quite similar with those resulted from SPSS AMOS 21 analysis. Since Mplus analysis is considered as the best approach for categorical data, all SEM analysis results reported in this study are those that are produced by Mplus Version 7 except for Initial F2 Model which failed to reach convergency in Mplus. For this model, the results from SPSS AMOS 21 analysis is

reported. Another advantage of using Mplus is that the program also provides the item characteristic curves (ICC) for all items for each model. The plots of the ICCs can be found in Appendix I below.

Standardized structural coefficients

Standardized structural coefficient estimates are based on standardized data. Standardized estimates can be used in comparing direct effects on a given endogenous variable in a single-group study. Like in ordinary least square regression, the standardized weights are used to compare the relative importance of the independent variables. The interpretation is similar to regression: if a standardized structural coefficient is 2.0, then the latent dependent will increase by 2.0 standard units for each unit increase in the latent independent. In AMOS, the standardized structural coefficients are labeled "standardized regression weights." In comparing models across samples, however, unstandardized coefficients are used (Bryne, 2010).

Goodness of fit criteria was used to determine whether the five models for each Form developed in this study fit the data well. A comparison analysis of goodness of fit statistics of the five models for each Form of instrument is presented in the results chapter.

In order to link the results from CTT item analysis and the exploratory and confirmatory factor analysis, the item characteristic curves (ICC) of items in each level are presented for each SEM models. The ICCs were developed by the Mplus Version 7 program that uses the Item Response Theory (IRT)-based method to create the ICCs. A brief explanation about the IRT and ICC are presented below.

Item response theory (IRT) is a method widely used by testing specialists, especially on standardized tests. When using item response theory, the primary interest is whether examinees get each individual item correct or not, rather than in the raw test score. Since open-response items are difficult to score in a reliable manner, multiple-choice are preferred when using IRT.

As cited by R. K. Hambleton and R. W. Jones (1993), Allen Birnbaum introduced IRT models that are based on the idea of logistic statistical modeling and tests homogeneity. The popular IRT models, the one-, two-, and three-parameter logistic models make a strong assumption that individual test items measure a single construct (Hambleton & Jones, 1993). In this study the construct being measured is statistics knowledge (ability). IRT uses the items as the units of measure to obtain ability scores that are on the same scale that is called ability scale (Baker, 2001). A reasonable assumption underlying IRT is that each examinee responding to a test item possesses some amount of ability that can be represented by a numerical value, a score. An examinee's score placeS him or her somewhere on the ability scale. The ability score will be denoted by θ (theta). In this study, the ability scale consists of three levels: Level A, B, and C. At each ability level, there will be a certain probability that an examinee with that ability will give a correct answer to the item that is denoted by $P(\theta)$.

Considering $P(\theta)$ as a function of ability, the plot of the function will be a smooth, S-shaped curve. In the case of measuring students' developmental level in learning statistics, Figure 10 shows three hypothetical item characteristic curves that discriminate students into three levels, Level A, B, and C where item 1 measure an ability that is owned by Level A students (consequently it is also owned by Level B and Level C

students). On the other hand, item 2 measures the ability owned by Level B students (consequently it is also owned by Level C students but more likely it is not owned by Level A students). A similar explanation holds true for item 3 that is assigned to measure an ability that has to be owned by Level C students that more likely is not owned by Level A and Level B students.

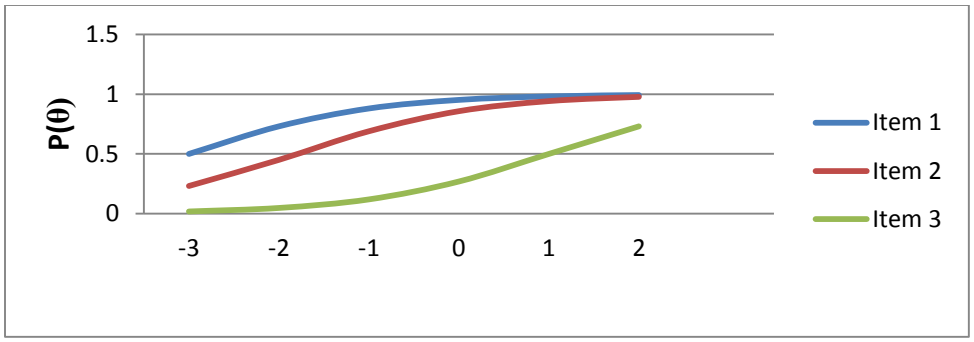


Figure 10. Item characteristic curves.

In recognizing item difficulty, Figure 11 shows three hypothetical item characteristic curves that all have the same level of discrimination but differ with respect to difficulty. In this case, item 3 is easier than items 1 and 2; meanwhile item 2 is easier than item 1. All three items are assigned to measure the ability of Level C students.

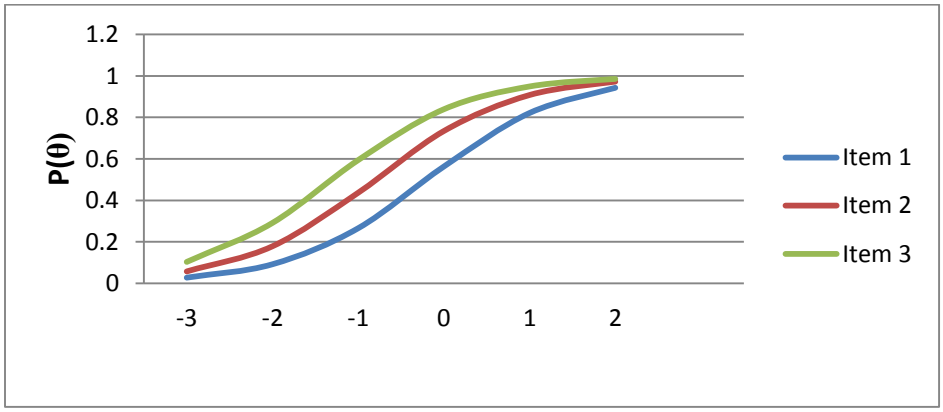


Figure 11. Level C item characteristic curves with different difficulty.

The last analyses conducted in this study were several ordinal regression analyses to investigate the relationships between students' GAISE developmental level and their school grade levels, their latest mathematics course taken, the survey form that they took, and their ages. These analyses were exploratory in nature, since no research question was addressed by this analysis. These analyses, however, are considered important since they will give significant information to uncover the relation among GAISE Levels and other factors that are considered influential for students' developmental level in learning, especially in learning statistics concepts.

Ordinal Regression Analysis

Marija J. Norusis (2011) explains that in ordinal logistic regression, the event of interest is observing a particular score or less. The ordinal logistic model for a single independent variable is given by the following equation:

$$\ln(\theta_j) = \alpha_j - \beta X$$

where $$\theta_j = \frac{prob\ (score \leq j)}{prob\ (score > j)} = \frac{prob\ (score \leq j)}{1 - prob\ (score \leq j)}$$

and *j* goes from 1 to the number of categories minus 1. In this study there are three categories of dependent variables involved: Level A, Level B, and Level C. So, *j* goes from 1 to 2. The minus sign before the coefficients for the explanatory variables is used so that larger coefficients indicate an association with larger scores. For example, a positive coefficient for a dichotomous factor means higher scores are more likely for the first category and a negative coefficient means that lower scores are more likely. For a continuous variable, a positive coefficient means when the values of the variable increase, it is then likely that the larger scores increases. An association with higher

scores means smaller cumulative probabilities for lower scores, since they are less likely to occur. Each log odds of $prob\ (score \leq j)$ has its own α_j term but the same coefficient β. That means that the effect of the independent variable is the same for different logit functions, an assumption that has to be checked. The α_j terms, called the threshold values, are seldom of much interest (Norusis, 2011).

All data analyses conducted in this study have primary goals to answer the research questions and to provide evidence of the validity and reliability of item responses. To achieve these goals, an anticipation of possible threats to validity is necessary. The description of the identified threat is presented below.

Threats to Validity

There are two possible threats that are identified; the threats and anticipation plan are listed below:

- Absent or unclear conceptual match evident between the instrument and the intended results. To anticipate this threat a search for expert panelists to align items with the Pre-K-12 GAISE Framework and school mathematics curriculum was initiated. The experts not only helped in validating the items' contents and appearance, but also helped align items with their appropriate levels suggested by the literature. The experts' opinions were extremely helpful for revising the instruments.
- The long time needed to evaluate items could decrease the response rates of the surveys. The less the number of items, the more likely that the experts/ students involved in this study are able to assist; considering time consumed in reviewing items. Optimizing the number of items that are large enough for a valid CTT-

based item analysis and SEM-based construct analysis but also small enough to guarantee that the time needed by the experts to validate the contents of the items is adequate resolved this threat.

IRB Exemption

According to federal regulation, the IRB may determine a research activity to be exempted where the only involvement of human subjects is in several categories. One of the categories is the research conducted in established or commonly accepted educational settings, involving normal educational practices. This study fulfills the exempt Categories of Research listed in 45 CFR, Part 46, Sec. 101 (b) for the following reasons: this study involves the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures or observation of middle- and high- school students behavior, and (i) information obtained is recorded so that the middle- and high-school students participating in this study cannot be identified directly. The students were identified using a number code, so that their responses were not linked to their personal information; (ii) since the students' responses only reflected their cognitive ability, then the use of their responses outside the research will not place the students at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, or reputation. Therefore, an IRB exemption for this study was requested and has been approved. The IRB exemption number is EXP2012B6438.

Summary

There are two instruments that have been developed in this study: the expert survey instrument and the student survey instrument. Several items of the students' survey instrument were adapted from earlier studies (Garfield, 1991, 2003; Callingham &

Watson, 2005; Watson, 1997; Watson & Callingham, 2003; OECD, 2009). New items were developed based on Pre-K-12 GAISE Framework and Common Core State Standards guidelines.

The first expert survey was conducted during fall 2011. In the first survey, two experts were asked to judge the alignment of 40 multiple-choice items with statistics problem solving process components and GAISE levels suggested by Pre-K-12 GAISE Framework (Franklin, et al., 2007). They were also asked to give their opinion regarding the clarity of the items. Based on the experts' input, some items were revised and others were replaced. The forty items were classified to measure the GAISE levels (Level A, B, and C) following experts' opinions.

The forty items were divided into three forms to ensure that participants only required less than one hour to respond to all items in each Form. The items were piloted during summer 2012. Pilot study data were analyzed using CTT-based analyses. Results of the pilot study were used to revise the items. Thirty six items were then used for the actual survey that was administered during fall 2012 to 797 middle and high school students in Central Texas. Several ways to estimate the validity of the scores acquired in the measurement were conducted, including face validity, content validity, and construct validity (Bornstein, 2003; Vogt, 2007).

In order to answer the first question, descriptive statistics of students' difficulty indices across all process components were analyzed. From the literature and experts' review, four models that describe the alignment between items and levels and the relation among levels were developed. A Confirmatory Factor Analysis using Structural Equation

Modeling (SEM) approach was conducted in which model fit testing was applied to examine to what extent the models fit the data.

The results were used to support or reject the hypotheses of the developmental progression of statistical investigation process components that were explained by the models. The best model for each form was chosen and used as the guideline for aligning items with their appropriate levels. Descriptive statistics of students' performances in each level were then analyzed. These analyses were conducted to answer the second and the third research question. The results described whether the developmental levels suggested by the Pre-K-12 GAISE Framework remain true for each statistical investigation process component.

To answer the fourth research question, several psychometric and statistical analyses were conducted that included CTT and SEM analyses using SPSS AMOS 21 (IBM, 2012) and Mplus Version 7 (Muthen & Muthen, 2012) programs. Some threats of validity were identified. This study fulfilled the exempt Categories of Research listed at 45 CFR, Part 46, Sec. 101 (b), therefore, an IRB exemption for this study was requested and has been approved.

CHAPTER IV

RESULTS

In this chapter, the results of data analyses conducted in this study will be presented. The results address the following research questions:

1. How and how much do middle school and high school students understand statistical concepts that are related to the investigation cycle (formulating questions, collecting data, analyzing data, and interpreting result)?
2. What are the learning trajectories that describe the developmental progression for different concepts and statistical investigation processes?
3. To what extent do students' understandings of statistical concepts develop similarly across developmental levels?
4. Given the structure of the progressions observed in performance of different levels, to what extent can students' developmental levels be diagnosed reliably and validly?

Literature reviews, expert reviews, and two important documents serve as the bases in developing the instruments to measure students' developmental levels and learning trajectories of statistics in this study. The two documents are (1) the Pre-K-12 GAISE Framework (Franklin, et al., 2007) and (2) the K–12 Common Core State Standards in Mathematics (National Governors Association Center for Best Practices

and Council of Chief State School Officers, 2010). The Pre-K-12 GAISE Framework comprehensively addresses student learning objectives and gives detail guidance for instruction and assessments in the areas of statistics and probability that are consistent with research on statistical learning. The framework presents the statistical problem solving process and then, suggests how this process can be presented at each of three developmental levels (Levels A, B, and C). The framework considers a two-dimensional model of a conceptual structure for learning statistics. The first dimension consists of components of the statistical problem-solving process that involves four components: question formulation, data collection design and implementation, data analysis, and interpretation (Franklin, et al., 2007). The second dimension includes the three developmental levels of statistical education; Levels A, B, and C. The framework also provides an additional emphasis on understanding the role of variability in the problem solving process (Franklin, et al., 2007).

The K–12 Common Core State Standards in Mathematics (CCSS-M) contains ambitious expectations for statistics in grades 6-12. At Grade 6, in the CCSS-M, students are expected to develop understanding of statistical variability and summarize and describe distributions. At Grade 7, students are expected to use random sampling to draw inferences about a population and draw informal comparative inferences about two populations. At Grade 8, students are expected to investigate patterns of association in bivariate data. At the high school level, students are expected to be able to interpret categorical and quantitative data, make inferences and justify conclusions, understand conditional probability and the rules of probability, and use probability to make decisions.

Groth and Bargagliotti (2012) suggested that the learning expectation of the CCSS-M for statistics naturally fall under the Pre-K-12 GAISE Framework, thus allowing practitioners to use the Pre-K-12 GAISE Framework as a roadmap to help implement the CCSS-M. In addition, Pre-K-12 GAISE Framework is a compelling supplement to the CCSS-M because it offers the following ideas that are not contained in the CCSS-M: (1) pedagogical approaches for statistics; (2) meaningful statistical connection; (3) developmental trajectories for students' statistical learning; and (4) enhancement of the curriculum prescribed by the CCSS-M (Groth & Bargagliotti, 2012).

The instrument developed in this study is targeted to investigate Grade 6 -12 students' developmental level and learning trajectory of statistics. The three GAISE levels (Levels A, B, and C) are considered latent variables within a latent variable framework that determine the responses to the instrument items which are considered observed variables. Using Structural Equation Modeling (SEM) approach, four structural equation models were developed to explain the relationships among items and GAISE levels. The development of the models was based on the GAISE and the Common Core State Standards in Mathematics (CCSS-M), pilot study, and experts' opinions. It was expected that the structural equation models fit the data well, so that an inference can be made about students' developmental levels and learning trajectory in statistics. In this chapter, results on model fit test analysis will be presented.

The following description explains the organization of this chapter. First, sample and general information of participants are described. Second, the descriptions of all items in the instrument are presented. Third, the classical test theory analysis results of the instrument developed in this study are discussed. The description includes reports on

internal consistency score reliability of the instrument used including the relationship between reliability and validity, the difficulty indices of the items, and also point-biserial indices of each item are displayed. Fourth, establish validity evidence, the results of a SEM-based CFA, are presented. Fifth, analyses of students' performances at each level are discussed.

Sample

The total number of participants in this study was 797 students, where 649 (81.43 %) students were from one middle school and 148 (18.57 %) students were from one high school. Both schools are located in the Central Texas area. Both schools also have similar demographics. Table 3 below exhibits the number of participants by grade levels.

Table 3

Number of Participants by School Grade Level

Middle School Grade Levels	Number of Participants	High School Grade Levels	Number of Participants
Grade 6	220 (27.60 %)	Grade 9	18 (2.26 %)
Grade 7	206 (25.85 %)	Grade 10	84 (10.54 %)
Grade 8	223 (27.98 %)	Grade 11	25 (3.14 %)
		Grade 12	21 (2.63 %)
Total Middle School	649 (81.43 %)	Total High School	148 (18.57 %)

Percentages of middle school students from all grade levels are quite uniform. This pattern is not the same for high school participants. More than half of high school participants are Grade 10 students. However, proportions of Grade 9, 11, and 12 student participants are also uniform.

Table 4 shows the classification of participants based on the most recent mathematics courses they had taken at the time of the instrument administration. Regular mathematics courses in Middle School were categorized as academic courses; meanwhile Pre-AP Mathematics Courses, Algebra I, and Geometry taken in Middle School were categorized as advanced Middle School mathematics courses. Additionally, regular mathematics courses in High School, Algebra I & II, and Geometry were categorized as academic courses; meanwhile Pre-Calculus, Calculus I, II, and III, and Mathematics Model taken in High School were categorized as advanced High School mathematics courses.

Table 4

Number of Participants by Latest Mathematics Courses Taken

Courses Taken in Middle School	Number of Participants	Courses Taken in High School	Number of Participants
Academic	228 (28.61 %)	Academic	107 (13.43 %)
Advanced	421 (52.82 %)	Advanced	41 (5.14 %)
Total Middle School	649 (81.43 %)	Total High School	148 (18.57 %)

Most of middle school student participants (421 of 649 students) were taking advanced mathematics courses; meanwhile only 41 of 148 high school student participants were taking advanced high school mathematics courses. How these facts influence students' responses of the instrument was investigated.

The survey instruments were divided into two forms; Form 1 and Form 2. The number of participants who took Form 1 is 140; meanwhile the number of participants who took Form 2 is 657. Table 5 displays the classification of participants who took Form 1. From Table 4.3 it can be seen that the proportion of middle school and high

school students who took Form 1 were not too different compared to those who took Form 2. As can be seen in Table 6, only about 10% of students who took Form 2 were high school students, and almost half of high school students taking Form 2 were Grade 10 students.

Table 5

Number of Participants Taking Form 1

Middle School Grade Levels	Number of Participants	High School Grade Levels	Number of Participants
Grade 6	3 (2.14 %)	Grade 9	8 (5.71 %)
Grade 7	34 (24.29 %)	Grade 10	55 (39.29 %)
Grade 8	21 (15 %)	Grade 11	11 (7.86 %)
		Grade 12	8 (5.71 %)
Total Middle School	58 (41.43 %)	Total High School	82 (58.57 %)

This unbalanced proportion of the number of middle school and high school participants might affect the outcomes of this study because the instrument was designed to include items that are more likely to only be responded correctly by high school students. Item analysis and confirmatory factor analysis of both forms are presented in the later sections.

Table 6

Number of Participants Taking Form 2

Middle School Grade Levels	Number of Participants	High School Grade Levels	Number of Participants
No response	5 (.8 %)	Grade 9	10 (1.5 %)
Grade 6	211 (32.1 %)	Grade 10	29 (4.4 %)
Grade 7	173 (26.3 %)	Grade 11	14 (2.1 %)
Grade 8	202 (30.7 %)	Grade 12	13 (2.0 %)
Total Middle School	591 (89.95 %)	Total High School	66 (10.05 %)

Validity

As mentioned in the previous chapter, validity is possibly the most important criterion for the quality of an instrument. The term validity denotes whether an instrument measures what it purports to measure. In this study, several ways to estimate the validity of an instrument were conducted, including content validity, face validity, and construct validity (Bornstein, 2003; Vogt, 2007).

The instrument was reviewed for content validity by sending the 40 items developed in fall 2011 to three experts in statistics and statistics education. The experts were asked to align the items with GAISE Levels and one expert also aligned the items with statistical process components suggested by the Pre-K-12 GAISE Framework (Franklin, et al., 2007). Experts were provided with the Pre-K-12 GAISE Framework summary, and were invited to indicate which item, if any, were unclear. The feedback from the experts then was used to revise the items. The expert survey was also conducted for face validity of the instrument. After revision, for face validity and to improve the quality of the instrument, the items were piloted on 19 middle school and high school students who participated in the Summer Math Camp organized by the Texas Mathworks and to 66 undergraduate students during summer 2013. A classical test theory (CTT) item analysis was applied to the data of students' responses in the pilot study. The results are displayed in Table 7. Four highlighted items are those that were removed from the instrument. Those items were excluded because they had a too large or too low percentage correct (difficulty indices) and too low point-biserial indices. The item would have low discrimination if it was so difficult that almost everyone got it wrong or guessed, or so easy that almost everyone got it right. On the other hand, the low point-

Table 7

Pilot Study Item Analysis Results

ITEM	PCT	LEVEL	PT BISERIAL	ITEM	PCT	LEVEL	PT BISERIAL
1	89.7	B	0.601**	21	35.7	C	0.643**
2	24.1	B	0.190	22	32.1	A	0.339
3	93.1	B	0.026	23	92.9	A	-0.198
4	27.6	B	0.456*	24	46.4	B	0.533**
5	20.7	B	0.202	25	21.4	B	0.069
6	37.9	C	0.401*	26	25.0	C	0.204
7	37.9	B	0.518**	27	93.1	A	0.255
8	31.0	B	0.530**	28	82.8	B	-0.010
9	65.5	B	0.333	29	48.3	A	0.552**
10	34.5	B	0.327	30	79.3	B	0.424*
11	100.0	A	0.000	31	34.5	B	0.528**
12	72.4	A	0.357	32	31.0	C	0.253
13	93.1	A	0.407*	33	82.8	A	0.223
14	32.1	B	0.318	34	69.0	C	-0.253
15	67.9	B	0.283	35	41.4	C	0.511**
16	71.4	C	0.449*	36	86.2	B	0.311
17	35.7	B	0.400*	37	34.5	C	0.433*
18	60.7	A	0.296	38	20.7	C	0.105
19	78.6	A	0.590**	39	31.0	C	0.176
20	82.1	A	0.326	40	75.9	B	0.197

biserial coefficient of an item indicates that the correlation between students' responses to the item and their total scores is not strong. This means the power of the item to discriminate students based on their ability was weak. One item in the pilot study that has small point-biserial (item 34) was not excluded from the instrument since it is a level C item that is needed in order to have a good proportion of number of items for each level in the instrument. This item was also adopted from a large scale study and has been proven to be a good item. The low point-biserial of this item in the pilot study was more

likely caused by the changes of multiple choice distractors. In the actual survey this item is identified as ITEM15 that actually has very good difficulty index (.42) and point-biserial (.496). The decision to keep this item proved to be a good decision.

Items 38 and 39 were kept in the instrument because both are Level C items that needed to be kept for balancing the number of items in the instrument. Especially for Item 38, several changes in the wording should improve the quality of this item. Since item 40 had a good difficulty index (75.9) and its point-biserial index (.197) was just slightly smaller than .2, this item was also kept in the instrument. A bad decision was made for keeping Item 11 that had 100% correct response. This item, identified as ITEM01 in the actual instrument also had 100% percentage correct response in the actual survey that was discarded from the analysis due to its zero variance.

Construct validity in this study was tested by conducting the SEM-based confirmatory factor analysis (CFA) by performing model fit testing using SPSS AMOS 21 program. The initial model was developed from the first expert survey results. The second model was developed by input from Expert 1, and the third model was developed by input from Expert 2. The fourth model was developed by exploring many combinations of Expert 1's and Expert 2's alignments of items with GAISE levels. The best model was chosen: the combination model.

The results showed that the models fit the data well which indicates that the instrument actually measured the three constructs that were intended to be measured, in this case, the GAISE Level A, Level B, and Level C. This construct validity test results are thoroughly discuss in the confirmatory Factor Analysis section.

Expert Survey Results

Two experts reviewed the thirty-six items in the instrument during spring 2013. The experts aligned each item with its appropriate developmental level suggested by the Pre-K-12 GAISE Framework. In several items both experts agreed with the aligned developmental levels, however they also disagreed with some of the items. Since the experts did not meet with each other during the survey, there was no consensus made about the alignments of items with GAISE Levels. This affected the validity of the instrument because one item could have been aligned to two different levels by the two experts. So, the accuracy of the instrument to measure students' developmental level in statistics was questionable. The inter-rater reliability of the instrument was also considered very low. As mentioned previously, an instrument that is not reliable must be invalid. To handle this validity issue, a structural model was developed for each expert's alignment then later both models were compared with a model that was developed by conducting an exploratory modeling using all possible combinations of both experts' item-level alignments. The best combination model was chosen as the model that best followed one of or both experts' opinions and had the best goodness of fit indices. The validity issues that were related to the inter-rater reliability of the instrument should have been taken into serious consideration for future development of the instrument. The alignment of items into GAISE Levels following the combination model is presented in Table 8 below. With the validity issues in mind, all results of data analyses conducted in this study is reported in the following section.

Table 8

Item Descriptions and Experts' Developmental Level Alignment

	ITEM	Description	Level
Form 1	ITEM 01	Analyzing pictograph of students' transportation modes to go to school.	A
	ITEM 02	Choosing the right box to get a blue marble if one box contains 6 red & 4 blue marbles and another box contains 60 red & 40 blue marbles.	B
	ITEM 03	Choosing the best collecting data method for estimating the proportion of residents who support the increased tax.	B
	ITEM 04	Estimating the number of fish in a farmer's dam by tagging 200 fish and finding out that 25 of 250 fish taken from the dam in the next day are tagged.	C
	ITEM 05	Choosing a statistical question from four different questions.	A
	ITEM 06	Predicting which sequence is most likely to result from flipping a coin 5 times.	B
	ITEM 07	Estimating the probability of winning a prize in a game booth by getting an even number in a spinner containing five of six even numbers and then picking a black marble from a bag containing 6 black and 14 white marbles.	A
	ITEM 08	Choosing a statistical question based on Miller MS basic health information data.	B
	ITEM 09	Judging the appropriateness of a sampling plan of a survey to study students' feeling about Miller MS Cafeteria's food.	B
	ITEM 10	Choosing the most appropriate description of possible outcomes in flipping a coin 10 times versus flipping a coin 100 times.	B
	ITEM 11	Judging a report presented by a TV reporter who said that a bar graph showed the number of robberies increases significantly only by looking at the height of the bars without considering other factors.	A
	ITEM 12	Choosing a more reliable recommendation by comparing recommendations from Consumer Reports and three friends about the performance of two different brands of cars.	B
	ITEM 13	Determining the most likely list of number of red candies taken in five trials of taking 10 candies from a bowl containing 20 yellow, 50 red, and 30 blue candies.	C
	ITEM 14	Cody plays a game involving two half-black and half-white spinners; A player wins if both arrows of the spinners land on black. Determine whether Cody's belief that he has a 50-50 chance of winning is correct or not.	A
	ITEM 15	What conditions need to be fulfilled to determine when change in response variable X causes a change in predictor variable Y?	C
	ITEM 16	Given three graphs represented data of experiments conducted by students, determine which graph is more likely made up.	C
	ITEM 17	Making inference of the graph of scores on a science test taken by two groups of students.	B
	ITEM 18	Inferring the correlation between getting lung disease and smoking cigarette based on survey data given in two by two way table	C

Table 8 continued

Form 2	ITEM 20	Determining probability of winning a recent state lottery awards of two people: Bill and Bob, given the condition that Bill has not won a single prize and Bob just won a $20 prize last week.	C
	ITEM 21	Determining the probability of picking a red candy from a bag of colored candies where the number of each colored candy in the bag is given in a dot plot.	A
	ITEM 22	Determining the most likely outcome of tossing a coin for the fifth time, given the condition that in four successive tosses, a fair coin lands heads up each time.	C
	ITEM 23	Determining the most likely events in throwing three dice given four possible outcomes (assessing students' knowledge of the theoretical probability of throwing three dice).	C
	ITEM 24	Determining the appropriate method to approximate the weight of an object measured several times; given the list of measured weights: 6.3, 6.0, 6.0, 15.3, 6.1, 6.3, 6.2, 6.15, and 6.3.	A
	ITEM 25	Determining the best method to collect data to determine which group of students can jump farther, boys or girls.	A
	ITEM 26	Determining the price of milk in April 2004, given a graph of monthly prices of milk in the United States from 2003 to 2012.	A
	ITEM 27	Determining which data has larger standard deviation, given two sets of data represented by two histograms.	B
	ITEM 28	A small sample (500 of 1,000) is taken from a large school and a larger proportion of sample (20 of 300) is taken from a small school. Which sample does give a strange proportion of boys (80 %) given the condition that both schools have the same percentages of boys and girls?	B
	ITEM 29	Determining which survey is the best, given several sampling methods conducted by four students, Shannon, Jake, Adam, and Claire.	B
	ITEM 30	Determining the average class size of fifth-grade classrooms in a town given the averages class size of fifth grade classrooms of all schools in the town which have various numbers of fifth grade classrooms.	A
	ITEM 31	Determining the most appropriate method to collect data to find out whether beans grow faster in the dark or in the light.	B
	ITEM 32	Determining the most appropriate variable for the horizontal axis of a histogram, given four possible variables.	A
	ITEM 33	Determining the probability of choosing another boy after choosing 2 boys of 20 students consisting of 10 boys and 10 girls.	A
	ITEM 34	Determining the association of getting lung cancer and cigarette smoking from an experiment in England where the frequency proportions of lung cancer cases for each level of smoking are given.	B
	ITEM 35	Given a two by two count of students who like/dislike rock or rap, determine the association between like/dislike rock and like/dislike rap.	B
	ITEM 36	Making inference of a survey about students' favorite desserts by simulating the survey using random odd and even integers 100 times to conclude whether the result that shows 58% of students like ice cream is due to chance variation alone or not.	C

Descriptive Analysis Results

In order to answer the first research question in this study, the distribution of items into Pre-K-12 GAISE process components is needed. Table 9 shows the classification of items based on the correspondent process components and their difficulty indices.

Table 9

Distribution Item Based on Pre-K-12 GAISE Process Component

Process Component	ITEM Number	Mean	Standard deviation	ITEM Number	Mean	Standard deviation
Formulating Questions	5	.45	.499	8	.32	.467
Collecting Data	3	.56	.499	13	.38	.487
	25	.61	.487	27	.47	.499
	29	.33	.472	31	.58	.513
Analyzing Data	1	1.00	.000	4	.22	.417
	9	.29	.453	11	.36	.483
	16	.28	.448	19	-	-
	21	.42	.494	24	.24	.427
	26	.51	.500	30	.21	.494
	32	.31	.462	33	.28	.448
Interpret Results	6	.81	.396	7	.66	.474
	14	.57	.497	15	.42	.496
	17	.36	.482	18	.19	.396
	34	.55	.498	35	.40	.490
	36	.24	.425			
Nature of Variability	2	.54	.501	10	.14	.344
	12	.39	.489	20	.41	.493
	22	.63	.482	23	.15	.358
	28	.22	.414			

As a reminder, the first question is: “how and how much do middle school and high school students understand statistical concepts that are related to the investigation cycle (formulating questions, collecting data, analyzing data, and interpreting result)?” Figure 10 presents the box plots of difficulty indices for all process components. The

difficulty index of an item is equal to 1 – percentage correct responses of the item. For example, the difficulty index of ITEM 5 is 1 - .45 = .55.

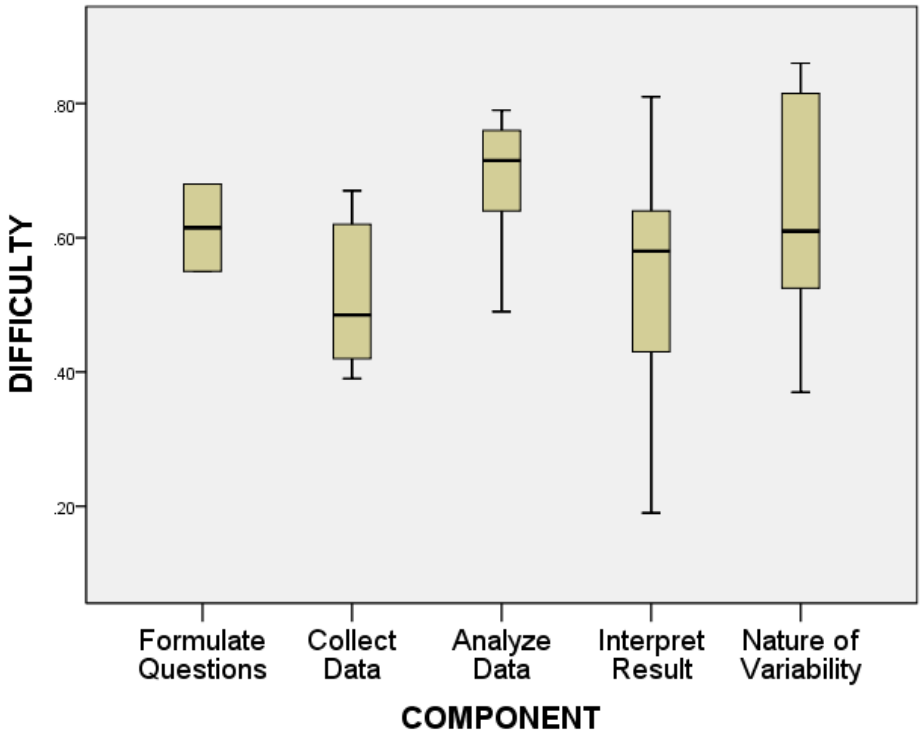


Figure 12. Boxplots of difficulty index of items across process components.

From the table and the box plots, it can be seen that more than 30 % of students answered items in formulating questions and collecting data correctly. In analyzing data, it was found that less than 30 % of participants answered six of 12 items (ITEM 4, 9, 16, 24, 30, and 33) in this process component correctly. This means that students did not perform as well as in the previous process components. Students performed quite well in interpreting results. Only two of nine items (ITEM 18 and 36) were responded to correctly by less than 25 % of participants. All other items in this process component were answered correctly by more than 38% of the participants. In Nature of Variability, it was found that three of seven items had low correct responses (ITEM 10, 23, and 28).

According to an expert who contributed in this study, these three items actually assessed students' understanding on natural and chance variability and variability within a group and variability between groups. These results suggest that students revealed a lack of understanding of the analyzing data process component and the nature of variability.

Even though Figure 10 shows that the mean of difficulty indices of Analyze Data was less than the means of difficulty indices in other process components, by the Kruskal-Wallis test it was found that there was no evidence that the means of difficulty indices of the five components were different. The Kruskal-Wallis test was conducted after the Q-Q plots of the difficulty indices residuals revealed that the difficulty index data was positively skewed. With significance level .13, it was likely that the large value of mean of difficulty indices in Analyze Data process component was due to chance alone. This result, however, might also have been caused by a sample size that was small (36).

To answer the second research question: "What are the learning trajectories that describe the developmental progression for different concepts and statistical investigation processes?" the alignment of items with GAISE Levels suggested by the combination of experts' opinion was applied. The detailed analyses are presented below. Before we answer the second research question, first we investigated whether the difficulty indices of items across levels showed a tendency suggested by the Pre-K-12 GAISE Framework where the higher the level of the items, the smaller the percentage correct answer of the items. This is explained by the fact students who have developed into level B must be able to answer all Level A items correctly which leads to the higher percentage correct of Level A items.

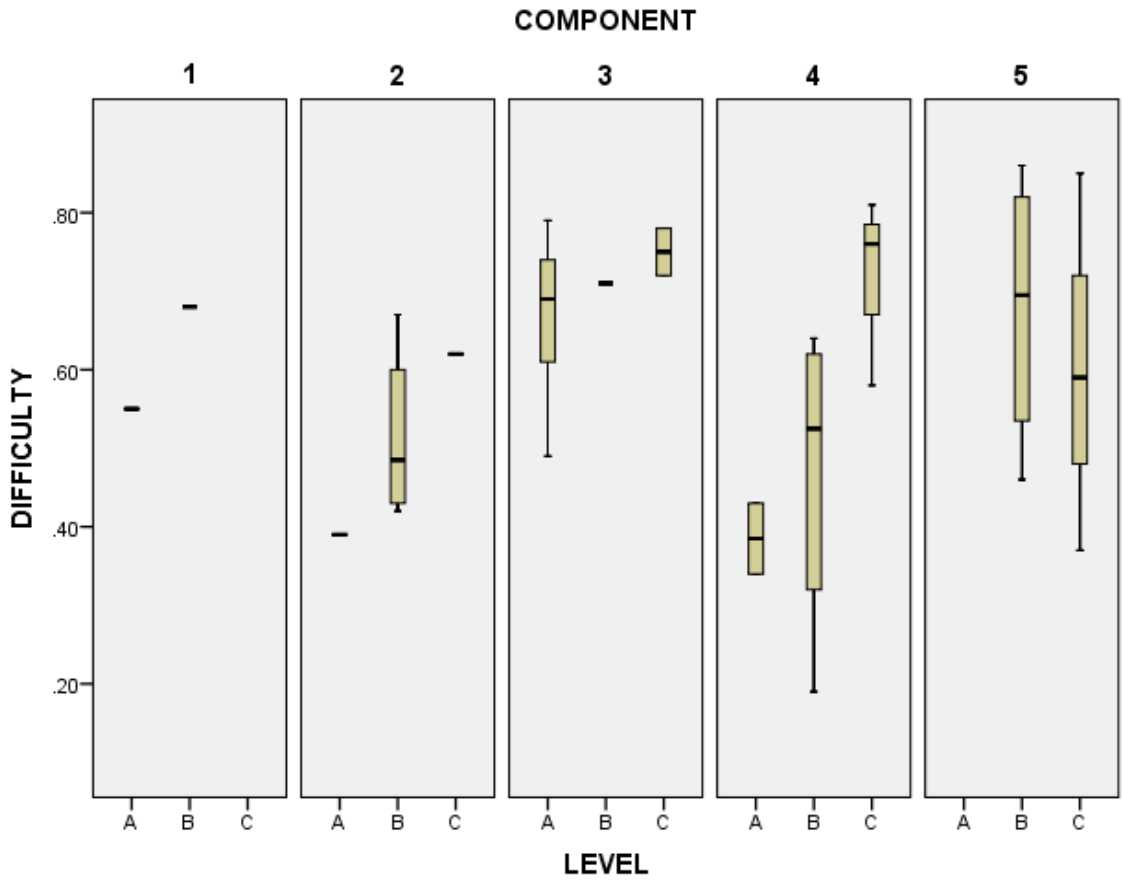


Figure 13. Boxplots of difficulty index of items across levels.

Component 1 represents Formulating Question, Component 2 represents Collecting Data, Component 3 represents Analyzing Data, Component 4 represents Interpreting Result, and Component 5 represents Nature of Variability process component.

Figure 11 shows that the higher the developmental levels of the items the higher the difficulty indices of the items in all process components except for the Nature of Variability component. These results agree with Pre-K-12 GAISE Framework's suggestion that students develop their understanding through three developmental levels. Level C items should have higher difficulty levels than items from lower developmental levels, since only students who have developed into Level C could answer Level C items

correctly. Likewise, Level B items should have lower difficulty level than Level C items, since not only Level B students are more likely to respond to Level B items correctly, but also Level C students that cause the percentage correct of level B items tend to be higher than Level C items, hence the difficulty indices of Level B items tend to be smaller than those of Level C items.

It is interesting to investigate how the difficulty indices of items from different process components behave in each form. Figure 12.a displays the data of difficulty indices of items in FORM 1. It was found that the higher the Level, the higher the mean of item difficulty indices.

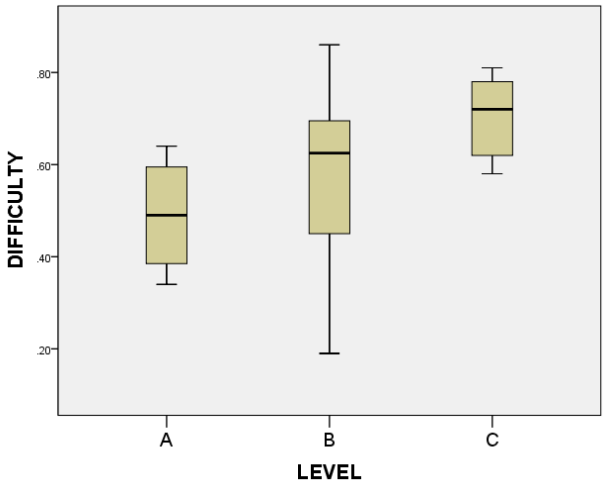


Figure 14. Boxplots of difficulty index of items in Form 1.

A different phenomenon was shown by the data of students' responses of Form 2 shown in Figure 12.b. Figure 12.b shows that Level A items tended to have higher difficulty indices than Level B and Level C items. This phenomenon disagrees with the Pre-K-12 GAISE Framework' which hypothesizes that a student who has developed into Level B should have gone through Level A stage as explained before. This phenomenon

seems to disagree with the Pre-K-12 GAISE Framework and needs to be investigated further.

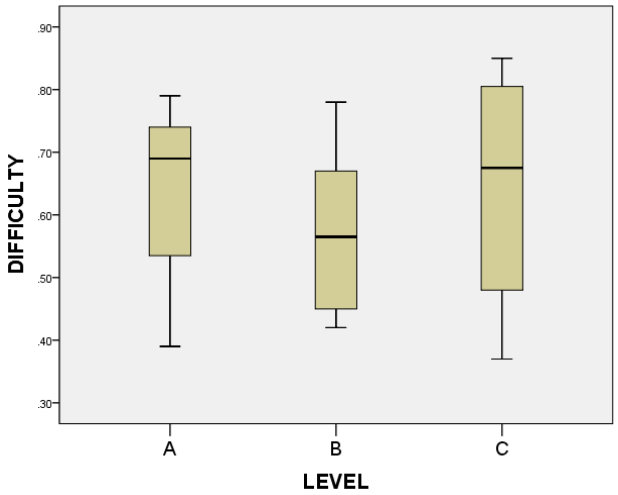


Figure 15. Boxplots of difficulty index of items in Form 2.

Now, the analyses focus on answering the second research question: to investigate the learning trajectory of the statistical investigation process components. It was found that for the collecting data process component, there were four Level B items and one Level A item, and also one Level C item. Comparing the difficulty indices of the items across levels, it was found that in Form 1, one Level B item (ITEM 03) had higher difficulty index (56 %) than the Level C item (ITEM 13) with percentage correct 38%; meanwhile in Form 2, the Level A item had percentage correct 61% meanwhile Level B items had percentage correct range from 33% to 58%. This suggests a tendency that the higher the level of an item, the fewer students can answer the item and this finding agrees with the Pre-K-12 GAISE Framework's suggestion.

For analyzing data process component, it was also found that Level A items had higher percentage correct than Level B and Level C. Level B item also had higher percentage correct than Level C items.

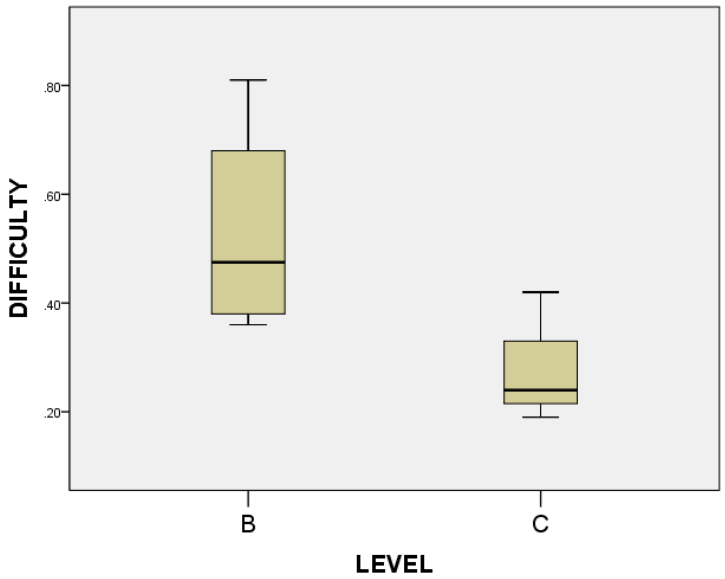


Figure 16. Boxplot of difficulty index of interpreting results item for Level B and C.

For the interpreting result process component, it was found that the higher the level of the items, the higher the difficulty level of the items that were indicated by lower percentages correct answer of the items. Figure 13 displays the boxplots of difficulty indices of Level B and Level C items in the interpreting result process component that explain this tendency.

For the understanding variability items, however, Level B items tended to have lower difficulty indices than Level C items, which can be seen in Figure 14. This indicates that students' understanding of the nature and focus of variability might not develop similarly as their understanding of the statistical process component: formulating questions, collecting data, analyzing data, and interpreting result. This result, yet, might

also be caused by the alignment of items into levels that were determined by combining experts' opinions.

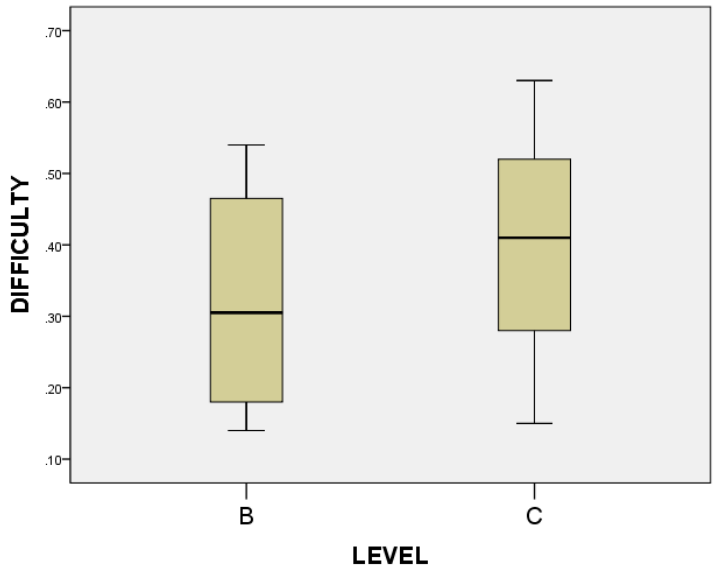


Figure 17. Boxplots of difficulty index of understanding variability items.

The description above also answered the third research question: To what extent do students' understandings of statistical concepts develop similarly across developmental levels?", by showing that students' understanding on formulating questions, collecting data, analyze data, and interpret results tend to develop similarly across developmental levels. The data showed a tendency that students develop their understanding following the developmental level suggested by the Pre-K-12 GAISE Framework for those process components. For nature of variability, however, the tendency was different. Students seemed to have developed into Level C mastery meanwhile their Level B mastery had not been developed yet. Further investigations are

needed considering a limitation of this study, for instance, there are not enough items in each process component for supporting a robust statistical analysis.

Although the boxplots showed the tendency that the higher the level the higher the item difficulties for each process component, a non-parametric test to compare means of item difficulties in each process component revealed that the difference is not significant. A Kruskal-Wallis test to compare means of item difficulties across levels for Formulate Questions revealed the p-value equal 1. Another Kruskal-Wallis test to compare means of item difficulties across levels for Collect Data Process component also revealed the p-value equal .287. Likewise, a Kruskal-Wallis test to compare means of item difficulties across levels for Analyze Data Process component revealed the p-value equal .499. Two Mann-Whitney U tests to compare means of item difficulties across levels for Interpret Result and Nature of Variability Process Component revealed the p-value equal .229 and .629 respectively. Therefore, the conclusion that students develop their statistical understanding through hierarchical levels, Level A, Level B, and Level C in each process component was not supported by the item difficulties data. The small number of items for each process components and for each level might have caused the large p-value. Further studies that involve more items would give more convincing results.

Another analysis to compare the means of item difficulties for all process components in each level also revealed that item difficulties of all process components in each level had no significant differences. Figure 14.a shows the boxplots of item difficulties for all process components for all three levels. A Kruskal-Wallis test to compare means of item difficulties for all process components for Level A showed the p-value of .084. This concluded that the differences among the means of item difficulties of

all process components in Level A were not significant. Similarly, a Kruskal-Wallis test to compare means of item difficulties for all process components for Level B showed the p-value of .231. This concluded that the differences among the means of item difficulties of all process components in Level B were also not significant. A similar test was conducted for Level C, and the result lead to the same conclusion: the differences among the means of item difficulties of all process components in Level C were also not significant (p-value = .891).

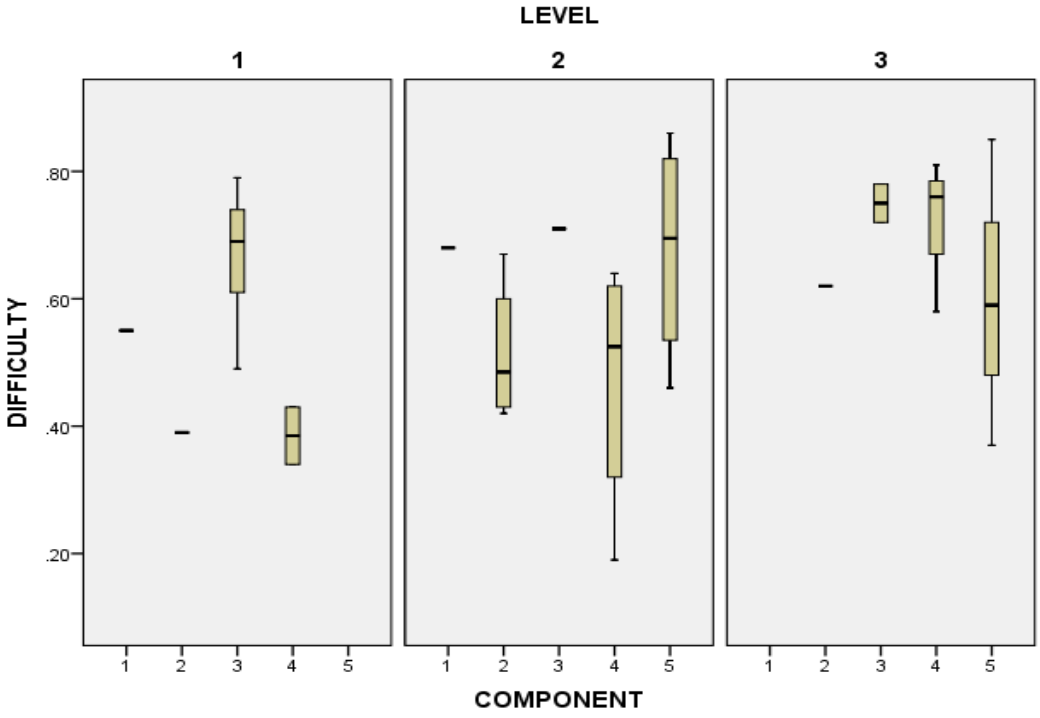


Figure 18. Boxplots of item difficulties for each process components for each level. Component 1 represents Formulate Question, Component 2 represents Collect Data, Component 3 represents Analyze Data, Component 4 represents Interpret Result, and Component 5 represents Nature of Variability process component.

As described in the previous chapter, to answer the fourth research question, several psychometric and statistical analyses were conducted. First, Classical Test Theory

(CTT) analysis was conducted in order to examine the quality of items in the instrument. All analyses conducted in this study had a primary goal: to show to what extent the instrument developed in this study validly measured what it was supposed to measure, in this case, the developmental level of learning statistics. Therefore, the Structural Equation Modeling (SEM) analysis results were also useful as construct validity evidence by showing that the instrument indeed measured three different factors, in this case GAISE Level A, B, and C. The analysis showed to what extent this hypothesis was supported by the data. In the following discussion, the CTT and SEM analyses were provided.

Classical Test Theory Analysis Results

Preliminary analysis started by analyzing the responses using classical test theory (CTT). Table 10 displays the descriptive statistics and point-biserial indices of all items in Form 1. Participant's response of an item was coded as 1 if it was correct and 0 otherwise. Therefore, the mean values in the table represent the percentages of correct responses of the items. It can be seen in the table that Item 01, and 06 have low difficulty indices since they were responded correctly by more than 60% of the participants; on the other hand, ITEM 04, 10, and 18 have high difficulty indices, since less than 25% of participants answered them incorrectly. ITEM 01 will not be used for further analysis, since all students answered the item correctly. The implication of an item with zero variance (i.e. a 100% correct response rate) is that the item is not useful for discriminating among students. An additional factor leading to the removal of this item is because we cannot conduct further analysis in SPSS using data from this item due to its zero variance.

It was also found that ITEM 09 and 10 had point-biserial indices negative that indicated these items were potentially bad and needed to be removed from the instrument. ITEM 06 had a point-biserial that was slightly less than .2. Items with point-biserial indices that were greater or equal to .2 were considered reasonable items (Kline, 2005). All other items in Form 1 have point-biserial indices that were greater than .2.

Table 10

Descriptive Statistics and Point-biserial Indices of Items in Form 1

ITEM	N	Mean (p)	Std. Deviation	Point-biserial
ITEM 01	140	1.00	.000	0
ITEM 02	140	.54	.501	.505**
ITEM 03	140	.56	.499	.515**
ITEM 04	140	.22	.417	.545**
ITEM 05	140	.45	.499	.580**
ITEM 06	140	.81	.396	.236**
ITEM 07	140	.66	.474	.211*
ITEM 08	139	.32	.467	.444**
ITEM 09	140	.29	.453	-.029
ITEM 10	140	.14	.344	-.201*
ITEM 11	140	.36	.483	.484**
ITEM 12	137	.39	.489	.459**
ITEM 13	137	.38	.487	.204*
ITEM 14	138	.57	.497	.521**
ITEM 15	137	.42	.496	.251**
ITEM 16	138	.28	.448	.186*
ITEM 17	136	.36	.482	.390**
ITEM 18	135	.19	.396	.252**

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Table 11 displays the descriptive statistics and point-biserial indices of all items in Form 2. It can be seen in the table that ITEM 22 and 25 had low difficulty indices since

they were responded correctly by more than 60% of the participants; on the other hand, ITEM 23, 28, 30, and 36 had high difficulty indices, since less than 25% of participants answered them incorrectly. Since some items in the instruments were intended to measure higher development levels in learning statistics than those that were possessed by several participants, the facts that there were items with high and low difficulty indices were understandable. It does not mean the items were problematic.

Table 11

Descriptive Statistics and Point-biserial Indices of Items in Form 2

	N	Mean	Std. Deviation	Point-biserial
ITEM 20	655	.41	.493	.460**
ITEM 21	654	.42	.494	.495**
ITEM 22	655	.63	.482	.433**
ITEM 23	651	.15	.358	.147**
ITEM 24	650	.24	.427	.347**
ITEM 25	644	.61	.487	.409**
ITEM 26	647	.51	.500	.402**
ITEM 27	645	.47	.499	.102**
ITEM 28	646	.22	.414	.171**
ITEM 29	648	.33	.472	.410**
ITEM 30	646	.21	.405	.178**
ITEM 31	641	.58	.513	.409**
ITEM 32	648	.31	.462	.157**
ITEM 33	635	.28	.448	.438**
ITEM 34	635	.55	.498	.360**
ITEM 35	636	.40	.490	.418**
ITEM 36	640	.24	.425	.191**

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Table 11 also shows that there were six items with point-biserial less than 0.2. Those items were ITEM 23 (.147), ITEM 27 (.102), ITEM 28 (.171), ITEM 30 (.178),

ITEM 32 (.157) and ITEM 36 (.191). Considering their low point-biserial indices, it is possible that those items are problematic and might have to be removed from the instrument. Further analyses regarding these items will be discussed in the following sections.

Table 12 displays the internal consistency reliability indices (KR20 or Cronbach's Alpha) of both forms. As explained in the previous chapter, if the data is dichotomous, the value of Cronbach's Alpha (α) is exactly the same as KR20. Cronbach's Alpha ranges from 0 to 1, with a 0 indicating no test reliability and a fraction close to 1 indicating high test reliability. If the items in a test are correlated to each other, the value of alpha is increased. However, a high coefficient alpha does not always mean a high degree of internal consistency. This is because alpha is also affected by the length of the test. If the test length is too short, the value of alpha is reduced (Nunnally & Bernstein, 1994). Thus, to increase alpha, more related items testing the same concept should be added to the test (Tavakol & Dennick, 2011). It is also important to note that alpha is a property of the scores on a test from a specific sample of test. Therefore investigators should not rely on published alpha estimates and should measure alpha each time the test is administered.

Table 12

Internal Consistency Reliability Statistics

Form	Cronbach's Alpha	N of Items
Form 1	.521	17
Form 2	.509	17

The SPSS output computes the reliability coefficient for the test excluding one item at a time. If the reliability increases when an item is deleted, that indicates that the item is problematic and reduces test reliability instead of increasing it. Table 13 presents the list of Crohnbach's Alpha coefficients excluding one item at a time.

Table 13

Internal Consistency Reliability Statistics After Deleting One Item

ITEM Deleted	Cronbach's Alpha if Item Deleted	ITEM Deleted	Cronbach's Alpha if Item Deleted
ITEM02	.471	ITEM20	.465
ITEM03	.469	ITEM21	.456
ITEM04	.460	ITEM22	.472
ITEM05	.451	ITEM23	.515
ITEM06	.520	ITEM24	.488
ITEM07	.529	ITEM25	.474
ITEM08	.483	ITEM26	.479
ITEM09	.572	ITEM27	.543
ITEM10	.575	ITEM28	.526
ITEM11	.473	ITEM29	.475
ITEM12	.480	ITEM30	.515
ITEM13	.538	ITEM31	.475
ITEM14	.466	ITEM32	.530
ITEM15	.523	ITEM33	.469
ITEM16	.531	ITEM34	.492
ITEM17	.497	ITEM35	.475
ITEM18	.514	ITEM36	.519

It can be seen in Table 13 that ITEM 09 and ITEM 10 were problematic items in Form 1, because the Crohnbach's alpha coefficient of Form 1 increased significantly if these items were deleted. In fact, when these two items were excluded, the Crohnbach's alpha coefficient of Form 1 increased to 0.622 (see Table 14). ITEM 27 and ITEM 32

could have been considered as slightly problematic items in Form 2; when these two items were excluded the Crohnbach's alpha coefficient of Form 2 increased to 0.571 (see Table 14); a slightly increasing Crohnbach's alpha coefficient. With Cronbach's alphas that were .521 and .509 as presented in Table 4.10, we can infer that the reliability indices of the instrument forms were quite low.

Table 14

Internal Consistency Reliability Statistics After Deleting Two Items

Form	ITEM Deleted	Cronbach's Alpha	N of Items
Form 1	ITEM 09 & 10	.622	15
Form 2	ITEM 27 & 32	.571	15

It is common that an investigator reports a small value of Crohnbach's alpha as an indication of the low quality of an instrument. However, before discarding the instrument, an investigation on the homogeneity or unidimensionality of the instrument can help to understand whether the low Crohnbach's alpha is caused by the low quality of the instrument or by the heterogeneity or multidimensionality of the instrument. Internal consistency investigation is interested with the interrelatedness of the instrument items. The concept of reliability assumes that unidimensionality exists in a sample of test items and if this assumption is violated it does cause a major underestimate of reliability (Tavakol & Dennick, 2011). If an instrument has more than one concept or construct, it may not make sense to report alpha for the instrument as a whole as the larger number of questions will inevitable inflate the value of alpha. In principle therefore, alpha should be calculated for each of the concepts, in this case for each level, rather than for the entire test or scale (Nunnally & Bernstein, 1994). The implication for summative examination

containing heterogeneous, case-based questions is that alpha should be calculated for each case.

In the next sections it is shown that the instrument used in this study was in fact multidimensional. Therefore, the low Crohnbach's alpha values of each instrument form might have been caused by the multidimensionality of the instrument. Further analyses are needed in measuring the reliability of the instrument.

The relationship between Score Reliability and Validity

Score reliability refers to the state where the scores of items in the instrument are consistent in multiple measurement attempts using the same instrument. On the other hand, validity refers to the condition where the items precisely measure what they are purport to measure. So, if an instrument is a valid measurement device, then the scores of its measurement should be consistent for multiple measures. In other words, a valid instrument must produce reliable scores for many measurement attempts. An instrument that produces reliable scores for repeated measurement, however, is not necessarily valid.

As explained previously, in this study, we only measured the internal consistency reliability of students' scores on the items involved in the instrument. Since there was only small number of items per level, a high reliability coefficient of each level was less likely to exist. This indicates that the validity of the instrument is still questionable, even though several efforts to assure the validity of the instrument had been applied. Adding more items for measuring each GAISE level might give higher internal consistency reliability of the instrument that will enhance the confidence that the instrument developed in this study is accurately measure students' GAISE Levels as intended.

Confirmatory Factor Analysis Results

The Pre K-12 GAISE Framework (Franklin, et al., 2007) suggests that students develop their understanding of statistical concepts through three levels (Levels A, B, and C). Students should develop into Level A before moving to Level B and then to Level C. Using instrument items as observed variables, measurement models that define GAISE Levels as latent variables are developed and then using the relationship among Levels explained by the Pre-K-12 GAISE Framework, a structural equation model is developed.

In the following descriptions, a thorough investigation to the regression weight estimates of all models will be presented. The discussion starts by analyzing the Initial Model involving items in Form 1, called Initial F1 Model. This model was developed based on expert's opinion in the first expert survey. Two other models were developed based on Expert 1 and Expert 2 opinions. The experts' opinion were gathered during the second expert survey. The fourth model was developed by conducting exploratory modeling where several combinations of experts' opinions were tested. The best combination model was chosen as the representation model and the alignments of items and levels suggested by the model were considered as the most appropriate alignment. The same processes were applied for the SEM models of Form 2.

Results for Form 1

Figure 19 presents the structural equation diagram of the initial theoretical structural equation model that was developed based on experts' suggestion during pilot study. The model that is called the Initial F1 Model consists of three unobserved latent variables, Level A, Level B, and Level C, seventeen observed variables (ITEM02 – ITEM18) and 19 residual error associated with an observed variable (E1-E19).

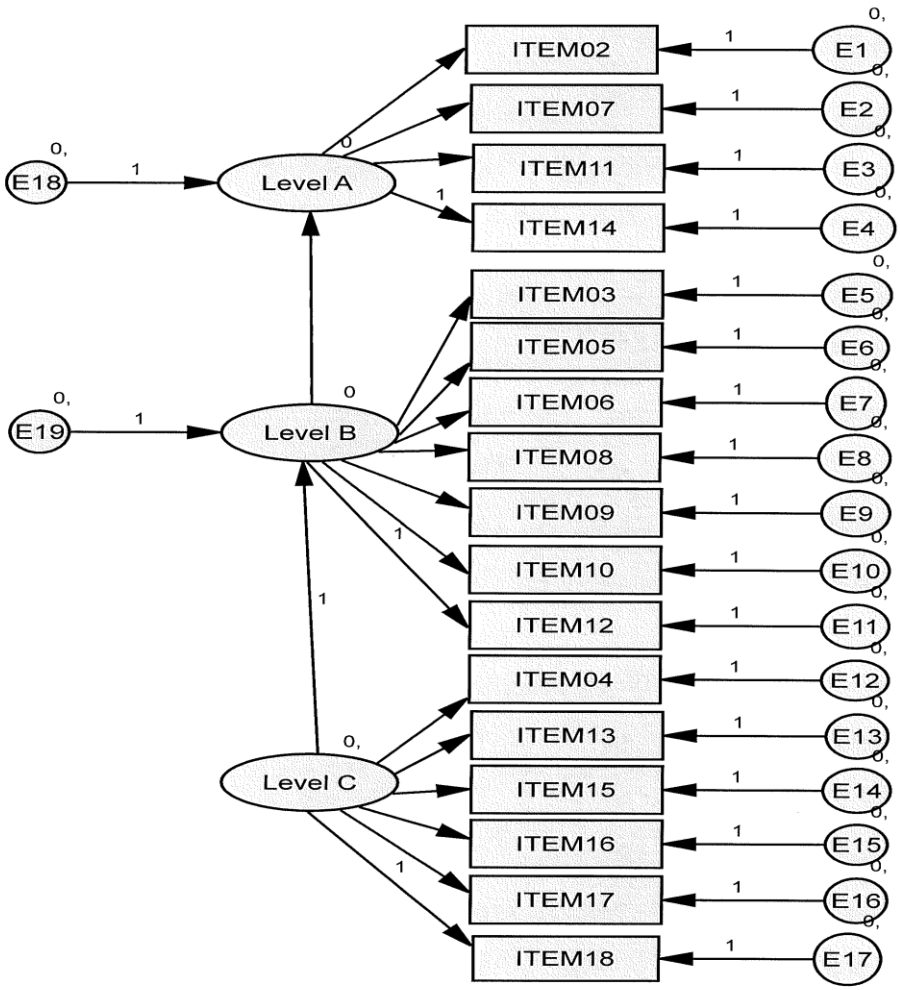


Figure 19. Initial F1 Model.

As mentioned earlier, a structural equation model consists of measurement models that describe the relation among observed variables and unobserved latent variables, and structural model that describe the relation among unobserved latent variables. The measurement models involved in this study consisted of simple standard regression models that took into account the errors of measurement. The following equation model explains the relations that are displayed by Figure 19 for Initial F1 Model.

The measurement models are specified as below:

ITEM02 = a_1Level A + E1

ITEM07 = a_2Level A + E2

ITEM11 = a_3Level A + E3

ITEM14 = 1*Level A + E4

ITEM 03 = b_1 * Level B + E5

ITEM 05 = b_2 * Level B + E6

ITEM 06 = b_3 * Level B + E7

ITEM 08 = b_4 * Level B + E8

ITEM 09 = b_5 * Level B + E9

ITEM 10 = b_6 * Level B + E10

ITEM 12 = 1* Level B + E11

ITEM 04 = c_1Level C + E12

ITEM 13 = c_2Level C + E13

ITEM 15 = c_3Level C + E14

ITEM 16 = c_4Level C + E15

ITEM 17 = c_5Level C + E16

ITEM 18 = 1*Level C + E17

The structural model for this Initial Model is specified as:

Level B = 1 * Level C + E18;

Level A = d_2Level B + E19

This model suggests that students' responses of ITEM 02, 07, 11, and 14 were affected by their Level A developmental stage in learning statistics and by the errors of measurement of the four items. Likewise, students' responses of ITEM 03, 05, 06, 08, 09, 10, and 12 were affected by their Level B developmental stage in learning statistics and by the errors of measurements of the seven items. Similarly, students' responses of ITEM 04, 13, 15, 16, 17, and 18 were affected by their Level C developmental stage in learning statistics and by the errors of measurements of the six items. Students' Level A understanding was affected by their Level B understanding. Likewise, their Level B understanding was also affected by their Level C understanding. SPSS AMOS 21 program helped us in estimating all parameters involved in this model. Table 15 displays the regression weights of the Initial F1 Model.

Table 15

Regression Weights of Initial F1 Model

	Estimate	S.E.	C.R.	P
Level B ← Level C	1.401	.783	1.791	.073
Level A ← Level B	.830	.183	4.528	***
ITEM 02 ← Level A	1.000			
ITEM 07 ← Level A	.084	.203	.412	.680
ITEM 11 ← Level A	.851	.214	3.976	***
ITEM 14 ← Level A	.931	.202	4.598	***
ITEM 03 ← Level B	1.000			
ITEM 05 ← Level B	.922	.179	5.143	***
ITEM 06 ← Level B	.289	.203	1.422	.155
ITEM 08 ← Level B	.673	.187	3.606	***
ITEM 09 ← Level B	- .379	.178	-2.132	.033
ITEM 10 ← Level B	-.889	.178	-5.003	***
ITEM 12 ← Level B	.927	.172	5.392	***
ITEM 04 ← Level C	1.000			
ITEM 13 ← Level C	.012	.193	.063	.950
ITEM 15 ← Level C	.126	.188	.671	.502
ITEM 16 ← Level C	.146	.207	.708	.479
ITEM 17 ← Level C	.711	.196	3.634	***
ITEM 18 ← Level C	.449	.237	1.892	.059

***. Correlation is significant at less than 0.001 level (2-tailed).

As can be seen in Table 15, the factor loadings of Initial F1 Model are listed as regression weights, where the columns display the parameter estimates (Column 2, Estimate), standard error (Column 3, S. E.), critical ratio (Column 4, C. R.), and p-value

(Column 5, P). As can be seen on Table 15, 6 items in Initial F1 Model have insignificant parameter estimates (ITEM 06, 07, 09, 13, 15, and 16). This might indicate whether the model is wrong or the item is not measuring what it is supposed to measure to indicate whether it should or should not be removed from the instrument. Further analysis is needed.

Table 16 shows the standardized factor loadings of Initial F1 Model that are listed as standardized regression weights. These standardized estimates will be used to estimate latent variables Level A, Level B, and Level C for Initial F1 Model.

Table 16

Standardized Regression Weights of Initial F1 Model

	Estimate		Estimate
Level B ← Level C	1.136	ITEM 09 ← Level B	- .274
Level A ← Level B	.909	ITEM 10 ← Level B	- .643
ITEM 02 ← Level A	.660	ITEM 12 ← Level B	.670
ITEM 07 ← Level A	.055	ITEM 04 ← Level C	.670
ITEM 11 ← Level A	.562	ITEM 13 ← Level C	.007
ITEM 14 ← Level A	.615	ITEM 15 ← Level C	.074
ITEM 03 ← Level B	.723	ITEM 16 ← Level C	.086
ITEM 05 ← Level B	.667	ITEM 17 ← Level C	.417
ITEM 06 ← Level B	.209	ITEM 18 ← Level C	.263
ITEM 08 ← Level B	.487		

From Table 16 it is found that standardized factor loadings of ITEM 06, 07, 09, 10, 13, 15, and 16 are lower than 0.2 that indicate these items might not aligned with the levels intended or the quality of the items are questionable. Among those items only

ITEM 10 that has significant p-value (see Table 15). Analysis on model fit index might also reveals whether this result is caused by item quality or by model chosen that is not fitted with the data. Model fit analysis will be discussed in the following subsection.

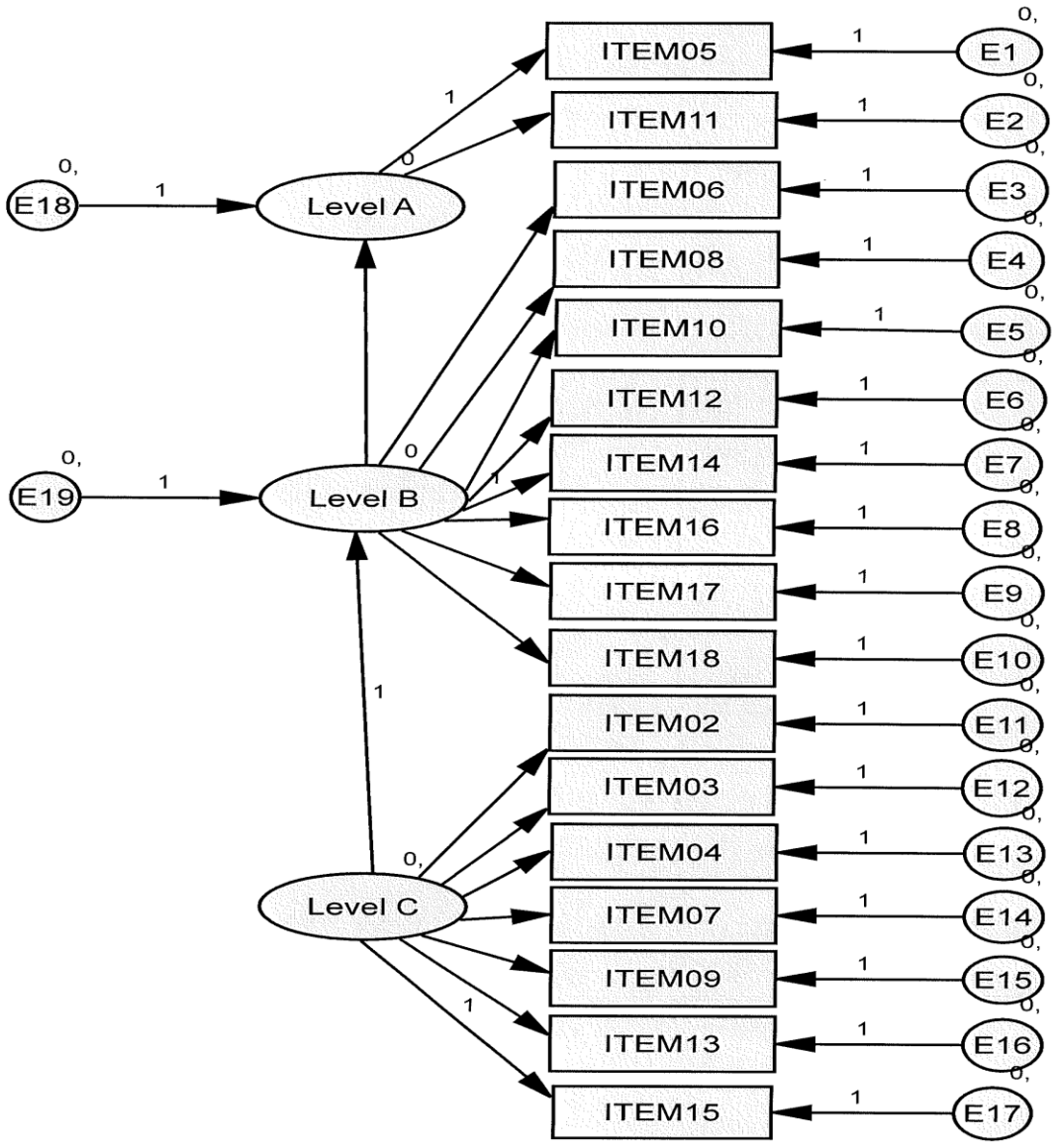


Figure 20. Expert 1F1 Model.

Table 17

Regression Weights of Expert 1F1 Model

	Estimate	S.E.	C.R.	P
Level B ← Level C	.334	.232	1.442	.149
Level A ← Level B	3.198	2.222	1.439	.150
ITEM 05 ← Level A	1.000			
ITEM 11 ← Level A	.803	.192	4.173	***
ITEM 06 ← Level B	1.000			
ITEM 08 ← Level B	2.307	1.657	1.392	.164
ITEM 10 ← Level B	-3.055	2.070	-1.476	.140
ITEM 12 ← Level B	3.170	2.196	1.443	.149
ITEM 14 ← Level B	2.773	1.943	1.427	.154
ITEM 16 ← Level B	0.453	.680	.666	.506
ITEM 17 ← Level B	2.177	1.585	1.373	.170
ITEM 18 ← Level B	1.372	1.301	1.054	.292
ITEM 02 ← Level C	1.000			
ITEM 03 ← Level C	1.150	.238	4.829	***
ITEM 04 ← Level C	1.042	.233	4.462	***
ITEM 07 ← Level C	.089	.205	.435	.663
ITEM 09 ← Level C	- .429	.210	-2.041	.041
ITEM 13 ← Level C	.023	.198	.114	.909
ITEM 15 ← Level C	.120	.192	.626	.531

***. Correlation is significant at less than 0.001 levels (2-tailed).

The second expert panel conducted during the actual study gathered suggestions from two experts (Expert 1 and Expert 2) about the alignment of each instrument item with the GAISE Level that it assesses. Expert 1 aligned ITEM 05, and 11 with Level A,

ITEM 06, 08, 10, 12, 14, 16, 17, and 18 with Level B, and ITEM 02, 03, 04, 07, 09, 13, and 15 with Level C. Figure 20 shows the structural equation model developed by Expert 1's opinion. Table 17 displays the factor loadings of Expert 1F1 Model represented by Figure 20. It was found that ITEM 06, 07, 09, 13, 16, and 18 in Expert 1F1 Model had insignificant parameter estimates. This means that the items might be problematic and need to be excluded from the instrument. Comparing to the results of Initial F1 Model (see Table 15); it was found that ITEM 06, 07, 09, 13, and 16 had insignificant estimates in both models. Comparing their difficult levels and point-biserial indices, it was found that ITEM 09 and 16 had low point-biserial indices (-.029 and .186); meanwhile ITEM 06, 07, and 13 also had slightly low point-biserial indices (.204 - .236). This indicates that ITEM 09 and ITEM 16 might not be good items in the instrument; meanwhile the appropriateness of ITEMS 06, 07, and 13 are questionable.

Compared to Initial F1 Model's regression weights (see Table 16), ITEM 06 aligned with Level B in both models, ITEM 13 and ITEM 16 aligned with Level C in both models. These indicate that ITEM 06, 13, and 16 might not have aligned with their assigned levels in both models, or maybe the three items were just not good items in the instrument. Further comparisons are needed and the discussions will be presented after all four models have been displayed. Table 18 displays the standardized factor loadings of Expert 1F1 Model. From this standardized factor loadings it was found that item 06, 07, 09, 10, 13, and 16 had low factor loadings that indicated that the items might not have matched with the level intended or their qualities were not satisfying.

Table 18

Standardized Regression Weights of Expert 1F1 Model

	Estimate		Estimate
Level B ← Level C	.999	ITEM 17 ← Level B	.456
Level A ← Level B	1.044	ITEM 18 ← Level B	.287
ITEM 05 ← Level A	.642	ITEM 02 ← Level C	.626
ITEM 11 ← Level A	.515	ITEM 03 ← Level C	.720
ITEM 06 ← Level B	.209	ITEM 04 ← Level C	.652
ITEM 08 ← Level B	.483	ITEM 07 ← Level C	.056
ITEM 10 ← Level B	-.640	ITEM 09 ← Level C	-.269
ITEM 12 ← Level B	.664	ITEM 13 ← Level C	.014
ITEM 14 ← Level B	.581	ITEM 15 ← Level C	.075
ITEM 16 ← Level B	.095		

From Table 17 and Table 18, there are strong indications that ITEM 09 and ITEM 10 were not aligned with Level B. It is more likely that these two items were bad items, since their point-biserial indices were also negative (-.029 and -.201, see Table 10). ITEM 16 also had a low point-biserial index, and since its loading factors to Level C in Initial F1 Model and to Level B in Expert 1F1 were insignificant, further analysis of this item is needed. We will carefully look at this item's factor loading in the other models. We will also carefully investigate how ITEM 06, 07, 13, 15, and 18 load to the GAISE levels in the next two structural equation models.

Figure 21 displays the structural equation model developed by Expert 2's opinion for Form 1. As can be seen in the diagram ITEM 05, 07, 08, 13 and 14 aligned with Level A. ITEM 02, 03, 06, 09, 11, 12, 16, and 17 aligned with Level B, meanwhile ITEM 04, 10, 15, and 18 aligned with Level C.

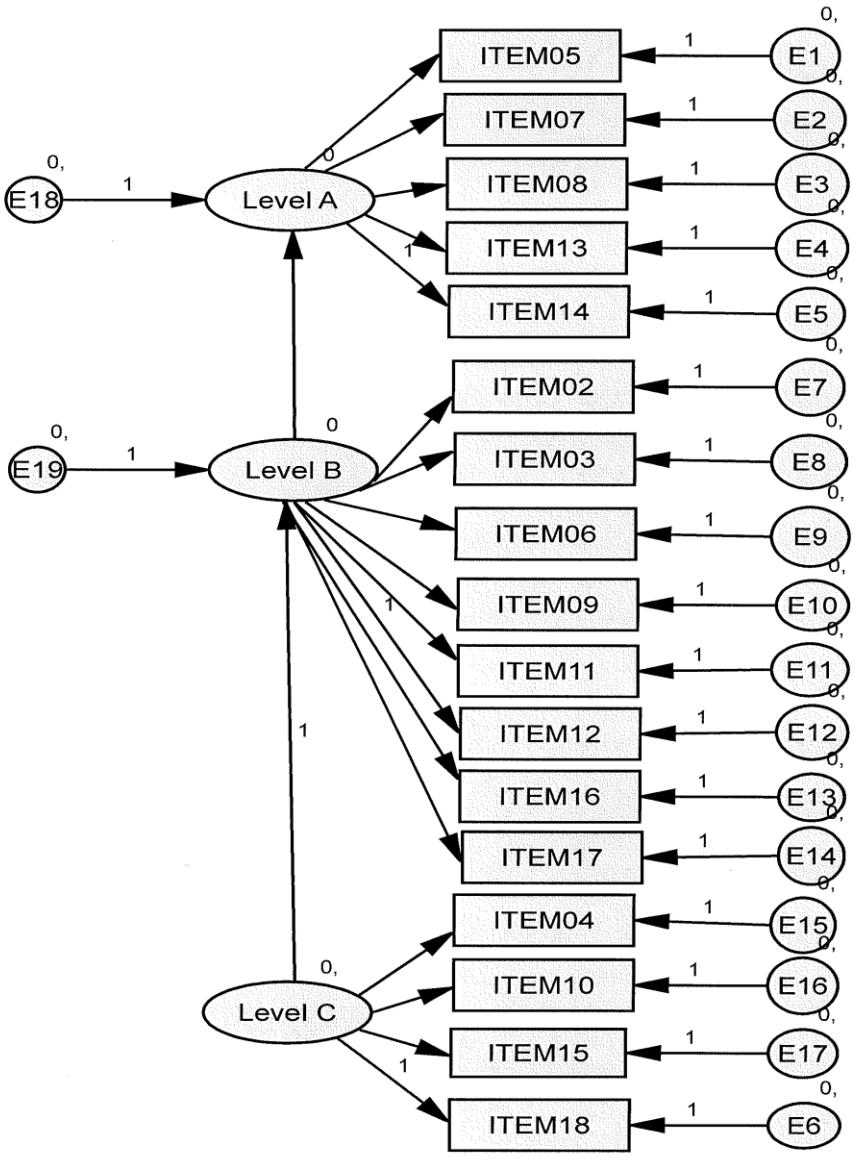


Figure 21. Expert 2F1 Model.

Table 19 displays the factor loadings and their significance levels of variables in Expert 2F1 Model (see Figure 21).

Table 19

Regression Weights of Expert 2F1 Model

	Estimate	S.E.	C.R.	P
Level B ← Level C	1.842	1.210	1.522	.128
Level A ← Level B	1.007	.241	4.170	***
ITEM 05 ← Level A	1.000			
ITEM 07 ← Level A	.086	.186	.462	.644
ITEM 08 ← Level A	.720	.199	3.622	***
ITEM 13 ← Level A	.032	.182	.175	.861
ITEM 14 ← Level A	.866	.191	4.529	***
ITEM 02 ← Level B	1.000			
ITEM 03 ← Level B	1.148	.238	4.818	***
ITEM 06 ← Level B	.338	.227	1.491	.136
ITEM 09 ← Level B	-.429	.210	-2.042	.041
ITEM 11 ← Level B	.849	.211	4.024	***
ITEM 12 ← Level B	1.065	.231	4.622	***
ITEM 16 ← Level B	.154	.206	.747	.455
ITEM 17 ← Level B	.731	.212	3.442	.001
ITEM 04 ← Level C	1.000			
ITEM 10 ← Level C	-.992	.198	-5.001	***
ITEM 15 ← Level C	.128	.188	.683	.495
ITEM 18 ← Level C	.443	.232	1.910	.056

***. Correlation is significant at less than 0.001 levels (2-tailed).

It was found that ITEM 06, 07, 09, 13, 15, and 16 in Expert 2F1 Model had insignificant factor loadings. These items were those that also had insignificant factor loadings in the previous two models. Investigating their point-biserial indices, it was found that those items had low point-biserial indices. It was good to see how these items loaded to their aligned levels in the last model in this study. If these items consistently have insignificant factor loadings to the three GAISE levels for all models, it is more likely that these items are not good for the instrument.

Comparing these items' difficulty and point-biserial indices (see Table 10), it was found that ITEM 09 and 16 had low point-biserial indices (-.029 and .186 respectively) with moderate difficulty indices (.29 and .28 respectively); meanwhile ITEM 06, 07, 13, and 15 had unsuspicious difficulty and point-biserial indices. It will be necessary to investigate ITEM 09 and 16 further. In the next discussion, a model developed based on Expert 2 suggestions will be discussed.

Table 20

Standardized Regression Weights of Expert 2F1 Model

	Estimate		Estimate
Level B ← Level C	1.425	ITEM 09 ← Level B	-.269
Level A ← Level B	.881	ITEM 11 ← Level B	.533
ITEM 05 ← Level A	.717	ITEM 12 ← Level B	.668
ITEM 07 ← Level A	.062	ITEM 16 ← Level B	.097
ITEM 08 ← Level A	.516	ITEM 17 ← Level B	.458
ITEM 13 ← Level A	.023	ITEM 04 ← Level C	.485
ITEM 14 ← Level A	.621	ITEM 10 ← Level C	-.482
ITEM 02 ← Level B	.627	ITEM 15 ← Level C	.062
ITEM 03 ← Level B	.720	ITEM 18 ← Level C	.215
ITEM 06 ← Level B	.212		

Table 20 displays the standardized regression weights of Expert 2F1 Model. ITEM 07 and 13 had low factor loadings to Level A; ITEM 06, 09, and 16 also had low factor loadings to Level B; meanwhile ITEM 10 and 15 had low factor loadings to Level C. Since ITEM 06, 07, 13, and 15 had reasonable difficulty and point-biserial indices, this low factor loadings might be signs of wrong model. It will be necessary to investigate the last model that is based on a combination of both experts' suggestions.

The last model for Form 1 is developed though exploratory modeling by combining both experts' suggestion. This Combination F1 Model was the best combination model developed after several modeling attempts. Figure 22 displays the structural equation model diagram of Combination F1 Model.

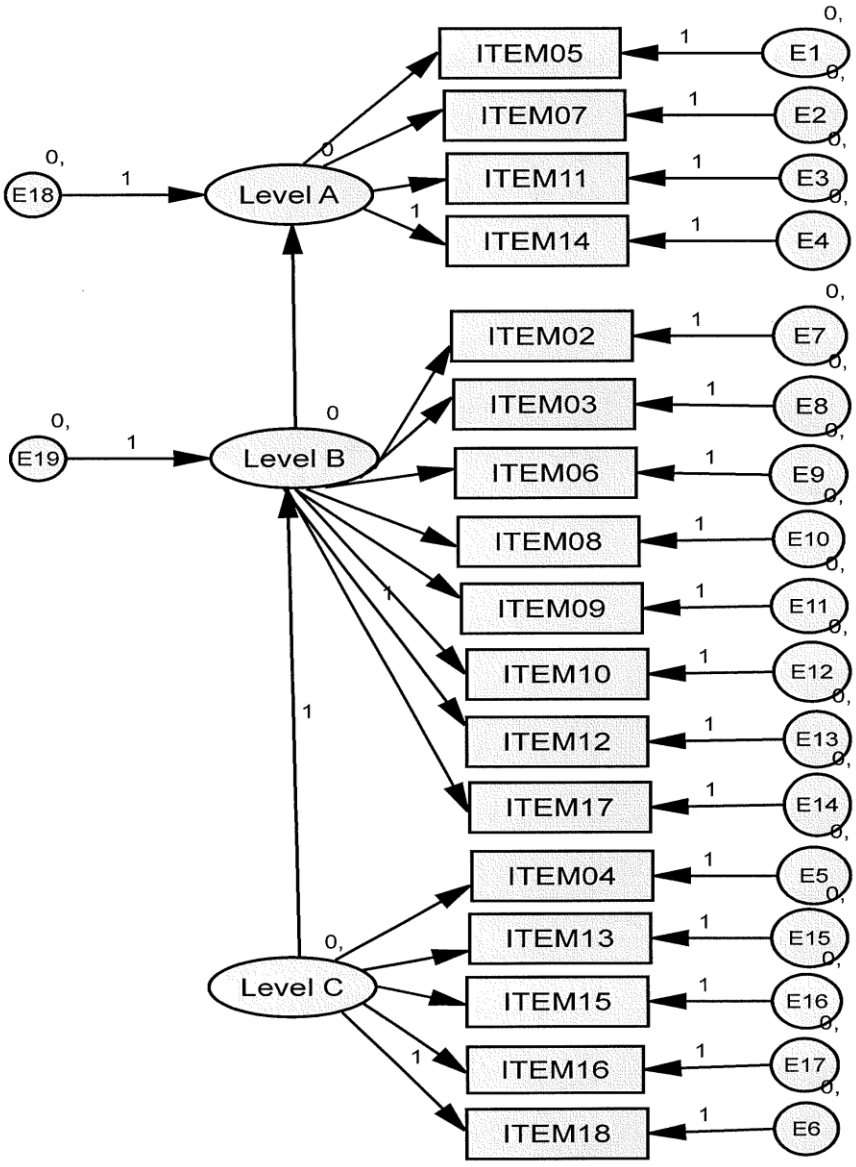


Figure 22. Combination F1 Model.

Table 21 displays the regression weights of Combination F1 Model. It was found that ITEM 06, 07, 09, 13, 15, and 16 in Combination F1 Model also had insignificant parameter estimates.

Table 21

Regression Weights of Combination F1 Model

	Estimate	S.E.	C.R.	P
Level B ← Level C	.954	.746	1.279	***
Level A ← Level B	.967	.228	4.242	***
ITEM 05 ← Level A	1.000			
ITEM 07 ← Level A	.079	.185	.427	.670
ITEM 11 ← Level A	.794	.194	4.083	***
ITEM 14 ← Level A	.873	.195	4.480	***
ITEM 02 ← Level B	1.000			
ITEM 03 ← Level B	1.151	.241	4.782	***
ITEM 06 ← Level B	.333	.227	1.463	.144
ITEM 08 ← Level B	.772	.236	3.269	.001
ITEM 09 ← Level B	-.438	.210	-2.088	.037
ITEM 10 ← Level B	-1.025	.233	4.393	***
ITEM 12 ← Level B	1.070	.232	4.605	***
ITEM 17 ← Level B	.728	.212	3.438	.001
ITEM 04 ← Level C	1.000			
ITEM 13 ← Level C	.015	.190	.082	.935
ITEM 15 ← Level C	.115	.186	.618	.537
ITEM 16 ← Level C	.148	.203	.728	.467
ITEM 18 ← Level C	.443	.237	1.866	.062

***. Correlation is significant at less than 0.001 levels (2-tailed).

Table 22 presents the standardized factor loadings of each relation among variables. It was found that ITEM 06, 07, 09, 10, 13, 15, and 16 had low factor loadings. It is clear form Table 15 - 22 that Item 06, 07, 09, 10, 13, 15, and 16 had insignificant parameter estimates and/or low standardized factor loadings in all models, even though

they were assigned to different levels in the models. Further investigation on these items should be conducted to conclude whether these items should be removed from the instrument or whether the results were caused by small sample size (140).

Table 22

Standardized Regression Weights of Combination F1 Model

	Estimate		Estimate
Level B ← Level C	.995	ITEM 09 ← Level B	-.278
Level A ← Level B	.834	ITEM 10 ← Level B	-.651
ITEM 05 ← Level A	.736	ITEM 12 ← Level B	.680
ITEM 07 ← Level A	.058	ITEM 17 ← Level B	.462
ITEM 11 ← Level A	.584	ITEM 04 ← Level B	.662
ITEM 14 ← Level A	.643	ITEM 13 ← Level C	.010
ITEM 02 ← Level B	.635	ITEM 15 ← Level C	.076
ITEM 03 ← Level B	.731	ITEM 16 ← Level C	.098
ITEM 06 ← Level B	.211	ITEM 18 ← Level C	.293
ITEM 08 ← Level B	.490		

All models showed that the relation between latent variables Level B and Level C as well as Level A and Level B were significant. All standardized factor loadings between two latent variables were larger than 0.6. This indicates that the theory that students develop their understanding starting from Level A then to Level B and finally reach Level C is confirmed by all four models.

Results for Form 2

Now, we consider the models developed for ITEMS 20–36 from Form 2. Figure 23 displays the first model, called Initial F2 Model.

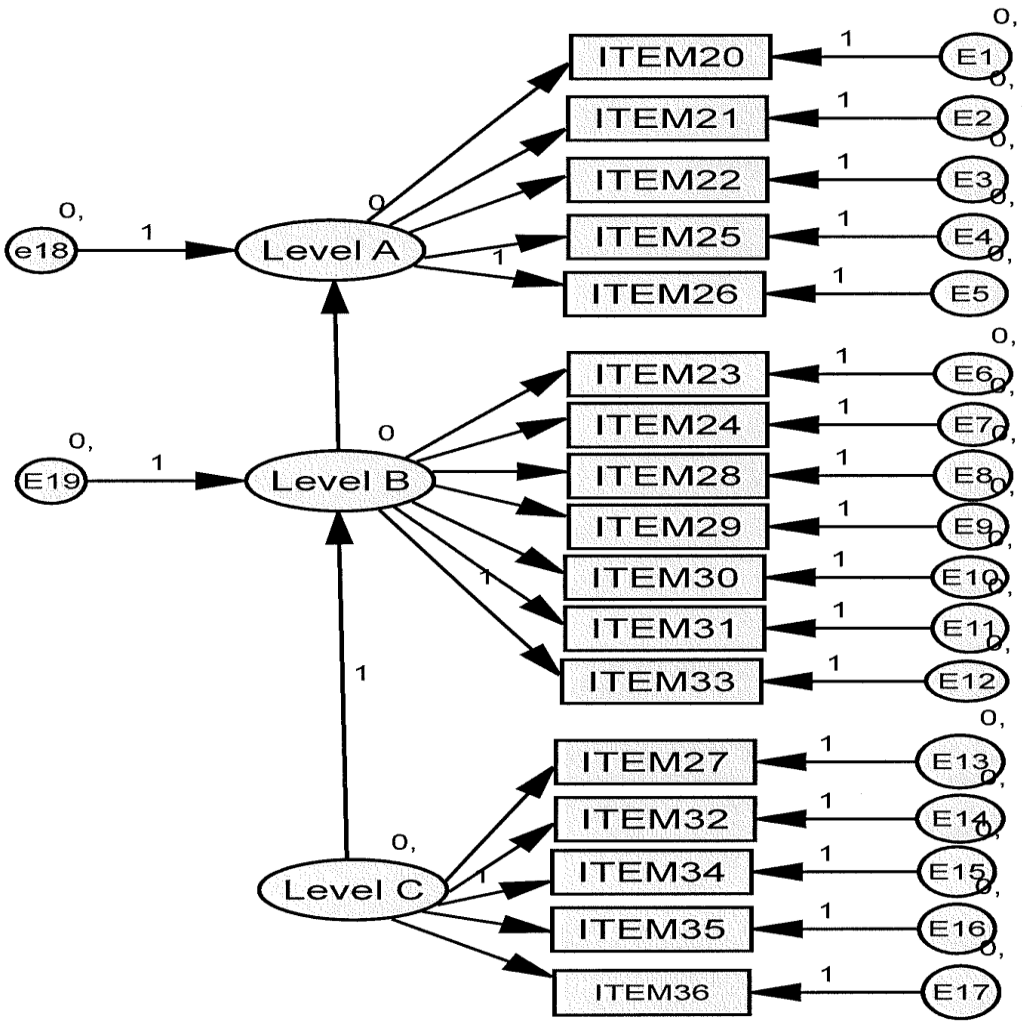


Figure 23. Initial F2 Model.

Similar to the previous tables for Form 1, in Table 23, the factor loadings of the Initial Model of Form 2 are listed as regression weights, where the columns display the parameter estimates (Column 2, Estimate), standard error (Column 3, S.E.), critical ratio

(Column 4, C.R), and p-value (Column 5, P). It is found that ITEMS 23, 28, and 30 have insignificant factor loadings to Level B.

Table 23

Regression Weights of Initial F2 Model

	Estimate	S.E.	C.R.	P
Level B ← Level C	1.000			
Level A ← Level B	.771	.154	5.009	***
ITEM 20 ← Level A	1.458	.249	5.866	***
ITEM 21 ← Level A	1.631	.270	6.036	***
ITEM 22 ← Level A	1.392	.239	5.822	***
ITEM 25 ← Level A	1.017	.200	5.075	***
ITEM 26 ← Level A	1.000			
ITEM 23 ← Level B	-.077	.096	-.802	.422
ITEM 24 ← Level B	.552	.128	4.314	***
ITEM 28 ← Level B	.010	.111	.093	.926
ITEM 29 ← Level B	.692	.146	4.756	***
ITEM 30 ← Level B	.048	.108	.444	.657
ITEM 31 ← Level B	1.040	.176	5.895	***
ITEM 33 ← Level B	1.000			
ITEM 27 ← Level B	-.475	.191	-2.484	.013
ITEM 32 ← Level C	-.145	.166	-.874	.382
ITEM 34 ← Level C	1.000			
ITEM 35 ← Level C	1.277	.277	4.612	***
ITEM 36 ← Level C	.014	.152	.092	.927

***. Correlation is significant at less than 0.001 levels (2-tailed).

ITEMS 32 and 36 also have insignificant factor loadings to Level C. Compared to their point-biserial indices (see Table 11) it was found that these items also had low point-biserial indices. This indicates that these items are problematic and need to be investigated further to decide whether they should be removed or kept in the instrument.

Table 24 shows the standardized factor loadings of Initial F2 Model that are listed as Standardized Regression Weights. It was found that ITEM 23, 27, 28, 30, 32, and 36 had low factor loading to the assigned GAISE levels.

Table 24

Standardized Regression Weights of Initial F2 Model

	Estimate		Estimate
Level B ← Level C	.829	ITEM 29 ← Level B	.278
Level A ← Level B	.875	ITEM 30 ← Level B	.023
ITEM 20 ← Level A	.494	ITEM 31 ← Level B	.384
ITEM 21 ← Level A	.551	ITEM 33 ← Level B	.422
ITEM 22 ← Level A	.483	ITEM 27 ← Level B	-.150
ITEM 25 ← Level A	.349	ITEM 32 ← Level C	-.049
ITEM 26 ← Level A	.334	ITEM 34 ← Level C	.317
ITEM 23 ← Level B	-.041	ITEM 35 ← Level C	.410
ITEM 24 ← Level B	.245	ITEM 36 ← Level C	.005
ITEM 28 ← Level B	.005		

As mentioned previously, these items also had low point-biserial indices. Investigating the problematic items more carefully, it was found that ITEMS 23, 28, 30, and 36 also had low difficulty indices (.15 – .24, see Table 11). It is possible that their low point-biserial indices were due to a guessing factor. On the other hand, ITEMS 27

and 32 were items with medium difficulty levels (.47 and .31, see Table 11) and both assessed students' understanding of interpreting data presented by histograms. This information is interesting since understanding histograms is one of the difficult concepts in statistics, even for students in college level introductory statistics courses (Meletiou & Lee, 2002). If 47% and 31% middle and high school students answered these two items correctly but the point-biserial of these items was low, it is also possible that this result was affected by guessing factors.

Careful analyses on these items should be conducted to ensure whether the items should be removed from the instrument, or should be modified, or whether an administration of the instrument to high school participants is necessary before discarding the items from the instrument. Since only around 10% of participants who took Form 2 were high school students, it was more likely that not many students who had progressed to Level C took the survey. Therefore, there was a higher probability that many students guessed the answer correctly that reflected in the moderate difficulty levels of ITEMS 27 and 32, but resulted in the low point-biserial indices for these items. It is interesting to see the result of analysis of the model developed following Expert 1's opinions. Figure 24 displays the structural equation model, called Expert 1F2 Model.

As mentioned before, two experts gave suggestions on the alignment of items into GAISE levels. Figure 4.12 presents the structural equation model developed based on suggestion of Expert 1 that is called Expert 1F2 Model. From Figure 4.5 and Figure 4.6, it can be seen that ITEM 32 were assigned as Level A item in Initial model, meanwhile in Expert 1F2 Model it was assigned as a Level C item. It is interesting to investigate how this item loads to the levels assigned in these two models. Considering ITEM 27, in the

previous model ITEM 27 was assigned to Level B; meanwhile in the Expert 1F2 Model, it was assigned to Level C.

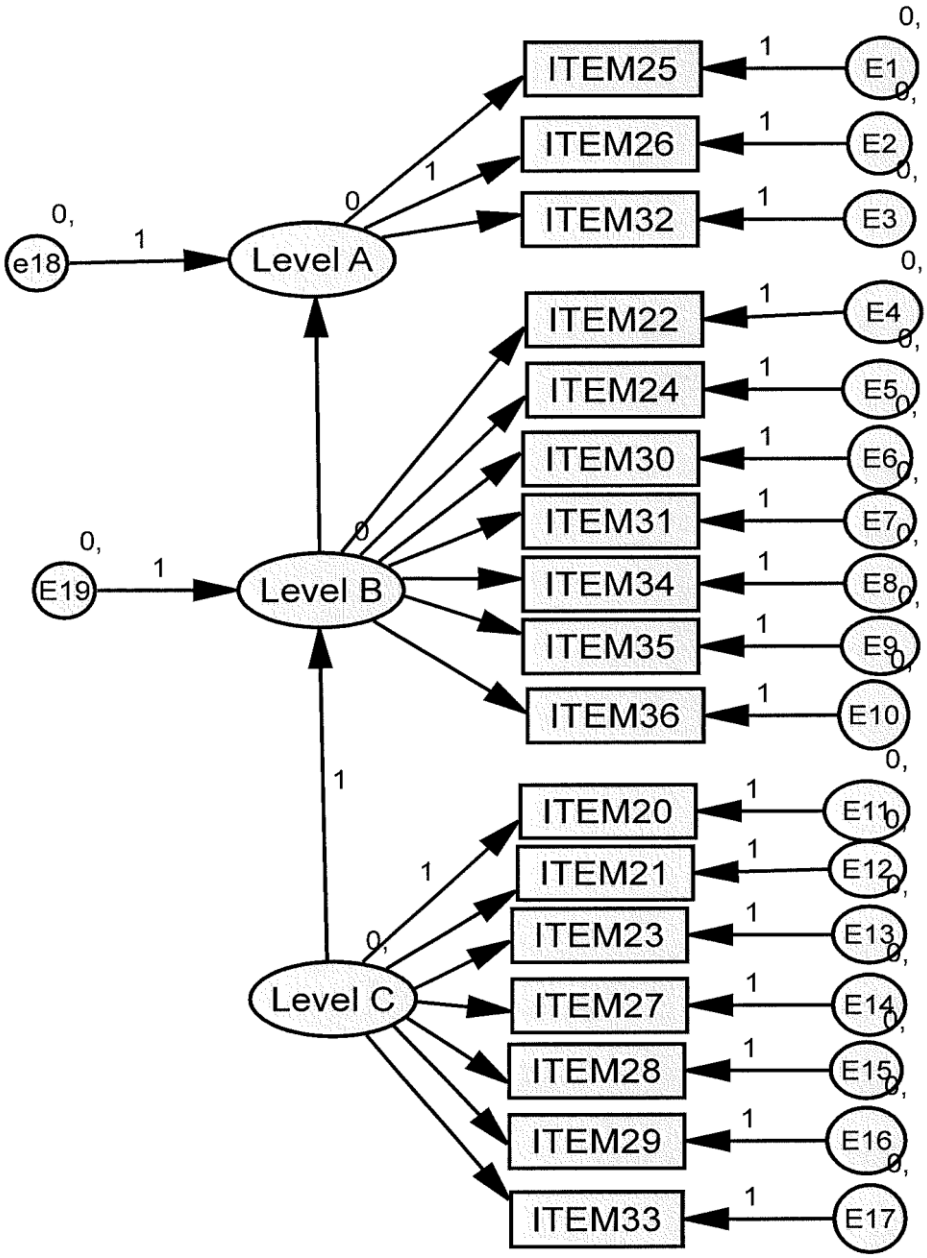


Figure 24. Expert 1F2 Model.

Table 25 displays the regression weights of Expert 1F2 Model. It can be seen in the table that factor loading of ITEM 27 into Level C is quite significant.

Table 25

Regression Weights of Expert 1F2

	Estimate	S.E.	C.R.	P
Level B ← Level C	.983	.124	7.901	***
Level A ← Level B	.747	.106	7.063	***
ITEM 25 ← Level A	1.000			
ITEM 26 ← Level A	.909	.161	5.636	***
ITEM 32 ← Level A	-.122	.136	-.897	.370
ITEM 22 ← Level B	1.000			
ITEM 24 ← Level B	.538	.124	4.351	***
ITEM 30 ← Level B	.049	.121	.403	.687
ITEM 31 ← Level B	.802	.109	7.344	***
ITEM 34 ← Level B	.519	.101	5.132	***
ITEM 35 ← Level B	.686	.108	6.339	***
ITEM 36 ← Level B	.025	.116	.215	.830
ITEM 20 ← Level C	1.000			
ITEM 21 ← Level C	1.094	.121	9.061	***
ITEM 23 ← Level C	-.121	.125	-.970	.332
ITEM 27 ← Level C	-.278	.100	-2.773	.006
ITEM 28 ← Level C	-.016	.105	-.154	.878
ITEM 29 ← Level C	.567	.107	5.295	***
ITEM 33 ← Level C	.860	.122	7.071	***

**. Correlation is significant at less than 0.001 levels (2-tailed).

It was also found that ITEM 23, 28, 30, 32, and 36 in Expert 1F2 Model had insignificant estimates, the same items that had insignificant estimates in the Initial F2 Model. It seems ITEM 32 that was assigned as a Level A item in Initial F2 Model and as a Level C item in Expert 1F2 Model had insignificant factor loading no matter what Level it was assigned to. However, further analyses should been conducted, especially considering the possibility that many the participants had not been progressed to Level C yet.

Table 26 showed standardized regression weights of Expert 1F2 Model.

Table 26

Standardized Regression Weights of Expert 1F2 Model

	Estimate		Estimate
Level B ← Level C	.981	ITEM 35 ← Level B	.418
Level A ← Level B	.860	ITEM 36 ← Level B	.015
ITEM 25 ← Level A	.529	ITEM 20 ← Level C	.609
ITEM 26 ← Level A	.480	ITEM 21 ← Level C	.666
ITEM 32 ← Level A	-.065	ITEM 23 ← Level C	-.074
ITEM 22 ← Level B	.609	ITEM 27 ← Level C	-.169
ITEM 24 ← Level B	.328	ITEM 28 ← Level C	-.010
ITEM 30 ← Level B	.030	ITEM 29 ← Level C	.345
ITEM 31 ← Level B	.489	ITEM 33 ← Level C	.523
ITEM 34 ← Level B	.316		

Items that had small regression weight estimates are ITEMS 23, 27, 28, 30, 32, and 36, the same items that had small regression weights in Initial F2 Model. It was interesting to see whether the model develop by Expert 2's opinions gave similar results or not.

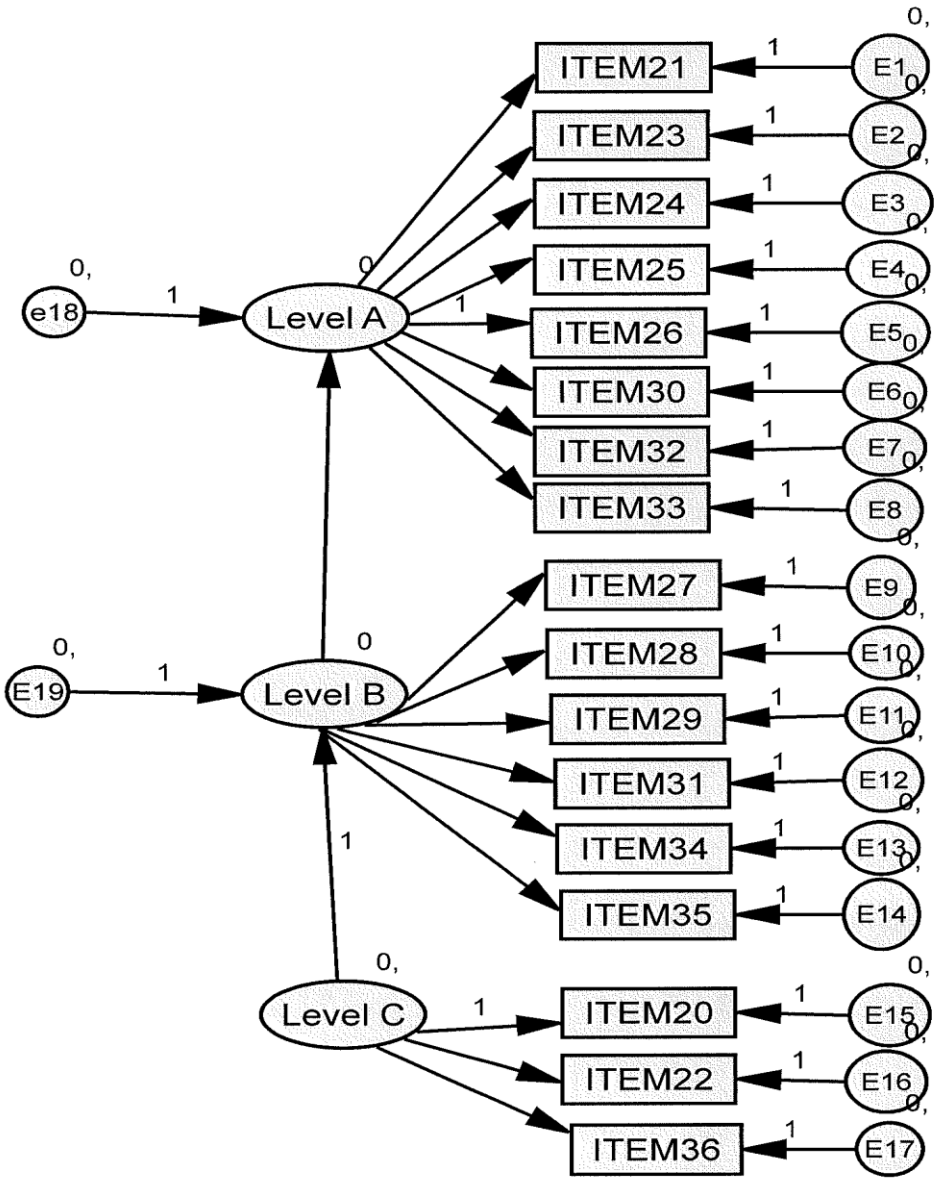


Figure 25. Expert 2F2 Model.

Figure 25 shows the structural equation model developed from Expert 2 opinions. Similar with Expert 1, Expert 2 also assigned ITEM 32 to Level A but assigned ITEM 27 to Level B. Table 27 displays the regression weights of Expert 2F2 Model.

Table 27

Regression Weights of Expert 2F2 Model

	Estimate	S.E.	C.R.	P
Level B ← Level C	-.167	.064	-2.620	.009
Level A ← Level B	-4.014	1.443	-2.782	.005
ITEM 21 ← Level A	1.000			
ITEM 23 ← Level A	-.091	.115	-.791	.429
ITEM 24 ← Level A	.499	.110	4.530	***
ITEM 25 ← Level A	.709	.091	7.822	***
ITEM 26 ← Level A	.647	.095	6.786	***
ITEM 30 ← Level A	.047	.110	.433	.665
ITEM 32 ← Level A	-.090	.099	-.905	.365
ITEM 33 ← Level A	.783	.107	7.335	***
ITEM 27 ← Level B	1.000			
ITEM 28 ← Level B	.063	.382	.166	.869
ITEM 29 ← Level B	-2.067	.824	-2.507	.012
ITEM 31 ← Level B	-2.928	1.070	-2.737	.006
ITEM 34 ← Level B	-1.906	.736	-2.590	.010
ITEM 35 ← Level B	-2.481	.914	-2.716	.007
ITEM 20 ← Level C	1.000			
ITEM 22 ← Level C	.998	.133	7.513	***
ITEM 36 ← Level C	.013	.105	.128	.898

***. Correlation is significant at less than 0.001 level (2-tailed).

It was also found that ITEM 23, 28, 30, 32, and 36 in Expert 2F2 Model had insignificant estimates, the same items that had insignificant estimates in the Initial F2 Model and Expert 1F2 Model. These continuous patterns suggested that the problems were not related to the assignments of items into Levels; for example ITEM 23 was

assigned into Level A (Expert 2F2 Model), Level B (Initial F2 Model), and Level C (Expert 1F2 Model), but all factor loadings of ITEM 23 to each level were not significant (see Table 23, 25, and 27). ITEM 23 assessed students' understanding of theoretical probability of the outcomes of throwing three dice simultaneously. It had a difficulty level of .15 that was very low that suggested that this concept was difficult for the students. This item also had point-biserial .147 that indicated this item's power in discriminating students' ability was low. This problematic item also has a low standardized regression weight (-.036) that can be seen in Table 28 below.

Table 28

Standardized Regression Weights of Expert 2F2 Model

	Estimate		Estimate
Level B ← Level C	-.709	ITEM 27 ← Level B	.170
Level A ← Level B	-1.013	ITEM 28 ← Level B	.011
ITEM 21 ← Level A	.675	ITEM 29 ← Level B	-.352
ITEM 23 ← Level A	-.061	ITEM 31 ← Level B	-.499
ITEM 24 ← Level A	.337	ITEM 34 ← Level B	-.325
ITEM 25 ← Level A	.479	ITEM 35 ← Level B	-.423
ITEM 26 ← Level A	.437	ITEM 20 ← Level C	.725
ITEM 30 ← Level A	.032	ITEM 22 ← Level C	.723
ITEM 32 ← Level A	-.061	ITEM 36 ← Level C	.010
ITEM 33 ← Level A	.529		

It was found that ITEMS 23, 27, 28, 30, 32, and 36 had low standardized regression weights. These items were the same items with low standardized regression weights with the previous models. Indication that these items should have been removed from the instrument is very strong. Likewise in Form 1, all models for Form 2 showed

that the relation between latent variables Level B and Level C as well as Level A and Level B were significant.

Now, it is interesting to combine experts' opinions and develop a structural equation model that will fit the data well. Figure 26 displays the structural equation model developed by combining experts' opinion.

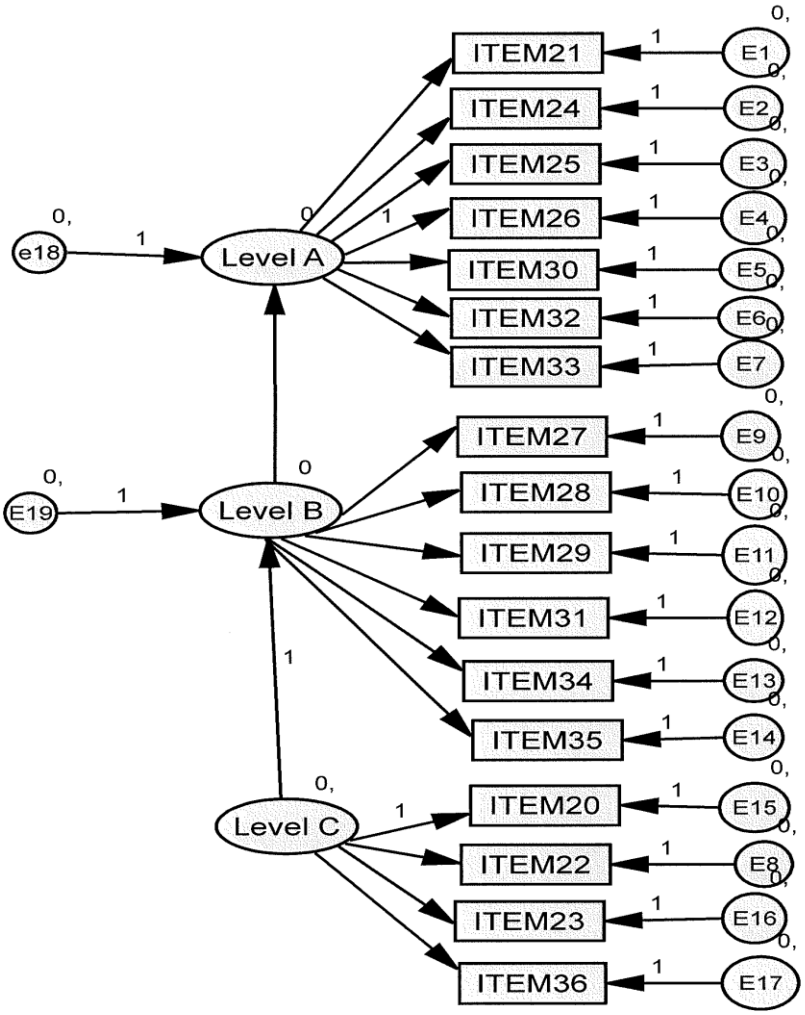


Figure 26. Combination F2 Model.

Table 29 presents regressions weight estimates of Combination F2 Model that is displayed by Figure 26.

Table 29

Regression Weights of Combination F2 Model

	Estimate	S.E.	C.R.	P
Level B ← Level C	-.160	.061	-2.619	.009
Level A ← Level B	-3.991	1.433	-2.785	.005
ITEM 21 ← Level A	1.000			
ITEM 24 ← Level A	.498	.110	4.514	***
ITEM 25 ← Level A	.708	.091	7.797	***
ITEM 26 ← Level A	.648	.096	6.784	***
ITEM 30 ← Level A	.047	.109	.425	.671
ITEM 32 ← Level A	-.091	.099	-.923	.356
ITEM 33 ← Level A	.782	.107	7.320	***
ITEM 27 ← Level B	1.000			
ITEM 28 ← Level B	.066	.381	.172	.863
ITEM 29 ← Level B	-2.054	.819	-2.510	.012
ITEM 31 ← Level B	-2.920	1.064	-2.745	.006
ITEM 34 ← Level B	-1.905	.733	-2.598	.009
ITEM 35 ← Level B	-2.476	.909	-2.722	.006
ITEM 20 ← Level C	1.000			
ITEM 22 ← Level C	1.002	.135	7.436	***
ITEM 23 ← Level C	-.173	.112	-1.544	.123
ITEM 36 ← Level C	.007	.104	.064	.949

***. Correlation is significant at less than 0.001 level (2-tailed).

Table 29 shows that ITEM 28, 30, 32, and 36 had insignificant factor loadings. Those items were also the ones that were problematic in the other models. This is

convincing evidence that these items might need to be removed from the instruments because of their poor qualities.

Investigating the standardized regression weight estimates of Combination F2 Model, seen in Table 30, it was found that ITEM 23, 27, 28, 30, 32, and 36 had low standardized factor loadings. Indications that these six items were problematic are clear, suggesting the removal of the items from the instrument and analysis of the reduced item instrument is needed.

Table 30

Standardized Regression Weights of Combination F2 Model

	Estimate		Estimate
Level B ← Level C	-.684	ITEM 28 ← Level B	.011
Level A ← Level B	-1.009	ITEM 29 ← Level B	-.352
ITEM 21 ← Level A	.678	ITEM 31 ← Level B	-.500
ITEM 24 ← Level A	.337	ITEM 34 ← Level B	-.326
ITEM 25 ← Level A	.480	ITEM 35 ← Level B	-.424
ITEM 26 ← Level A	.032	ITEM 20 ← Level C	.735
ITEM 30 ← Level A	.022	ITEM 22 ← Level C	.736
ITEM 32 ← Level A	-.062	ITEM 23 ← Level C	-.127
ITEM 33 ← Level A	.530	ITEM 36 ← Level C	.005
ITEM 27 ← Level B	.171		

The following analyses of reduced item instrument of Combination F1 Model and Combination F2 Model were conducted based on the previous analyses that showed ITEMS 06, 07, 09, 10, 13, 15, and 16 were problematic in Form 1 as well as ITEMS 23, 27, 28, 30, 32 and 36 that were problematic in Form 2. Figure 27 presents the reduced item Combination F1 Model.

Reduced combination models

Figure 27 presents the reduced item Combination F1 Model, that is the model developed from Combination F1 model by deleting problematic ITEMS 06, 07, 09, 10, 13, 15, and 16.

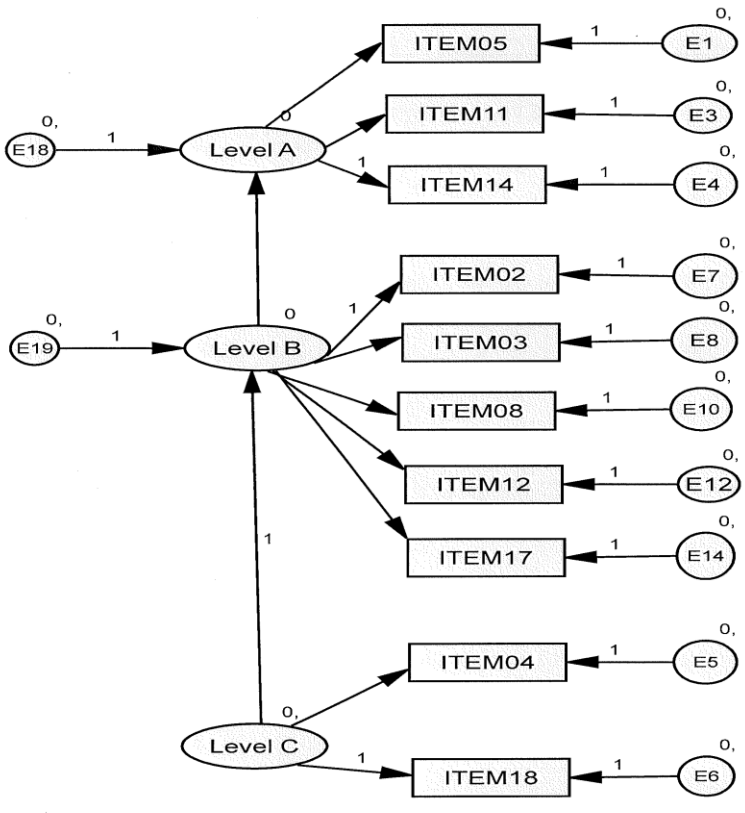


Figure 27. Reduced combination F1 Model.

The Reduced Combination F1 Model is not considered good from its psychometric properties due to its small number of items to measure three constructs. Especially, only two items are available to measure students' level C mastery. Kim and Mueller (1981) suggested that it is desirable to have three or more items per construct to ensure better measurement properties for each construct.

In Table 31, it can be seen that all parameter estimates to be reasonable and statistically significant at 0.05 levels, meanwhile the standard errors and critical ratios are also to be in good order.

Table 31

Regression Weights of Reduced Combination F1 Model

			Estimate	S.E	C.R.	P
Level_B	←	Level_C	.889	.684	1.300	.194
Level_A	←	Level_B	1.064	.250	4.254	***
ITEM05	←	Level_A	1.000			
ITEM11	←	Level_A	.775	.181	4.283	***
ITEM14	←	Level_A	.834	.178	4.676	***
ITEM02	←	Level_B	1.000			
ITEM03	←	Level_B	1.240	.266	4.664	***
ITEM08	←	Level_B	.741	.245	3.028	.002
ITEM12	←	Level_B	1.002	.243	3.028	***
ITEM17	←	Level_B	.688	.220	3.135	.002
ITEM04	←	Level_C	1.000			
ITEM18	←	Level_C	.442	.234	1.888	.059

***. Correlation is significant at less than 0.001 level (2-tailed).

Table 32 shows the standardized regression weights of Reduced Combination F1 Model. It can be seen that all regression weight estimates are reasonable which indicates that this model is good from SEM perspectives. It has been shown that the reduced item combination F1 Model gives a very promising result.

Table 32

Standardized Regression Weights of Reduced Combination F1 Model

			Estimate				Estimate
Level_B	←	Level_C	.989	ITEM03	←	Level_B	.766
Level_A	←	Level_B	.874	ITEM08	←	Level_B	.458
ITEM05	←	Level_A	.752	ITEM12	←	Level_B	.619
ITEM11	←	Level_A	.583	ITEM17	←	Level_B	.425
ITEM14	←	Level_A	.627	ITEM04	←	Level_C	.688
ITEM02	←	Level_B	.618	ITEM18	←	Level_C	.304

Mplus Version 7 also provides item characteristic curve (ICC) of all items in each level. Figure 28 shows the ICCs of items in Level A for Reduced Combination F1 Model.

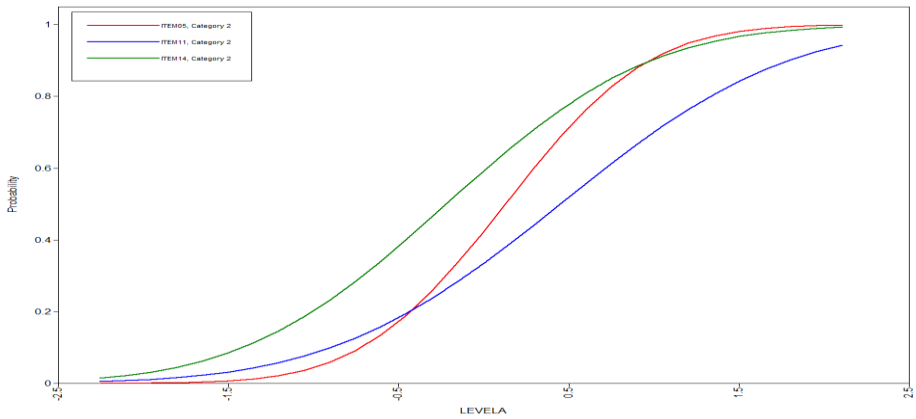


Figure 28. Item characteristic curves of Level A items in reduced combination F1 Model.

ICCs for ITEMS 05, ITEM 11, and ITEM 14 are presented in red, blue, and green respectively. Comparing the three curves, it is clear that ITEM 11 has the lowest probabilities to be answered correctly among the three items; meanwhile ITEM 14 has slightly higher probabilities than the other two items. Comparing the percentage correct responses of the three items represented in the previous section, it was found that ITEM 14 has the largest correct responses (57%), followed by ITEM 05 (45%) and ITEM 11 (36%). This fact is consistent with the phenomenon showed by the ICCs of the three items.

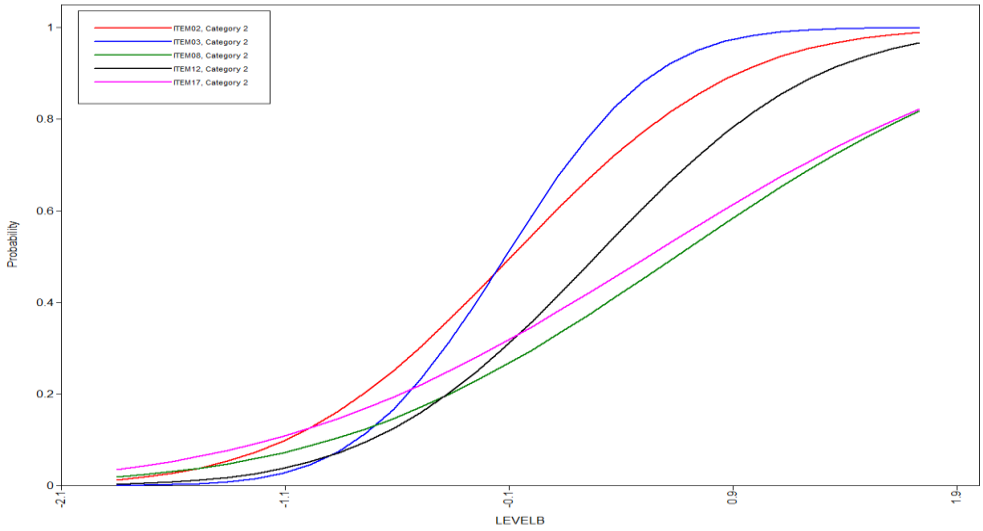


Figure 29. Item characteristic curves of Level B items in reduced combination F1 Model.

Comparing the slopes of the three curves displayed in Figure 29, it is clear that ITEM 05 has the highest slope among the three curves and ITEM 14 has higher slope than ITEM 11. Connecting the slopes of the ICCs with items' point-biserial, it was found that ITEM 05 had the largest point-biserial (.580) among the three items, followed by ITEM 14 (.521) and ITEM 11 (.484). In view of that, the connection between CTT item analysis results and SEM-based exploratory/confirmatory factor analysis results can be

established by analyzing the ICCs of all items for each level in each model. ICCs for Level B and C for Reduced Combination F1 Model are presented in Figure 29 and Figure 30 respectively.

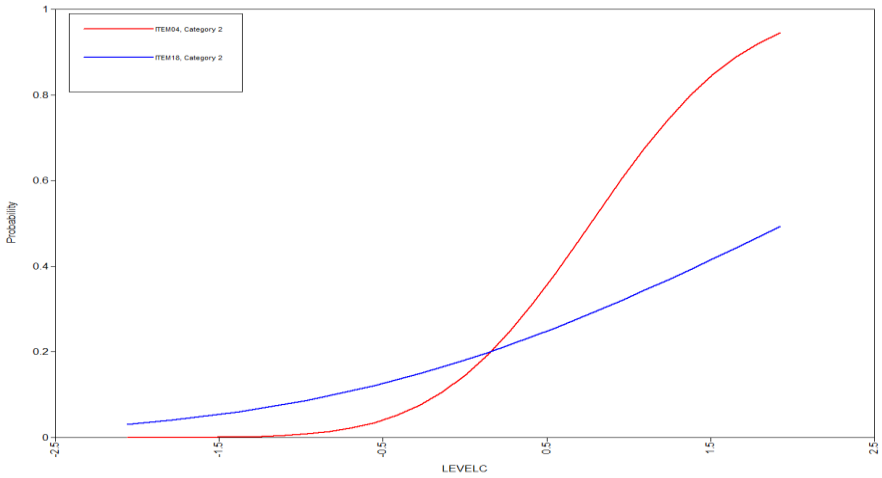


Figure 30. Item characteristic curves of Level C items in reduced combination F1 Model.

Now, by deleting the problematic items in Form 2, that are ITEMS 23, 27, 28, 30, 32 and 36 and use the Combination F2 Model as the base model, the Reduced Item Combination F2 Model was developed. Figure 31 shows the diagram of the Reduced Item Combination F2 Model. This model has its own flaws since it only has two items for Level C construct that is considered an undesired measurement model (Kim & Mueller, 1981). This flaw can be fixed by adding more items to the measurement model; however, the new SEM model must be tested for validity and items quality.

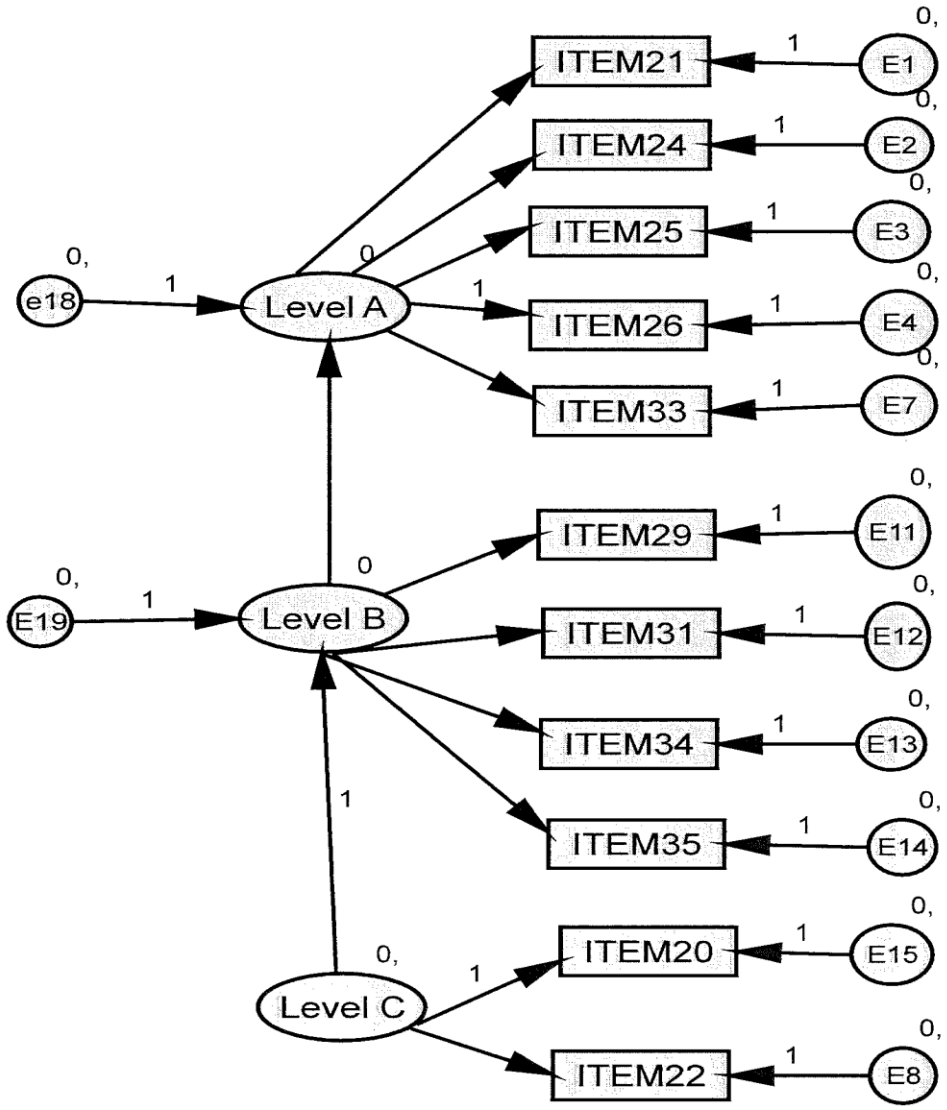


Figure 31. Reduced combination F2 Model.

In Table 33 it can be seen that all parameter estimates to be reasonable and statistically significant, meanwhile the standard errors and critical ratios are also in good order.

Table 33

Regression Weights of Reduced Combination F2 Model

			Estimate	S.E.	C.R.	P
Level B	←	Level C	.356	.076	4.671	***
Level A	←	Level B	1.834	.358	5.119	***
ITEM21	←	Level A	1.000			
ITEM24	←	Level A	.504	.112	4.512	***
ITEM25	←	Level A	.715	.092	7.750	***
ITEM26	←	Level A	.656	.097	6.790	***
ITEM33	←	Level A	.785	.109	7.231	***
ITEM29	←	Level B	1.000			
ITEM31	←	Level B	1.351	.264	5.118	***
ITEM34	←	Level B	.879	.207	4.249	***
ITEM35	←	Level B	1.167	.250	4.665	***
ITEM20	←	Level C	1.000			
ITEM22	←	Level C	.981	.133	7.396	***

***. Correlation is significant at less than 0.001 levels (2-tailed).

Table 34 shows the standardized regression weights of Reduced Combination F1 Model. It can be seen that all regression weight estimates are reasonable (.323 – 1.001) which indicates that this model is good.

Table 34

Standardized Regression Weights of Reduced Combination F2 Model

		Estimate			Estimate
Level_B	← Level_C	.708	ITEM29	← Level_B	.368
Level_A	← Level_B	1.001	ITEM31	← Level_B	.496
ITEM21	← Level_A	.673	ITEM34	← Level_B	.323
ITEM24	← Level_A	.340	ITEM35	← Level_B	.429
ITEM25	← Level_A	.481	ITEM22	← Level_C	.731
ITEM26	← Level_A	.442	ITEM20	← Level_C	.717
ITEM33	← Level_A	.528			

The ICCs of the items in Reduced Combination F2 Model for Level A, Level B, and Level C are presented in Figure 32-34 below.

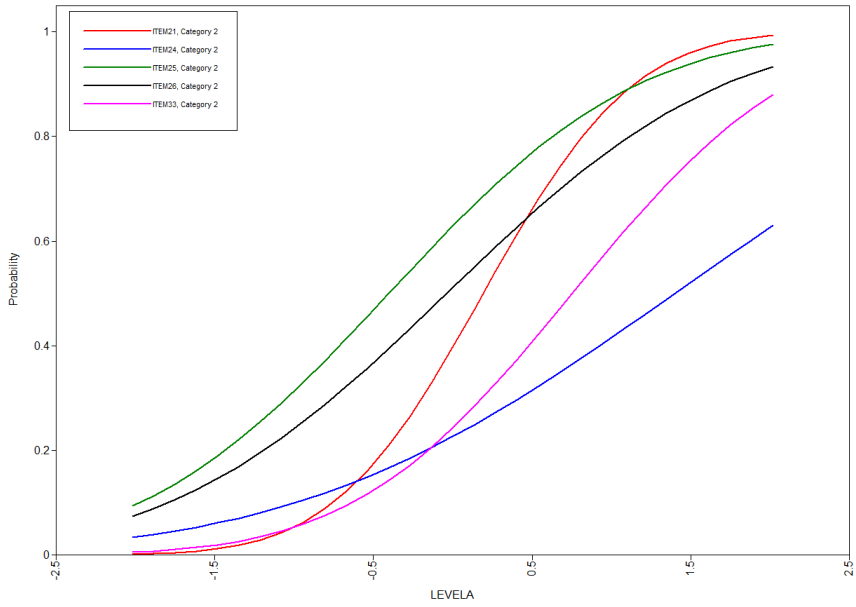


Figure 32. Item characteristic curves of Level A items in reduced combination F2 Model.

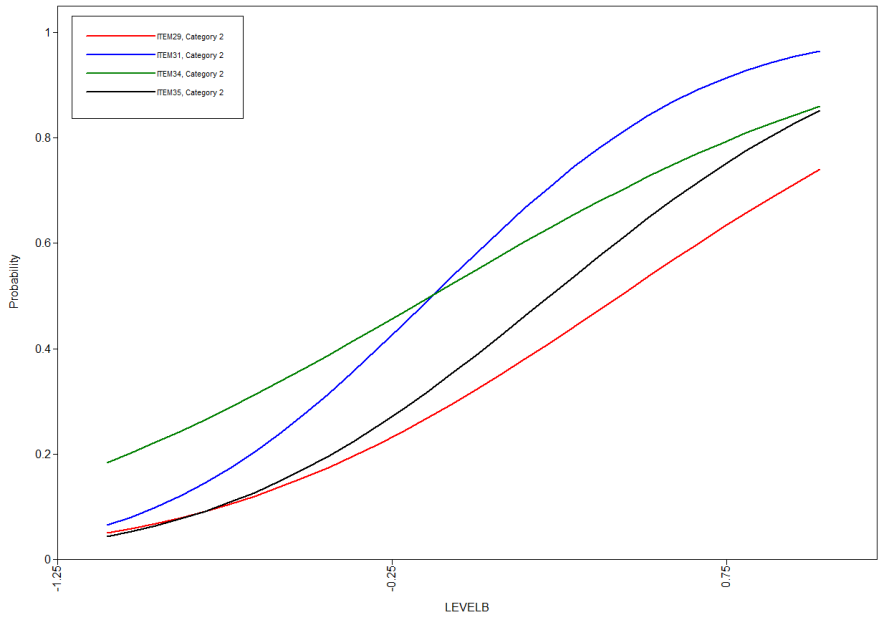


Figure 33. Item characteristic curves of Level B items in reduced combination F2 Model.

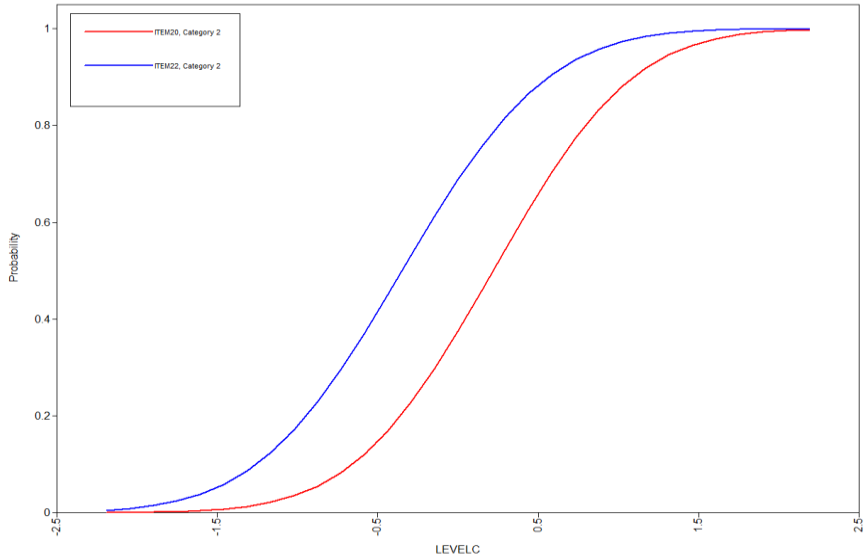


Figure 34. Item characteristic curves of Level C items in reduced combination F2 Model.

The two reduced models are very short considering that they have three constructs to be measured. Combining the two reduced models into one model might give a better quality for the instrument. The combined model, however, should be tested for its validity and item quality. To ensure the goodness of fit of the models used in this study, the testing fit results of the models are presented in the following section.

Testing fit results

A model fit analysis is presented using Structural Equation Modeling (SEM) analysis in SPSS AMOS version 21 and Mplus Version 7 programs. Standard fit statistics used to compare the four models are minimum discrepancy (CMIN) with degree of freedom (DF) value, and the χ^2/DF ratio, as well as its p-value, Comparative Fit Index (CFI), Tucker Lewis Index (TLI), and root mean square error of approximation (RMSEA). CMIN represents the Likelihood Ratio Test statistics, most commonly expressed as a χ^2 statistics. The null hypothesis (H_0) assumes that specification of the factor loadings, factor variances and co-variances, and error variances for the model under study are valid; the χ^2 test simultaneously tests the extent to which this specification is true. Bryne (2010) explained that the p-value associated with χ^2 represents the likelihood of obtaining a χ^2 value that exceeds the value when H_0 is true. Thus, cited Bollen (1989), Bryne (2010) concluded that the higher the probability associated with χ^2 the closer the fit between the hypothesized model (under H_0) and the perfect fit. χ^2/DF ratios less than 2.0 are indicators of good fit, and ratios greater than 2.0 but less than 3.0 are indicators of modest fit (Purpura & Lonigan, 2013). CFI and TLI values of greater than .95, and RMSEA value of less than or equal to .05 are indicators of good fit (Purpura & Lonigan, 2013). In addition, cited by Bryne (2010), Hu and Bentler

(1999) recommended that an RMSEA value between .05 and .08 is an indicator of a moderate fitting. Table 35 displays all indicators of model fit for the five models developed for Form 1.

Table 35

Model Fit Indices of Form 1

Model	CMIN	DF	P	CMIN/DF	CFI	TLI	RMSEA
Initial	109.016	117	.688	.932	1.00	1.042	.000
Expert 1	109.332	117	.680	.934	1.00	1.041	.000
Expert 2	106.720	117	.742	.912	1.00	1.055	.000
Combination	107.582	117	.722	.920	1.00	1.050	.000
Reduced Combination	26.399	33	.785	.800	1.00	1.042	.000

In the table we can see that all five models of Form 1 fit the data well (CMIN/DF < 2.00; CFI > .95; TLI > .95; and RMSEA < .05). These results, however, are questionable since the sample size (140) is quite small and this tends to give insignificant p-values.

Table 36 displays all indicators of model fit for the five models developed for Form 2. In Table 4.33, it was found that the Initial Model of Form 2 did not fit the data well (P-value < 0.05; CFI < .95; and TLI < .95). Likewise, the Expert 1's Model of Form 2 also had P-value $\leq$ 0.05, CFI < .95, and TLI < .95. The Expert 2's Model and the Combination Model of Form 2 fit the data well. For Reduced Item Combination F2 Model, it was found that the p-value satisfactorily met the criteria of good fit models for .05 significant levels. All other criteria were fulfilled satisfactorily. Therefore, we can conclude that the reduced item models fit the data well. The discussion on latent variables model will be discussed in the next section.

Table 36

Model Fit Indices of Form 2

Model	CMIN	DF	P	CMIN/DF	CFI	TLI	RMSEA
Initial	155.142	118	.012	1.315	.914	.889	.022
Expert 1	156.956	117	.008	1.342	.940	.930	.023
Expert 2	140.798	117	.066	1.203	.964	.958	.018
Combination	136.340	117	.107	1.165	.971	.966	.016
Reduced Combination	63.504	42	.018	1.512	.968	.958	.028

To obtain latent variables distribution, we will use the Reduced Combination F1 & F2 Model's parameters to define the latent variable models. These models were chosen because they have very good standardized regression weights that give consistent results for the latent variable distribution (for instance, they do not have negative standardized regression weights that will affect the latent variable distribution).

This analysis was conducted to develop a model that will be useful in diagnosing students' developmental level in learning statistics. By developing the latent variable models, students' scores for each level can be generated. Based on these scores, students' developmental levels can be predicted. Detailed description of this analysis will be presented in the next subsection.

Continuous latent variables modeling

A latent variable model is a statistical model that relates each latent variable with one or more observed/latent variables. In this study models for latent variables Level A, Level B, and Level C were developed. The estimates of latent variable scores were calculated using the standardized regression weights of the best SEM models developed as explained in the previous sections. The estimates of latent variable scores were used to

describe the distribution of the latent variables: Level A, Level B, and Level C. All students were assigned to their Level A, Level B, and Level C scores. By estimating students' scores in each level, it was expected that we could describe their developmental phase in learning statistics. We could also infer the characteristics of group of students in each level. The description of the latent variable modeling framework is presented below.

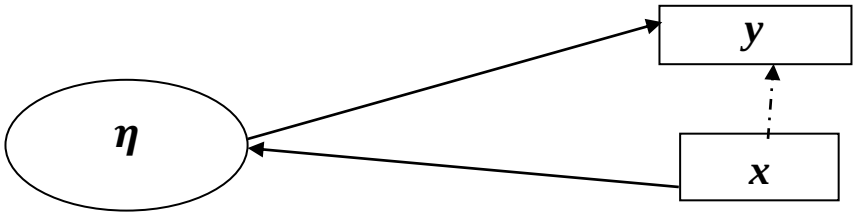


Figure 35. General modeling framework.

Muthen (2002) gives a brief overview of statistical analysis with latent variables. Muthen considers a special case of the general modeling framework shown in Figure 25. Muthen assumes that the latent variables and their indicators are continuous. In this study, however, the outcomes are binary meanwhile the latent variables are assumed to be continuous.

The framework is characterized by using continuous latent variables, denoted by the vector **η**, shown as an ellipse in Figure 35. In this model, latent variables are measured indirectly through multiple indicators ***x*** and ***y*** that capture different aspects of the constructs. Suppose ***y*** has *p* indicators, then the measurement part of the model is defined in terms of the *p*-dimensional binary outcome vector ***y***,

$$\boldsymbol{y} = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta} + \mathbf{K}\boldsymbol{x} + \boldsymbol{\epsilon}$$

where η is an *m*-dimensional vector of latent variables, $\boldsymbol{x}$ is a *q*-dimensional vector of covariates, $\boldsymbol{\epsilon}$ is a p-dimensional vector of residuals or measurement errors which are uncorrelated with other variables, $\boldsymbol{\nu}$ is a *p*-dimensional parameter vector of measurement intercepts, $\boldsymbol{\Lambda}$ is a $p \times m$ parameter matrix of measurement slopes or factor loadings, and $\mathbf{K}$ is a $p \times q$ parameter matrix of regression slopes. The structural part of the model is defined in terms of the latent variables regressed on each other and the *q*-dimensional vector $\boldsymbol{x}$ of independent variables,

$$\boldsymbol{\eta} = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\,\boldsymbol{x} + \boldsymbol{\zeta} \qquad (1)$$

Where $\boldsymbol{\alpha}$ is an *m*-dimensional parameter vector, $\mathbf{B}$ is an $m \times m$ parameter matrix of slopes for regressions of latent variables on other latent variables. $\mathbf{B}$ has zero diagonal elements and it is assumed that $\mathbf{I} - \mathbf{B}$ is non-singular. Furthermore, $\boldsymbol{\Gamma}$ is an $m \times q$ slope parameter matrix for regressions of the latent variables on the independent variables, and $\boldsymbol{\zeta}$ is an *m*-dimensional vector of residuals. For standardized latent variable estimates the *m*-dimensional parameter vector $\boldsymbol{\alpha}$ is equal zero.

Based on equation (1), for Form 1, standardized estimates of Level C can be calculated by taking the linear combination of observed variables (ITEM 04 and ITEM 18) with coefficients the slopes or regression weights of Level C on both items. A student who answered ITEM 04 and ITEM 18 correctly will have Level C score .688 * 1 + .304 * 1 = .992. Another student who answered ITEM 04 correctly but answered ITEM 18 incorrectly will have Level C score .688 * 1 + .304 * 0 = .688.

Since Level B has regression weight on Level C, then the standardized latent variable estimates of level B can be calculated as a linear combination of Level C, ITEM 02, ITEM 03, ITEM 08, ITEM 12, and ITEM 17 with regression weights of Level B on

Level C and of Level B on the five items are applied as the coefficients of the linear combination. Using the same argument, Level A estimates can be calculated as a linear combination of Level B, ITEM 05, ITEM 11, and ITEM 14 with the coefficients of the linear combination are the regressions weights of Level A on Level B and on the three items. The value of each ITEM is binary, 1 if students answered the item correctly and 0 if the item was answered incorrectly.

The explanation about the standardized latent variable estimates in this study can be summarized as the following:

Level C = .688 * ITEM 4 + .304 * ITEM 18

Level B = .989 * Level C + .618 * ITEM 02 + .766 * ITEM 03 + .458 * ITEM 08 + .619* ITEM 12 + .425 * ITEM 17

Level A = .874 * Level B + .752 * ITEM 05 + .583 * ITEM 11 + .627 *ITEM 14

It was found that Level C values range from .00 to .99, Level B values ranged from .00 to 3.87, and Level A values ranged from .00 to 5.34. There were 88 participants (65.2 %) who answered the two Level C items incorrectly; meanwhile there were 16 (12%) and 7 (5.3 %) participants who incorrectly answered all items in Level B and Level A respectively. Since the mode of all three levels were zero then we can infer that most participants who took Form 1 did not develop into Level C.

The descriptive statistics of the standardized distributions of the latent variables Level A, Level B, and Level C of Form 1 are presented in Table 37.

Table 37

Descriptive Statistics of Distributions of Levels' Scores of Form 1

		LEVEL_C	LEVEL_B	LEVEL_A
N	Valid	135	133	133
	Missing	5	7	7
Mean		.2063	1.4796	2.2011
Std. Error of Mean		.02755	.08972	.12037
Median		.0000	1.3850	1.9620
Mode		.00	.00	.00
Std. Deviation		.32008	1.03466	1.38818
Variance		.102	1.071	1.927
Skewness		1.264	.349	.236
Std. Error of Skewness		.209	.210	.210
Kurtosis		.158	-.779	-.964
Std. Error of Kurtosis		.414	.417	.417
Range		.99	3.87	5.34
Minimum		.00	.00	.00
Maximum		.99	3.87	5.34
Sum		27.86	196.79	292.75
Percentiles	10	.0000	.0000	.4003
	20	.0000	.4580	.7520
	30	.0000	.7660	1.2277
	40	.0000	1.0440	1.6641
	50	.0000	1.3850	1.9620
	60	.0000	1.7366	2.4888
	70	.3040	1.9859	3.1231
	80	.6880	2.4669	3.5586
	90	.6880	2.9480	4.1014

Students who already reached mastery in Level A but did not develop into Level B should have been able to answer all items in Level A. Therefore their Level A values should have reached 1.962, found by calculating .752 + .583 + .627. Since it is possible

that students made unintentional mistakes, then we considered allowing students to answer one Level A item incorrectly. If they made one mistake in answering Level A item, their Level A values would have reached at least 1.21 (1.962 -.752). Considering the cut points for 10 equal groups that are presented in the last rows of Table 37, it is reasonable if we take 1.21 as the cut off value for Level A. This means students who had Level A values that were equal or bigger than 1.21 were considered to have developed Level A mastery in learning statistics. From Table 37 we found that around 70% of participants who took Form 1 had developed Level A mastery in learning statistics.

To determine the cut off for Level B, we considered the equation for Level B. If a student who had developed into Level B but had not developed into Level C yet, their Level B scores should have reached 2.886 (.618 + .766 + .458 + .619 + .425). Again, if we assume that students could have made one unintentional mistake, then the minimum Level B scores should have reached 2.12 (2.886 - .766). From Table 37 we found that the smallest number that is larger than 2.12 for Level B was 1.9859. If we take 2.12 as the cut off number of Level B, then we have more than 20% of students who took Form 1 progressed to Level B. Since there were only two items for Level C, we chose the maximum value (.99) as the cut off for Level C that was reached by less than 10% of participants. The above discussion is summarized in Table 39.

The histogram of standardized distribution of Level C, level B, and Level A is presented in Figure 36 - Figure 38.

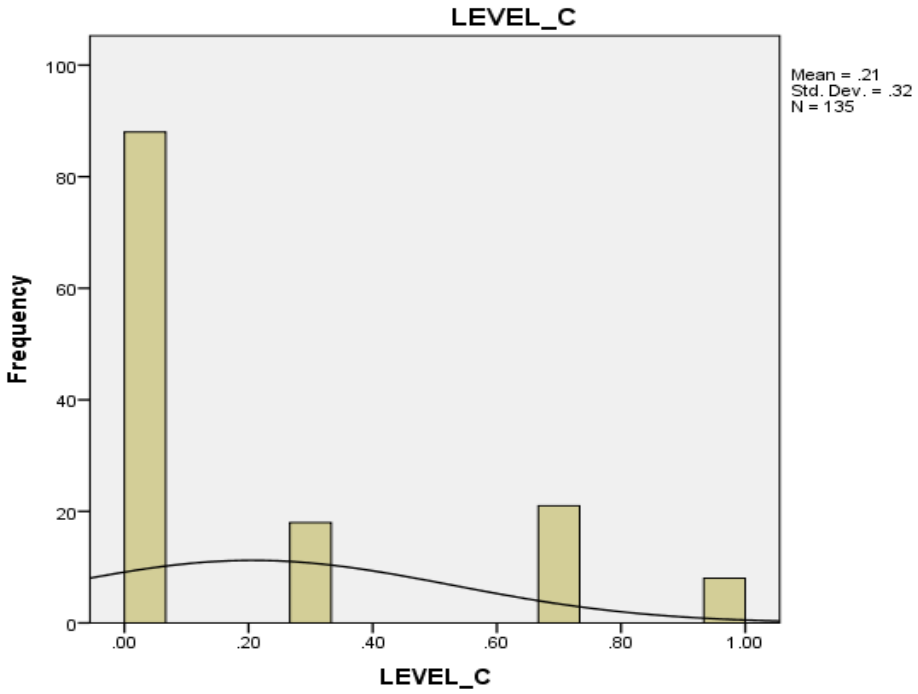


Figure 36. Distribution of Level C of Form 1.

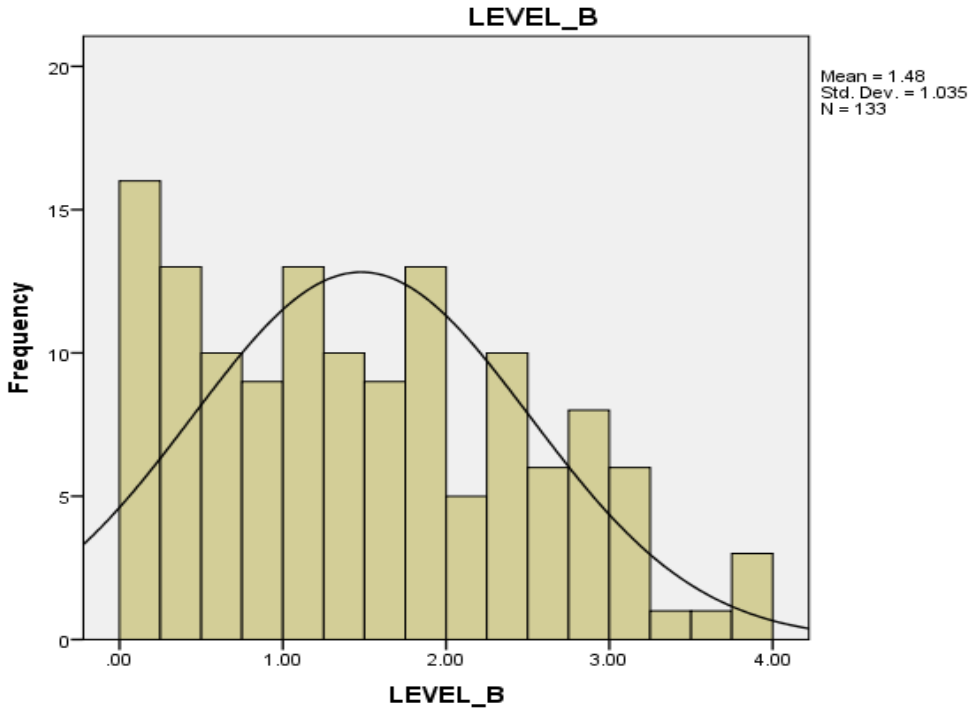


Figure 37. Distribution of Level B of Form 1.

Figure 39 shows the 3-dimensional positions of participants' Level A, B, and C, where a point (*a, b, c*) in the 3-dimensional coordinate represent a participant who has Level A value = *a*, Level B value = *b*, and Level C value = *c*. From the diagram it is clear that most participants were located in the lower scale of Level C but in the middle scale of Level B and higher scale of Level A.

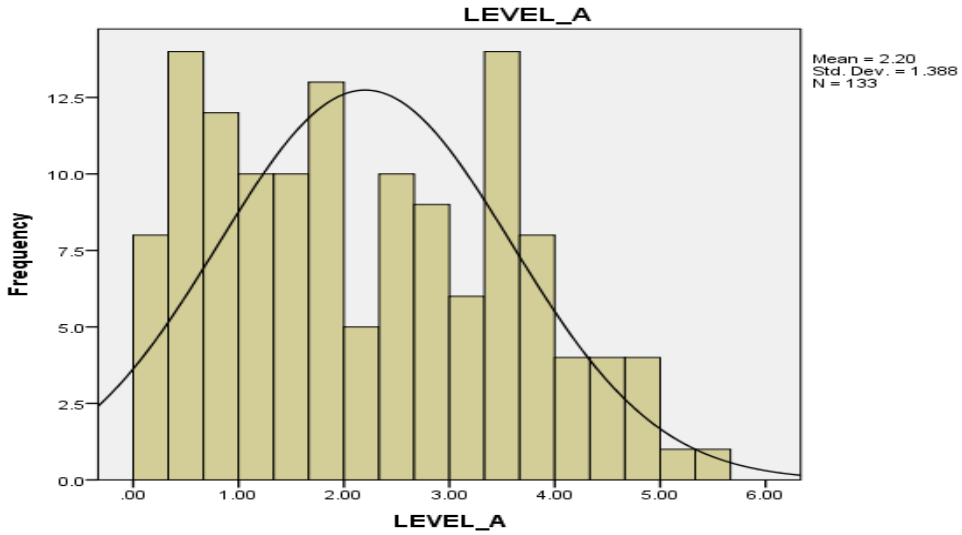


Figure 38. Distribution of Level A of Form 1.

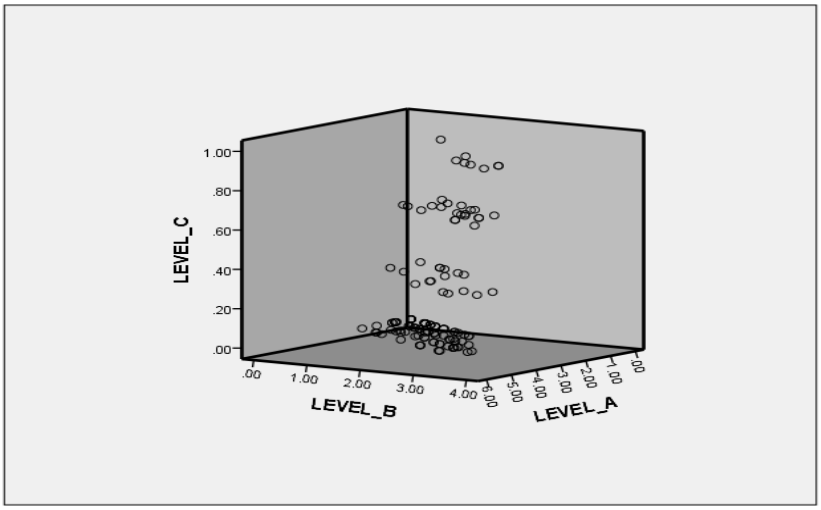


Figure 39. 3-Dimensional scatter plot of Form 1 participants' levels.

Based on standardized regression weights of combination model of Form 2 we define regression estimates for Level A, Level B, and Level C.

Level C = .717 * ITEM 20 + .731 * ITEM 22

Level B = .708 * Level C + .368 * ITEM 29 + .496 * ITEM 31 + .323*ITEM 34 + .429 * ITEM 35

Level A = 1.001 * Level B + .673 * ITEM 21 + .340 * ITEM 24 + .481 * ITEM 25 + .442 * ITEM 26 + .528 * ITEM 33

The descriptive statistics of the standardized distributions of the latent variables Level A, Level B, and Level C of Form 2 are presented in Table 38. It was found that Level C values ranged from .00 to 1.45, Level B values ranged from .00 to 3.25, and Level A values ranged from .00 to 5.11. There were 192 participants (29.3 %) who answered the two Level C items incorrectly; meanwhile there were 29 (4.7%) and 7 (1.2 %) participants who incorrectly answered all items in Level B and Level A respectively. Students who already reached mastery in Level A but had not developed into Level B should have been able to answer all items in Level A. Therefore their Level A values should have reached 2.464, found by calculating .673 + .340 + .481 + .442 + .528.

Since it is possible that students made unintentional mistakes, then we considered allowing students answer two Level A items incorrectly, considering there were five Level A items involved. If they made one mistake in answering Level A item, their Level A values would have reached at least 1.263 (2.464 - .673 - .528). Considering the cut points for 10 equal groups that are presented in the last rows of Table 38, it is reasonable if we take 1.263 as the cut off value for Level A. This means students who had Level A values that were equal or bigger than 1.263 were considered to have developed Level A

mastery in learning statistics. From Table 38 we found that more than 90% of participants who took Form 2 developed Level A mastery in learning statistics.

Table 38

Descriptive Statistics of Distributions of Levels' Scores of Form 2

		LEVEL_A	LEVEL_B	LEVEL_C
N	Valid	603	618	655
	Missing	54	39	2
Mean		2.3352	1.3038	.7598
Std. Error of Mean		.04791	.02799	.02249
Median		2.1957	1.3046	.7310
Mode		5.11	2.64	1.45
Std. Deviation		1.17640	.69585	.57563
Variance		1.384	.484	.331
Skewness		.306	.117	-.093
Std. Error of Skewness		.100	.098	.095
Kurtosis		-.609	-.723	-1.411
Std. Error of Kurtosis		.199	.196	.191
Range		5.11	3.25	1.45
Minimum		.00	.00	.00
Maximum		5.11	3.25	1.45
Sum		1408.11	805.78	497.67
Percentiles	10	.8540	.3680	.0000
	20	1.2804	.6910	.0000
	30	1.6114	.8640	.7170
	40	1.9314	1.0252	.7310
	50	2.1957	1.3046	.7310
	60	2.5709	1.4090	.7310
	70	3.0466	1.7556	1.4480
	80	3.4295	1.9502	1.4480
	90	3.9330	2.2732	1.4480

To determine the cut off for Level B, we considered the equation for Level B. If a student who developed into Level B but had not developed into Level C yet, their Level B scores should have reached 1.616 (.368 + .496 + .323 + .429) that represented the

student had reached Level A cutoff and answered all items in Level B correctly. Again, if we assume that students could have made one unintentional mistake, then the minimum Level B scores should have reached 1.293 (1.616-.323).

From Table 38 we also found that if we took 1.293 as the cut off value of Level B, more than 50% of participants who took Form 2 developed Level B Mastery. With a similar argument with Form 1 analysis, we chose 1.448 as the cut off score for Level C. This indicated to us that about 30% of participants developed Level C mastery. This result is questionable because only 10% of participants who took Form 2 were high school students who were hypothesized to have developed Level C mastery in learning statistics. Since there were only two Level C items that could be used in the final analysis, this finding needs to be analyzed further. The discussion about the cut off values of levels in both forms is summarized in Table 39.

Table 39

Acceptable Range Scores for All Levels

Form 1			Form 2		
Level A	Level B	Level C	Level A	Level B	Level C
1.21 – 5.34	2.12– 3.87	.99	1.263 – 5.11	1.12– 3.25	1.448

From Table 39 we can infer that a student who took Form 1, for instance student with ID # 19 and had level values written in 3-dimensional coordinates (3.37, 3.14, .69) where 3.37 was the Level A value, 3.14 was the Level B value, and .69 was the level C value, might have progressed to Level B, but had not developed into Level C yet. On the other hand student #34 who took Form 1 and had Level coordinates (4.57, 2.98, .99) had progressed to Level C. Both students were 11th graders, but it seems their grade level did

not align to the GAISE level. Further investigation is needed to uncover the relationship between school grade level and GAISE Level measured by the instrument.

When we investigated the level coordinates of student #122, a seventh grader, who took Form 2, we obtained the coordinates (1.30, .52, .73), and therefore this student should have been categorized as a Level A student, since the student's scores did not meet the cutoff for Level B and Level C. Another example, student #146, also a seventh grader who took Form 2 and had level coordinates (1.51, 1.03, 1.45) might have developed into Level A and Level C, but had not developed into Level B. This strange phenomenon might have been caused by other factors. As we mentioned before, the lack of items for Level C might have caused this anomalous result. On the other hand, student #48, a grade 10 student who took Form 2 and had level coordinates (1.31, 1.30, .72) might have developed into Level B. Based on this analysis, we assigned each student with their highest developed level if they also demonstrated that they had regressed to a lower level(s). For example, if they satisfied Level A and Level B cutoff but not Level C, they were considered to have progressed to Level B. But, if they progressed to Level B but not satisfied the cut off of Level A, then they were considered still at the Level A developmental level. Therefore for the anomalous case found for student #146 mentioned above, we assigned the students to Level A.

The histogram of standardized distribution of Level C, level B, and Level A of Form 2 is presented in Figure 40–Figure 42. From Figure 40, it was found that the lacking of items for Level C caused the scores for Level C to have very small variation; meanwhile for Level A and Level B the distributions looked normal.

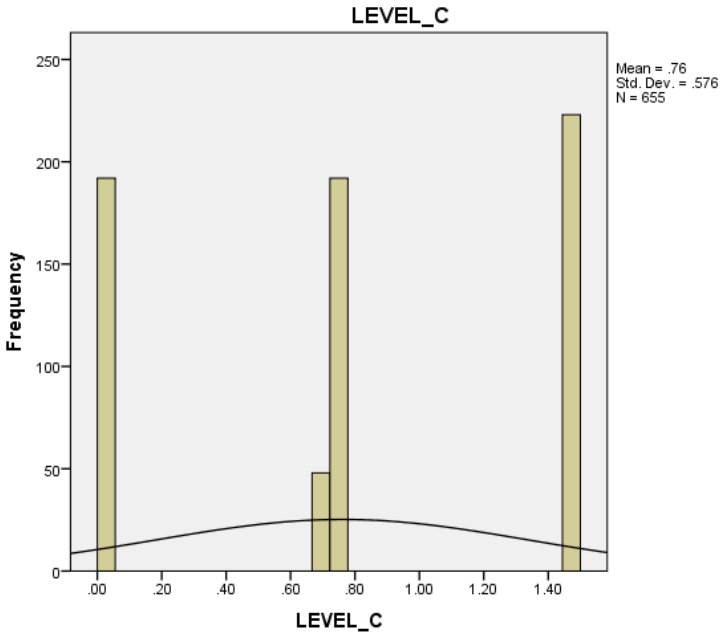


Figure 40. Distribution of Level C of Form 2.

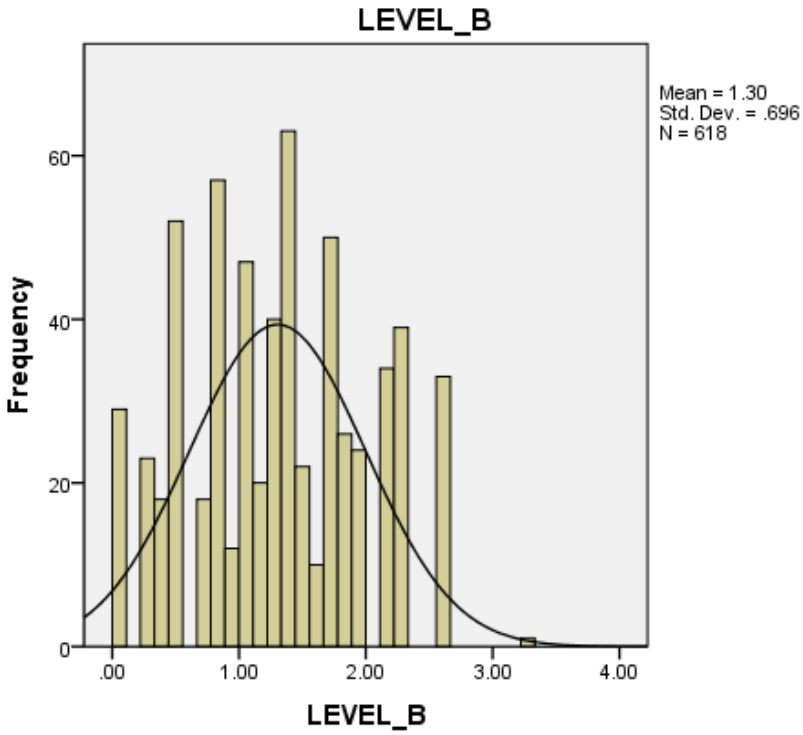


Figure 41. Distribution of Level B of Form 2.

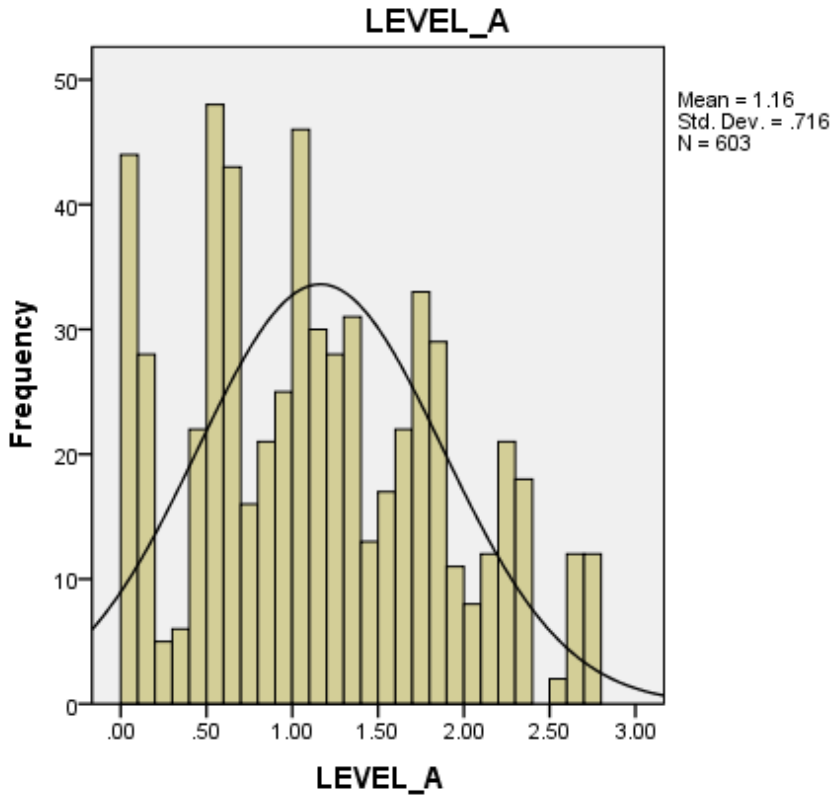


Figure 42. Distribution of Level A of Form 2.

Figure 43 shows the 3-dimensional positions of participants' Level A, B, and C, where a point (*a, b, c*) in the 3-dimensional coordinate represent a participant who has Level A value = *a*, Level B value = *b*, and Level C value = *c*. The 3-dimensional scatter plot shows linear progressions of scores along the scale of Level A, Level B, and Level C. From the plot it is clear that students could be categorized into three similar size groups based on their Level C scores.

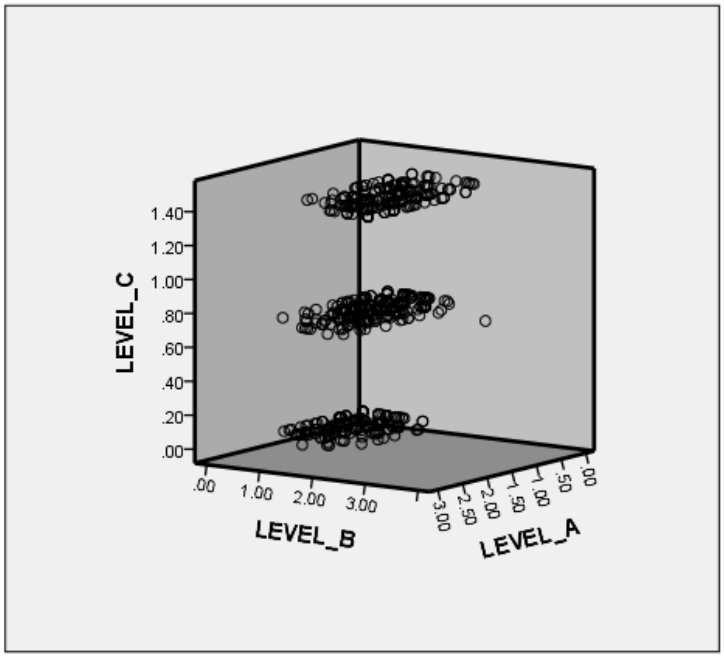


Figure 43. 3-Dimensional scatter plot of Form 2 participants' levels.

Table 40 shows the distribution of level alignment based on students' scores on all levels for Form 1. It can be seen in Table 40 that the cut off set for the alignment works almost perfectly for discriminating students who took Form 1.

Table 40

Form 1 Level Assignment

Reach Level A Score?	Reach Level B Score?	Reach Level C Score?	Level Assigned	Number of participants	Percentage
No	No	No	A	46	32.9%
Yes	No	No	A	57	40.7%
No	Yes	No	A	0	0%
No	Yes	Yes	A	0	0%
No	No	Yes	A	0	0%
Yes	No	Yes	A	1	0.7%
Yes	Yes	No	B	29	20.7%
Yes	Yes	Yes	C	7	5%
Total				140	99.9%

Some anomalous results, however, are found for Form 2 that can be seen in the highlighted lines in Table 41. About 6.4% of the participants who took Form 2 had scores that were inconsistent. Further investigation needs to be conducted for Form 2.

Table 41

Form 2 Level Assignment

Reach Level A Score?	Reach Level B Score?	Reach Level C Score?	Level Assigned	Number of participants	Percentage
No	No	No	A	137	20.9%
Yes	No	No	A	135	20.5%
No	Yes	No	A	9	1.4%
No	Yes	Yes	A	6	.9%
No	No	Yes	A	12	1.8%
Yes	No	Yes	A	15	2.3%
Yes	Yes	No	B	153	23.3%
Yes	Yes	Yes	C	190	28.9%
Total				657	100%

By aligning students' scores with their developmental level in learning statistics, the instrument developed in this study has been shown to be capable of diagnosing students' developmental level in learning statistics. The last analysis that was conducted in this study involved ordinal regression analysis to investigate the relationships between GAISE levels and students' school grade levels, latest mathematics courses taken, and ages that will be discussed in the next section.

Ordinal Regression Analysis Results

An ordinal regression analysis was conducted to evaluate whether there was an association between students GAISE Levels and four factors: (1) the forms that students took, (2) their latest mathematics courses, (3) their ages, and (4) their school grades. The

latest mathematics courses were categorized into three groups: (1) academic courses in middle school, (2) the advanced courses in middle school, and (3) the academic and advanced courses in high school. The ages of students were also categorized into three groups: (1) 11-13 years old. (2) 14-16 years old, and (3) > 16 years old. Students were categorized into two school grades, high school and middle school. The results of the ordinal regression analysis are presented in Table 42.

Table 42

Case Processing Summary of Ordinal Regression between Students' GAISE Levels and School Grades, Latest Mathematics Course Taken, Forms taken, and Ages

		N	Marginal Percentage
LEVEL	A	418	52.4%
	B	182	22.8%
	C	197	24.7%
School Grade	MS	649	81.4%
	HS	148	18.6%
Latest Math Course	Academic MS	229	28.7%
	Advanced MS	420	52.7%
	HS Math	148	18.6%
Form taken by students;	FORM 1	140	17.6%
	FORM 2	657	82.4%
Age	11-13 years old	564	70.8%
	14-16 years old	205	25.7%
	> 16 years old	28	3.5%
Valid		797	100.0%
Missing		0	
Total		797	

Table 42 shows that 52.4 % of students who participated in this study developed into Level A and 22.8 % of them developed into Level B and about 24.7% of the students

developed into Level C. Table 43 displays parameter estimates of each parameter in the model. Considering the location estimates presented in the table, it was found that among the four factors, only age did not have a significant relationship to GAISE Levels.

Table 43

Parameter Estimates of the Ordinal Regression Model between Students' GAISE Level and School Grades, Latest Mathematics Course Taken, Forms, and Ages

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	LEVEL A	-1.639	.392	17.479	1	.000	-2.407	-.870
	LEVEL B	-.510	.388	1.727	1	.189	-1.271	.251
Location	Middle School	-.732	.305	5.776	1	.016	-1.329	-.135
	High School	0[a]	.	.	0	.	.	.
	Academic MS	-1.058	.175	36.384	1	.000	-1.402	-.714
	Advanced MS	0[a]	.	.	0	.	.	.
	HS Math Course	0[a]	.	.	0	.	.	.
	FORM 1	-1.822	.257	50.120	1	.000	-2.327	-1.318
	FORM 2	0[a]	.	.	0	.	.	.
	11-13 years old	-.527	.475	1.230	1	.267	-1.458	.404
	14-16 years old	-.686	.417	2.707	1	.100	-1.504	.131
	> 16 years old	0[a]	.	.	0	.	.	.

Link function: Logit.

a. This parameter is set to zero because it is redundant.

All parameter estimates of the other three factors were significant with negative values that suggest: (1) middle school grades were associated with Level A and high school grades were associated with the higher GAISE levels; in other words, for a one unit increase in school grades (i.e. going from Middle School to High School), we would expect that the odds of being in Level B or Level C increase by a factor of 2.1 ($= e^{.732}$) , given all of the other variables in the model were held constant; (2) For an additional one unit increase in academic middle school mathematics course (i.e. going from Academic

MS to Advanced MS) we would expect that the odds of being in Level B or Level C increase by a factor of 3 ($= e^{1.1}$); and (3) For FORMs, we would say that for an additional one unit increase of FORMs (i.e. getting FORM 2 instead of FORM 1), we would expect that the odds of being in Level B or Level C increase by a factor of 6.2 ($= e^{.1832}$), given that all of the other variables in the model were held constant. These results infer that students' GAISE levels were influenced by their school grades and latest mathematics courses taken. The higher the school grades, the higher the GAISE Levels. Similarly, the more advanced mathematics courses students took, the higher their GAISE Levels. The results also suggest that students' GAISE Levels were influenced by the instrument form that they took. A student was more likely to be identified as a Level A student if he/she took FORM 1 than if he/she took FORM 2. This might indicate that both forms might not be parallel where FORM 1 is more sensitive in identifying Level A students meanwhile FORM 2 is more sensitive in identifying higher levels.

There are several R^2 -like statistics that can be used to measure the strength of the association between the dependent variable and the predictor variables. They are not as useful as the R^2 statistic in regression, however, since their interpretation is not straightforward.

Three commonly used statistics are:

- Cox and Snell R^2

$$R^2{}_{cs} = 1 - \left(\frac{L(B^0)}{L(\hat{B})}\right)^{\frac{2}{n}}$$

- Nagelkerke's R^2

$$R^2{}_N = \frac{R^2{}_{cs}}{1 - L(B^0)^{2/n}}$$

- McFadden's R^2

$$R^2{}_M = 1 - \left(\frac{L(\hat{B})}{L(B^0)}\right)$$

where $L(\hat{B})$ is the log-likelihood function for the model with the estimated parameters and $L(B^0)$ is the log-likelihood with just the thresholds, and *n* is the number of cases (sum of all weights) (Norusis, 2011). Table 43 displays the values of all pseudo *R*-square statistics for the ordinal regression model with predictor variable latest mathematics course taken.

Table 44

Pseudo R-square of the Ordinal Regression Model between Students' GAISE Level and School Grades, Latest Mathematics Course Taken, Forms, and Ages

Cox & Snell	.129
Nagelkerke	.148
McFadden	.068

Link function: Logit.

The Cox and Snell's pseudo R-squared indicates that about 13% of variation in students' GAISE levels was determined by one or more factors being considered that were School Grades, Latest Mathematics Course Taken, Forms taken, and Ages. Likewise, the Nagelkerke's and McFadden's pseudo R-squared indicates that 15% and 7% of variation in students' GAISE levels were determined by one or more factors mentioned above.

Summary

All four models developed for items in Form 1 fit the data well based on all goodness of fit criteria used in this study, that are p-value, CFI, TLI, and RMSEA. In Initial F1 Model, Expert 2 F1 Model, and Combination F1 Model, it is found that ITEMS

06, 07, 09, 10, 13, 15, and 16 were problematic due to their factor loadings that were insignificant or their standardized regression weights that were less than .2. For Expert 1F1 Model, ITEMS 06, 07, 09, 10, 13, 16, and 18 were also problematic due to the same reason with the three models mentioned before. These problematic items indicate that the items might have to be removed from the instrument.

All four models developed for items in Form 2 fit the data well based on RMSEA goodness of fit criterion. Combination F2 Model also fulfilled CFI criterion. In all models for Form 2, it was found that ITEM 23, 27, 28, 30, 32, and 36 were problematic due to their factor loadings that were insignificant or their standardized regression weights that were less than .2. These problematic items indicate the items might have to be removed from the instrument.

Based on the SEM analyses results, one new model was developed for each form. Reduced Combination F1Model was developed from Combination F1 Model by deleting problematic ITEMS 06, 07, 09, 10, 13, 15, and 16 and preserve the alignment of other items with the GAISE levels; meanwhile Reduced Combination F2 Model was developed from Combination F2 Model by deleting problematic ITEMS 23, 27, 28, 30, 32 and 36. The alignments of the other items with the GAISE levels were preserved. All parameter estimates of the reduced models were significant and the standardized regression weights of all factors in these two models were adequate (greater than .2).

Students' scores for each level were determined using the standardized regression weights estimates for each relation. Based on their level scores, each student was aligned with their developmental level. Ordinal regression analyses were conducted to investigate the relation between students' GAISE Levels and their school grade levels, latest

mathematics courses, and ages, respectively. It was found that only scores of Grade 7 were related to GAISE Levels. This explains that there was not enough evidence to conclude that GAISE Levels were parallel to school grade levels. In investigating relations between GAISE Level and mathematics course currently taken at the time of survey administration, the result indicated that there was not enough evidence to conclude that GAISE Levels were parallel to the latest mathematics course taken. Considering the ordinal regression model that relates students' GAISE levels and their ages, it was found that there may not be any relation between students' GAISE Levels and their ages.

CHAPTER V

CONCLUSIONS AND IMPLICATIONS

Motivated by the increase of content and rigor of statistics in school mathematics curriculum and the need to understand how students develop their understanding of statistical concepts and the learning trajectories of the concepts, this study has made an initial attempt to identify students' developmental levels in learning statistics using the Pre-K-12 GAISE Framework (Franklin, et al., 2007). This study also has attempted to provide empirical evidences of the learning trajectories of several statistical concepts per grade level as hypothesized by the K-12 Common Core State Standards in Mathematics (National Governors Association Center for Best Practices, and Council of Chief State School Officers, 2010) by examining the degree of association between the GAISE levels and school grade levels. This framework suggests that students develop their understanding of statistical investigation process (formulating questions, collecting data, analyzing data, interpreting results, and understanding the nature and focus on variability) through three hierarchical levels, Level A, Level B, and Level C. The CCSS-M provides learning trajectories of several statistical concepts at each grade level grounded from theories and empirical studies on teaching and learning statistics. These learning trajectories are aligned with the developmental levels of GAISE. In order to identify students' developmental level in understanding the statistical investigation process, an

instrument to measure students' developmental level in learning statistics was developed. The 36 items in the instruments were designed to measure students' developmental level suggested by the Pre-K-12 GAISE Framework (Franklin, et al., 2007). Two expert panels, each with two experts in statistics and statistics education, aligned the items into appropriate levels that they were supposed to measure suggested by Pre-K-12 GAISE Framework. The instrument items were divided into two forms and then administered to 649 middle school students and 148 high school students. Participants' responses were used to analyze the quality of the items using Classical Test Theory (CTT) analysis and also used to confirm the theoretical framework suggested by the Pre-K-12 GAISE Framework by applying Structural Equation Modeling (SEM) Analysis. Data of students' responses were then used to classify participants based on their developmental levels. Students' responses were also descriptively analyzed to summarize the learning trajectories of statistical concepts captured from students' responses to the instrument items. This chapter is intended to discuss the conclusions taken from the results found, and also their implications in the field of mathematics and statistics education research. This chapter is organized follows: first, a discussion on the conclusions of the results discussed in chapter IV is presented; second, the implication that this study might provide to the field of mathematics and statistics education research is discussed; and third, future directions on continuing the efforts that have been originated by this study are discussed.

Conclusions

The results presented in Chapter IV suggest that students' understanding in formulating questions was quite well where at least one third of the students could answer items to assess this process component correctly. For collecting data, on average, more

than half of the participants were able to respond to the items that assess this process component correctly. In analyzing data, students' performances were not adequate. About half of the items in these process components were answered incorrectly by more than 70% of participants. This result also indicates that items for this process component were difficult for participants. Since the experts have considered the items to be appropriate to measure GAISE Levels mastery of the students, the low correct responses might have been caused by students' lack of knowledge on the contents assessed. In addition, students did quite well in interpreting results. On average, more than 40% of the participants were able to correctly answer the items that assessed their understanding of interpreting results based on the given data representations. In Nature of Variability, it was found that three of seven items had correct responses that were less than 23% (ITEMS 10, 23, and 28). On average, however, students performed better in this process component than in analyzing data. According to an expert who contributed to this study, these three items actually assessed students' understanding on natural and chance variability, variability within a group, and variability between groups. Therefore, we can conclude that middle school and high school students who participated in this study still showed low level understanding of univariate variability that is expected to be understood by Level B students. It can be inferred by this result that students who have developed into Level B in other process components might not have developed into Level B in understanding variability.

When observing the means of difficulty indices of items in each process component, it was found that the difficulty indices for formulate questions, collect data, analyze data, and interpret results tended to agree with the Pre-K-12 GAISE Framework.

The means of Level A items tended to be higher than the means of Level B items, and the means of Level B items tended to be higher than the means of Level A items. These tendencies however, were not followed by the nature of variability component items. The means of Level B items for this process component tended to be lower than the means of Level C items. This indicates that students might have developed into Level C in nature of variability, but they had not fully developed into Level B. This also indicates that participants developed their understanding similarly across statistical investigation process components except for the nature of variability process component. The Kruskal-Wallis and Whitney-Mann U test, however, showed that the differences of the means were not statistically significant. These results might have been due to chance alone.

The results showed that in formulating questions, collecting data, analyzing data, and interpreting result process components, the patterns tended to be similar, where the lower the levels, the better the performance of students in general. On the other hand, in understanding variability, students tended to perform better in Level C items than in Level B items (see Figure 4.5). This anomaly suggests that either students' understanding of variability might not have followed the path suggested by the Pre-K-12 GAISE Framework and CCSS-M or the items used might not have aligned with the documents' suggestion. Further study on how students develop their understanding of variability needs to be conducted. Thorough investigations of items that measure variability in this study are also necessary to be conducted. For other process components, it was found that the learning trajectories of statistical concepts assessed by the instrument in this study seemed to agree with the Pre-K-12 GAISE Framework and CCSS-M. For example, in the formulating question process component, the Pre-K-12 GAISE Framework suggested that

in Level A students began to develop awareness of statistics question distinction that was assessed by ITEM05 and in Level B, students should have developed an increase of awareness of statistics question distinction by developing skills to start posing their own question. ITEM08 was developed to assess students' awareness to differentiate what kind of statistics question can be posed based on a middle school basic health information data. The result showed that ITEM 05 had a difficulty index of .45; meanwhile ITEM08 had difficulty index .32. This means ITEM05 was responded to correctly by more participants than ITEM08. Point-biserial of ITEM05 and ITEM08 were .580 and .444 respectively. Both point-biserial had .01 level of significant (2-tailed), hence their power in discriminating students based on their general performance in responding the survey items was high. ITEM05 was categorized as a Level A item and ITEM08 was categorized as a Level B item. This indicates that Level A item was responded to correctly by more students than Level B item. It is interesting to analyze whether this pattern was followed by other process components.

In the Collecting Data process component, it was found that ITEM 03, a level A item, had difficulty index .56 that was higher than ITEM 13 that was a Level C item with difficulty index .38. Thus, more students answered Level A item correctly than those who answered Level C item. The same pattern also showed by the other four collecting data items (ITEMS 25, 27, 29, and 31). ITEM 25, a level A item, was responded to correctly by more students than ITEM 27, 29, and 31 that were Level B items. By conducting an ordinal regression analysis between items' levels and their difficulty indices exclusively for items that measured students' understanding on the analyzing data and interpreting results process components, we found that the Level A items tended to have higher

difficulty indices than Level B items, and Level B items tended to have higher difficulty indices than Level C items. Based on this result, we can conclude that students who participated in this study developed their understanding quite similarly across developmental levels. This answers the third research question in this study. This result, however, might have been due to chance alone, since the Kruskal-Wallis and Whitney-Mann U tests revealed that the difference among the mean of difficulty indices of the GAISE Levels in each process component were not statistically significant.

From Confirmatory Factor Analysis (CFA) using Structural Equation Modeling (SEM) analysis, it was found that all models for Form 1 fit the data well. This suggests that the hypothesis that students developed their understanding through Level A, and then through Level B, and finally through Level C is supported by the data. The best model for Form 1 was the Combination Model with p-value .722, CFI >.95, TLI > .95, and RMSEA < .05. Analyzing the regression weight estimates, it was found that several items were problematic. Therefore, an SEM analysis was developed to investigate how these problematic items affected the hypothesized model. By removing the items only for Combination Model, the new model was called the Reduced Combination Model. All regression weight estimates of the last models were significant at .001 levels implying that the model fit the data well. This explains that the removed items might have been best to be excluded from the instrument. This suggestion, however, could cost on shorter length of the survey form that might make it impossible to calculate internal consistency reliability of items for each level.

The CFA of Form 2 also revealed that Combination F2 Model fit the data well. The Combination F2 Model that was developed by combining experts' opinions had p-

value .107. This small value did not weaken the conclusion since we had a large sample size (657) that caused the calculation of p-value to be small. Our confidence was supported by the CFI index of this model that reached .95, the strong model fit cutoff for CFI. As mentioned before a model that has CFI $\geq$ 0.95 is consider a fit model. With TLI = .966 and RMSEA = .016, it was convincing that this model fit the data well. This convincing result was followed by the Reduced Combination F2 Model, which was developed by deleting items that had insignificant factor loadings in Combination F2 Model. The Reduced Combination F2 Model had p-value .18, CFI = .968, TLI = .958, and RMSEA index =.03 that fulfilled the criteria of Model fit. It is convincing that the items that were included in the Reduced Combination F2 Model were good items in measuring students' developmental level as suggested by the Pre-K-12 GAISE Framework. This conclusion will benefit us in developing more items for each level in the future.

In determining the estimates for latent variables, it was found that the method to determine latent variable distribution in SEM Model explained in the previous chapter gave us estimates of Level A, Level B, and Level C scores. Using these estimates, we were able to align each student with his/her GAISE developmental level. This method will be very useful for several purposes such as diagnosing students' developmental level, evaluating instructional approaches by assessing students' developmental level before and after instruction episodes, or evaluating the effects of educational policy in mathematics and statistics education (for example in providing technology in the classroom) to students' developmental level in learning statistics. The instrument

developed in this study, then, will be very useful as a research tool to measure students' developmental levels suggested by the Pre-K-12 GAISE Framework.

For statistical investigation process components, we have shown that the learning trajectories suggested by the Pre-K-12 GAISE Framework seem to be supported by the data, except for the nature of variability process component. One research question that has not been fully addressed using empirical data is the learning trajectories that describe the developmental progression for different statistical concepts. This question, however, was addressed by describing the learning trajectories of statistical concepts from previous studies that can be found in the literature review.

From the ordinal regression analyses conducted to investigate the relation between students' GAISE Levels and their school grade levels, based on the regression parameter estimates it was found a tendency that the higher the school grades the higher the GAISE Levels. It was also found that regular middle school mathematics courses took by students tended to relate with GAISE Level A. Advanced middle school mathematics courses and high school mathematics courses were related to GAISE Level B. The results indicated that students' preparation in mathematics influence their developmental level in statistics. Other finding revealed that the use of FORM 1 influence the number of Level A students identified. This result indicates that FORM 1 and FORM 2 might not parallel. Further investigation is needed to reveal which form has more consistent results.

The findings that we discussed above do not have perfect confidence. Several limitations need to be considered. For that purpose, a description of the limitations of this study is discussed in the following section.

Limitations of the Study

The sampling method used in this study did not involve choosing the participants or the schools randomly. Therefore, the findings of this study cannot be generalized to all populations of middle and high school students. The findings can be used as a baseline in developing conjectures about how students develop their understanding of statistical concepts and about learning trajectories of several statistical concepts or statistical investigation process components.

The proportion of middle school and high school students who participated in this study was unbalanced. The proportion of students who took survey Form 1 and Form 2 were also unbalanced. These fact leads to the difficulties in comparing the results of item analysis of the two forms. Even though the results show a tendency that the results align with the theoretical frameworks that are used in this study, this tendency needs to be taken with caution.

The instrument developed in this study does not cover all statistical concepts that high school and middle school students should be able to do as suggested by the Pre-K-12 GAISE Framework and K-12 Common Core State Standards. Claiming that the instrument can precisely diagnose students' developmental level and learning trajectory in statistics would be misleading. Further studies need to be conducted to develop a more rigorous instrument to measure students' developmental level and learning trajectory in statistics.

Based on the limitations listed above, a discussion on future direction on research that we can or should do to overcome this limitation will be presented in the next section.

This discussion will be useful to address several questions that are left without convincing answers or new questions that arose during the implementation of this study.

Future Directions

Considering the conclusions and the limitations of this study, several ideas of future research that can be conducted have emerged. The ideas were developed in order to answers several questions that are surfaced through the whole implementation processes of this study. The ideas are described in the following paragraphs.

As mentioned before, to be able to generalize the results of the populations of middle school and high school students, a study that carefully chooses their participants using randomized methods, whether randomly sampling the schools that have similar characteristics or use other sampling techniques, will guarantee representative samples so that the findings can be generalized for a larger population of students. By conducting a carefully designed sampling method, the evidence of validity and reliability of the participants' scores can be established because several possible biases can be minimized by randomization. The results will determine whether the instrument to measure students' developmental levels and learning trajectory in statistics developed in this study is strong and has potential to serve many purposes, including: (1) as a diagnostic assessment instrument, (2) as an instrument to measure the effectiveness of instructional approach, or (3) as a research tool to investigate how middle school and high school students learn statistics.

Future studies that include more high school participants are necessary. A balanced proportion of middle and high school students participating in the survey will provide more convincing conclusions since the possible biases developed by unbalanced

proportions of participants can be eliminated. It is also necessary to administer Form 1 to a larger sample so that the psychometric analyses conducted with larger sample size will bring forth more convincing results.

A parallel form reliability test needs to be conducted so that the forms actually measure exactly the same constructs and subjects can be revealed. This test can be conducted by administering both forms to the same sample. If the results show that the forms are parallel, we can then start developing bank items to measure students' developmental levels and learning trajectory of statistics by adding items that measure statistical concepts that had not been included in the instrument. Creating bank items will give flexibility to stakeholders to use the items for different purposes that are related to measuring students' developmental level and learning trajectory in statistics.

Price (2013) described that in confirmatory factor analysis (CFA) perfect model-data fit is rarely obtained. Citing Bollen (1989), Price added that even when a perfect model-data fit is obtained, there are other possible models that fit the data perfectly and the model-data fit is easy to verify. This situation occurred in all five models of Form 1 developed in this study. Ideally, items load exclusively on theoretical factors, in other words, items are expected to have a zero loading for factors not supported by theory. This constraint, however, is not a requirement of a well-defined and useful factor structure. Not even for a simple structure. In this study, in the models for Form 1 it was found that there were items that had large factor loadings for different levels. For instance, ITEM 08 was assigned as a Level B item in Initial F1 Model with standardized factor loading .487 and the item was assigned to Level A in Expert 2F1 Model with slightly higher standardized factor loading .516. Another example was ITEM 17 where in Initial F1

Model it was assigned to Level C and had standardized factor loading.417, but it was assigned to Level B in Expert 2F1 Model with standardized factor loading .458. As discussed before, Initial F1 Model and Expert 2F1 Model both fit the data perfectly. Hence, it is necessary to investigate which factors contributing to a perfect model-data fit using more sophisticated analysis methods rather than claiming the model with the best fit indices as the best model as applied in this study. One method that can be applied is the second-order (hierarchical) confirmatory factor analytic (HCFA) model (Price, Preprint).

Last but not the least, several studies that focused on conducting more rigorous psychometric and statistical analyses on the recent and new data will strengthen the conclusions that have been found. For example, conducting a study implementing more robust regression analyses in investigating the relations among GAISE Levels and school grade levels, mathematics course currently taken, FORM taken, and ages, might reveal more valid results. Among all these possible future directions, most importantly, the final goal is to develop a better understanding on how middle and high school students develop their statistical knowledge and also a better understanding on the learning trajectories of statistical concepts. Several implications of this study that can be used in the efforts to reach this goal are presented in the next subsection.

Implications

The results have shown that the instrument developed in this study is able to diagnose students' developmental levels in learning statistics suggested by the Pre-K-12 GAISE Framework. Several recommendations are provided for all parties that intend to use the instrument for different purposes.

A. For practicality in the classroom.

Expecting teachers to diagnose their students' developmental levels in learning statistics and to apply different approaches for different groups of students seems impractical, not only because it will require significant preparation time for the teachers, but also because not many teachers have strong preparation skills in statistics. Putting too many expectations on teachers to differentiate their instructional approaches in teaching statistics will be too ambitious. It would be more reasonable if the diagnostic assessment is given to the students at school levels and then based on their developmental levels, students are assigned to have statistics lessons from teachers who are prepared to teach statistics lessons for their level.

If the information needed is only about students' developmental level, then an instrument that consists of ITEMS 02, 03, 04, 05, 08, 10, 11, 12, 17, and 18 from Form 1 can be used. If students answer two of the problems in ITEMS 05, 11, and 14 then classify the students as Level A students. If students answer at least two problems of this set of items then observe how they perform on the set of items (ITEM 02, ITEM 03, ITEM 08, ITEM 12, and ITEM 17). If students incorrectly answer more than one item from the set of items, then categorize the students as Level A students; otherwise if the students answer one of the two items (ITEM 04 and ITEM 18) incorrectly then categorize them as Level B students. If students answer ITEM 04 and ITEM 18 correctly, then categorize them as Level C students. Then assign different teachers to teach different groups of students whose levels have been accurately identified. This suggestion might be still unpractical in classroom levels if the number of students in the school is large.

B. As a diagnostic tool for research related purposes.

Use reduced Form 1 or 2 with the following rules:

- **Form 1**
    - Calculate each student's scores for each level as below:

      Level C = .688 * ITEM 4 + .304 * ITEM 18

      Level B = .989 * Level C + .618 * ITEM 02 + .766 * ITEM 03 + .458 * ITEM 08 + .619* ITEM 12 + .425 * ITEM 17

      Level A = .874 * Level B + .752 * ITEM 05 + .583 * ITEM 11 + .627 *ITEM 14

      The score of an item is 1 if the student answers the item correctly and 0 otherwise. For instance if John answers ITEM 4 correctly and answers ITEM 18 incorrectly, then John's Level C score is .688 (.688 * 1 + .304 * 0).
    - Use the following cutoffs to determine students' developmental level:

Form 1		
Level A	Level B	Level C
1.21 – 5.34	2.12– 3.87	.99

    - Students whose Level B scores are less than 2.12 and Level C scores less than .99 are categorized as Level A students.
    - Students whose Level A scores are in the 1.21–5.34 range, Level B scores in the 2.12–3.87 range, and Level C scores less than .99 are categorized as Level B students.

- Students whose Level A scores are in the 1.21–5.34 range, Level B scores in the 2.12–3.87 range and Level C scores equal to .99 are categorized as Level C students.
- If a student has Level B score that is less than 2.12 but has Level C score equal to .99, then the student cannot be categorized as a Level C student, instead, categorize the student as a Level A student.
- If a student has Level A score that is less than 1.21 but has Level B score in the 2.12–3.87 range, then the student cannot be categorized as a Level B student, instead, categorize the student as a Level A student.

- **Form 2**
  - Calculate each student's scores for each level as below:

    Level C = .717 * ITEM 20 + .731 * ITEM 22

    Level B = .708 * Level C + .368 * ITEM 29 + .496 * ITEM 31 + .323*ITEM 34 + .429 * ITEM 35

    Level A = 1.001 * Level B + .673 * ITEM 21 + .340 * ITEM 24 + .481 * ITEM 25 + .442 * ITEM 26 + .528 * ITEM 33

    The score of an item is 1 if the student answers the item correctly and 0 otherwise. For instance if Jane answers ITEM 20 incorrectly and answers ITEM 22 correctly, then Jane's Level C score is .731 (.717 * 0 + .731 * 1).
  - Use the following cutoffs to determine students' developmental level:

Form 2		
Level A	Level B	Level C
1.263 – 5.11	1.12– 3.25	1.448

- Students whose Level B scores are less than 1.12 and Level C scores less than 1.448 are categorized as Level A students.
- Students whose Level A scores are in the 1.263–5.11 range, Level B scores in the 1.12–3.25 range and Level C scores less than 1.448 are categorized as Level B students.
- Students whose Level A scores are in the 1.263–5.11 range, Level B scores in the 1.12–3.25 range, and Level C scores equal to 1.448 are categorized as Level C students.
- If a student has a Level B score that is less than 1.12 but has a Level C score equal to 1.448, then the student cannot be categorized as a Level C student, instead, categorize the student as a Level A student.
- If a student has Level A score that is less than 1.263 but has a Level B score in the 1.12–3.25 range, then the student cannot be categorized as a Level B student, instead, categorize the student as a Level A student.

- Do not use these forms to infer any knowledge of process component or statistics concepts, because there are not enough items in the instrument to make inferences about any knowledge of process component or statistics concepts.

C. As a research tool.

Use Form 1 and Form 2 with large number of participants and balanced proportion across forms. Add more items to each level and each process component to strengthen reliability and to be able to infer performance at each level and at each process component.

D. As an evaluation tool.

Use Form 1 or Form 2 to determine students' developmental levels using the rules described in term A above. Do not use these forms to infer about any knowledge of process component or statistics concepts, because there are not enough items in the instrument to make inferences about any knowledge of process component or statistics concepts

As a final statement, it is expected that the instrument developed in this study could be improved to be a stronger instrument to measure students' developmental level in learning statistics and in general; this study is also expected to contribute to any efforts to develop a better understanding on how middle and high school students develop their statistical knowledge and also a better understanding on the learning trajectories of statistical concepts.

APPENDIX A

LEARNING TRAJECTORY DISPLAY OF
THE COMMON CORE STATE STANDARDS FOR STATISTICS
(Confrey, Maloney, & Nguyen, 2010)

<table>
<tr><th colspan="5">STATISTICAL INVESTIGATIONS AND SAMPLING, DESCRIPTIVE STATISTICS (CENTRAL TENDENCY AND DISTRIBUTION), BIVARIATE DATA AND SCATTERPLOTS AND PROBABILITY (GRADES 6 – 8)
SAMPLING AND DESIGN, DATA DISTRIBUTIONS, PROBABILITY AND BIVARIATE DATA, LINEAR REGRESSION, AND CORRELATION (HIGH SCHOOL)</th></tr>
<tr><th>GRADE 6</th><th>GRADE 7</th><th>GRADE 8</th><th colspan="2">HIGH SCHOOL</th></tr>
<tr><th colspan="2">PROBABILITY</th><th colspan="2">PROBABILITY</th></tr>
<tr><td rowspan="2">N/A</td><td rowspan="2">7.SP.5 Understand that the probability of a chance event is a number between 0 and 1 that expresses the likelihood of the event occurring. Larger numbers indicate greater likelihood. A probability near 0 indicates an unlikely event, a probability around ½ indicates an event that is neither unlikely nor likely, and a probability of near 1 indicates a likely event.

7.SP.6 Approximate the probability of a chance event by collecting data on the chance process that produces it and observing its long-run relative frequency, and predict the approximate relative frequency given the probability. For example, when rolling a number cube 600 times, predict that a 3 or 6 would be rolled roughly 200 times, but probably not exactly 200 times.

7.SP.7ab Develop a probability model and use it to find probabilities of events. Compare probabilities from a model to observed frequencies; if the agreement is not good, explain possible sources of the discrepancy.</td><td rowspan="2">N/A</td><td>LEVEL 1</td><td>N/A</td></tr>
<tr><td>LEVEL 2</td><td>S-CP.2 Understand that two events A and B are independent if the probability of A and B occurring together is the product of their probabilities, and use this characterization to determine if they are independent.

S-CP.4 Construct and interpret two-way frequency tables of data when two categories are associated with each object being classified. Use the two-way table as a sample space to decide if events are independent and to approximate conditional probabilities. For example, collect data from a random sample of students in your school on their favorite subject among math, science, and English. Estimate the probability that a randomly selected student from your school will favor science given that the student is in 10th grade. Do the same for other subjects and compare the results.

S-CP.7 Apply the addition rule, P(A or B) = P(A) + P(B) – P(A and B), and interpret the answer in terms of the [uniform probability] model.</td></tr>
</table>

Learning trajectory display continued

	7.SP.7ab Develop a probability model and use it to find probabilities of events. Compare probabilities from a model to observed frequencies; if the agreement is not good, explain possible sources of the discrepancy. **a.** Develop a uniform probability model by assigning equal probability to all outcomes, and use the model to determine probabilities of events. **For example**, if a student is selected at random from a class, find the probability that Jane will be selected and the probability that a girl will be selected **b.** Develop a probability model (which may not be uniform) by observing frequencies in data generated from a chance process. **For example**, find the approximate probability that a spinning penny will land heads up or that a tossed paper cup will land open-end down. Do the outcomes for the spinning penny appear to be equally likely based on the observed frequencies? **7.SP.8abc** Find probabilities of compound events using organized lists, tables, tree diagrams, and simulation. **a.** Understand that, just as with simple events, the probability of a compound event is the fraction of outcomes in the sample space for which the compound event occurs. **b.** Represent sample spaces for compound events using methods such as organized lists, tables and tree diagrams. For an event described in everyday language (***e.g.,*** "rolling double sixes"), identify the outcomes in the sample space which compose the event. **c.** Design and use a simulation to generate frequencies for compound events. **For example**, use random digits as a simulation tool to approximate the answer to the question: If 40% of donors have type A blood, what is the probability that it will take at least 4 donors to find one with type A blood?			

Learning trajectory display continued

			LEVEL 3	**S-CP.3** Understand the conditional probability of ***A*** given ***B*** as $^{P(A \text{ and } B)}/_{P(B)}$, and interpret independence of ***A*** and ***B*** as saying that the conditional probability of ***A*** given ***B*** is the same as the probability of ***A***, and the conditional probability of ***B*** given ***A*** is the same as the probability of ***B***. **S-CP.5** Recognize and explain the concepts of conditional probability and independence in everyday language and everyday situations. **For example**, compare the chance of having lung cancer if you are a smoker with the chance of being a smoker if you have lung cancer. **S-CP.6** Find the conditional probability of ***A*** given ***B*** as the fraction of ***B's*** outcomes that also belong to ***A***, and interpret the answer in terms of the [uniform probability] model. **S-MD.1 (+)** Define a random variable for a quantity of interest by assigning a numerical value to each event in a sample space; graph the corresponding probability distribution using the same graphical displays as for data distributions.
			LEVEL 4	**S-CP.8 (+)** Apply the general Multiplication Rule in a uniform probability model, **P(A** and **B) = P(A)P(B | A)= P(B)P(A | B),** and interpret the answer in terms of the [uniform probability model. **S-MD.2 (+)** Calculate the expected value of a random variable; interpret it as the mean of the probability distribution. **S-MD.5ab (+)** Weigh the possible outcomes of a decision by assigning probabilities to payoff values and finding expected values. **a.** Find the expected payoff for a game of chance. **For example**, find the expected winnings from a state lottery ticket or a game at a fast-food restaurant. **b.** Evaluate and compare strategies on the basis of expected values. **For example**, compare a high-deductible versus a low-deductible automobile insurance policy using various, but reasonable, chances of having a minor or a major accident.

Learning trajectory display continued

			LEVEL 5	**S-CP.9** Use permutations and combinations to compute probabilities of compound events and solve problems. **S-MD.3 (+)** Develop a probability distribution for a random variable defined for a sample space in which theoretical probabilities can be calculated; find the expected value. **For example**, find the theoretical probability distribution for the number of correct answers obtained by guessing on all five questions of a multiple-choice test where each question has four choices, and find the expected grade under various grading schemes. S-MD.6 (+) Use probabilities to make fair decisions (***e.g.***, drawing by lots, using a random number generator)
			LEVEL 6	**S-MD.4 (+)** Develop a probability distribution for a random variable defined for a sample space in which probabilities are assigned empirically; find the expected value. For example, find a current data distribution on the number of TV sets per household in the United States, and calculate the expected number of sets per household. How many TV sets would you expect to find in 100 randomly selected households? **S-MD.7 (+)** Analyze decisions and strategies using probability concepts (e.g., product testing, medical testing, pulling a hockey goalie at the end of a game).
			LEVEL 7	**N/A**
			LEVEL 8	**N/A**

STATISTICAL INVESTIGATIONS AND SAMPLING, DESCRIPTIVE STATISTICS (CENTRAL TENDENCY AND DISTRIBUTION), BIVARIATE DATA AND SCATTERPLOTS AND PROBABILITY (GRADES 6 – 8)

SAMPLING AND DESIGN, DATA DISTRIBUTIONS, PROBABILITY AND BIVARIATE DATA, LINEAR REGRESSION, AND CORRELATION (HIGH SCHOOL)

<table>
<tr><th>GRADE 6</th><th>GRADE 7</th><th>GRADE 8</th><th colspan="2">HIGH SCHOOL</th></tr>
<tr><td colspan="3">BIVARIATE DATA AND SCATTERPLOTS</td><td colspan="2">BIVARIATE DATA, LINEAR REGRESSION, AND CORRELATION</td></tr>
<tr><td rowspan="8">N/A</td><td rowspan="8">N/A</td><td rowspan="8">8.SP.1 Construct and interpret scatterplots for bivariate measurement data to investigate patterns of association between two quantities. Describe patterns such as clustering, outliers, positive or negative association, linear association, and nonlinear association.

8.SP.2 Know that straight lines are widely used to model relationships between two quantitative variables. For scatterplots that suggest a linear association, informally fit a straight line, and informally assess the model fit by judging the closeness of the data points to the line.

8.SP.3 Use the equation of a linear model to solve problems in the context of bivariate measurement data, interpreting the slope and intercept. For example, in a linear model for a biology experiment, interpret a slope of 1.5 cm/hr as meaning that an additional hour of sunlight each day is associated with an additional 1.5 cm in mature plant height.

8.SP.4 Understand that patterns of association can also be seen in bivariate categorical data by displaying frequencies and relative frequencies in a two-way table. Construct and interpret a two-way table summarizing data on two categorical variables collected from the same subjects. Use relative frequencies calculated for rows or columns to describe possible association between the two variables. For example, collect data from students in your class on whether or not they have a curfew on school nights and whether or not they have assigned chores at home. Is there evidence that those who have a curfew also tend to have chores?</td><td>LEVEL 1</td><td>S-ID.5 Summarize categorical data for two categories in two-way frequency tables. Interpret relative frequencies in the context of the data (including joint, marginal, and conditional relative frequencies). Recognize possible associations and trends in the data.</td></tr>
<tr><td>LEVEL 2</td><td rowspan="3">S-ID.6abc (Levels 2 – 4) Represent data on two quantitative variables on a scatter plot, and describe how the variables are related. a. Fit a function to the data; use functions fitted to data to solve problems in the context of the data. Use given functions or choose a function suggested by the context. Emphasize linear, quadratic, and exponential models. b. Informally assess the fit of a function by plotting and analyzing residuals. c. Fit a linear function for a scatter plot that suggests a linear association.

S-ID 7 (Level 2 only) Interpret the slope (rate of change) and the intercept (constant term) of a linear model in the context of the data.

S-ID.8 (Level 4 only) Compute (using technology) and interpret the correlation coefficient of a linear fit.</td></tr>
<tr><td>LEVEL 3</td></tr>
<tr><td>LEVEL 4</td></tr>
<tr><td>LEVEL 5</td><td>S-ID.9 Distinguish between correlation and causation.</td></tr>
<tr><td>LEVEL 6</td><td>N/A</td></tr>
<tr><td>LEVEL 7</td><td>N/A</td></tr>
<tr><td>LEVEL 8</td><td>N/A</td></tr>
</table>

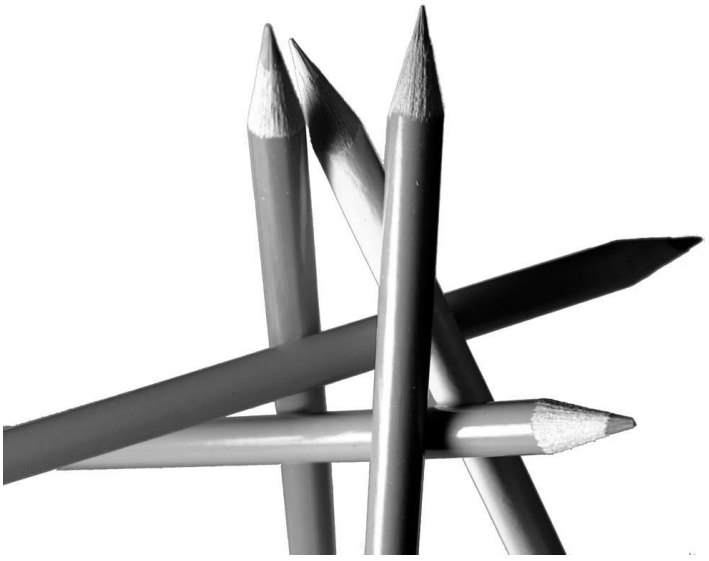
(+) indicates topics students should learn in order to take advanced mathematics courses

APPENDIX B

SURVEY INSTRUMENT FORM 1

TEXAS STATE UNIVERSITY
SAN MARCOS

Identifying Developmental Levels and Learning Trajectories of Statistics for Grade 6-12 Students: A survey



@ 2012 Developed by Rini Oktavia
Advisor: Dr. Maria Alejandra Sorto
Department of Mathematics
Texas State University-San Marcos

October 2012,

Dear Students.

This survey is part of a dissertation project for pursuing a doctoral degree in mathematics education from Texas State University – San Marcos. Your response will help develop an instrument to identify students' learning trajectories in statistics.

The survey has 18 multiple-choice questions that assess students' knowledge and skills on several statistical ideas.

Thank you for taking the time to complete the test. Your participation is truly appreciated.

Sincerely,

Rini Oktavia

Doctoral student in Mathematics Education
Texas State University - San Marcos

Form 1

School Grade : ☐ Grade 6 ☐ Grade 7 ☐ Grade 8

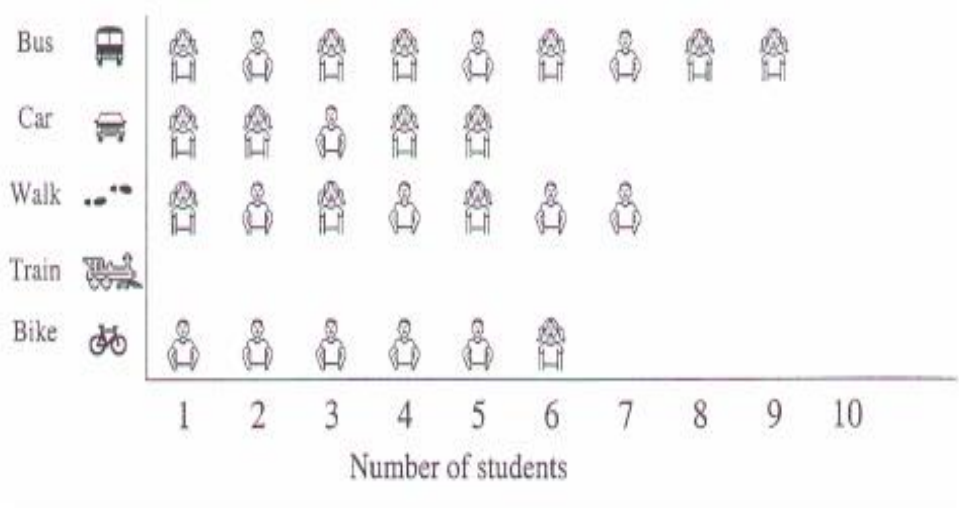
Age : ☐ 10 years old ☐ 11 years old ☐ 12 years old

☐ 13 years old ☐ 14 years old

Mathematics course currently taken: _____

Please circle one answer from several answer choices given in the following questions.

1. The following graph represents how children came to school one day.



How many children walk to school?

A. 9

B. 5

C. 7

D. 6

2. Box A and Box B are filled with red and blue marbles as follows. Each box is shaken. You want to get a blue marble, but you are only allowed to pick out one marble without looking. Which box should you choose, and why?

Box A	Box B
6 red 4 blue	60 red 40 blue

   A. Box B (with 60 red and 40 blue), because it contains more blue marbles.

   B. Box A (with 6 red and 4 blue), because the difference between the number of red and blue marbles is small.

   C. It doesn't matter, because Box B has ten times the amount in Box A.

   D. It doesn't matter, because both boxes have 40% blue marbles.

3. A city council wanted to estimate the proportion of residents of the city that would support an increase in taxes for education. A survey is conducted to ask residents whether they would support the increase tax or not. Of the following options, which data collection method will give **the most accurate** estimation?

   A. A sample is chosen by randomly select residents from the list of all residents of the city.

   B. A sample is chosen by randomly select residents from a certain area in the city.

   C. A sample is chosen randomly from government employees.

   D. A sample is chosen randomly from residents who have kids that are still in school.

4. A farmer wants to know how many fish there are in his dam. He took out 200 fish and tagged each of them, with a colored sign. He put the tagged fish back in the dam and let them get mixed with the others. On the second day, he took out 250 fish randomly and found that 25 of them were tagged. Estimate how many fish are in the dam.

    A. 250

    B. 500

    C. 1000

    D. 2000

5. Of the following questions, which one is a statistical question?
   Hint: A statistics question is a question that anticipates an answer based on data that vary.

    A. How tall is the tallest building in the world?

    B. How tall are adult men in the United States?

    C. How many students attend Miller Middle School in 2012?

    D. How many times a week do you practice soccer?

6. Which of the following sequences is **most likely** to result from flipping a fair coin five times?

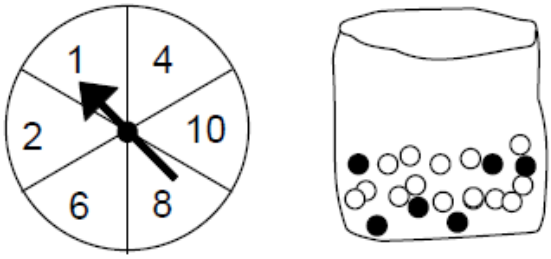
    A. H H H T T

    B. T H H T H

    C. T H T T T

    D. All three sequences are equally likely.

7. A game in a booth at a spring fair involves using a spinner first. Then, if the spinner stops on an even number, the player is allowed to pick a marble from a bag. The spinner and the marbles in the bag are represented in the diagram below.



Prizes are given when a black marble is picked. Sue plays the game once. How likely is it that Sue will win a prize?

A. Impossible.

B. Not very likely.

C. About 50% likely.

D. Very likely.

8. Of the following options, which one can be answered with a statistical investigation using Miller Middle School students' basic health information data?

A. What is the rate of obesity among students in the school?

B. Who is the tallest student in the school?

C. Is the overall health of middle school students declining in this country?

D. All of the above questions can be answered using statistical investigation.

9. The principal of Miller Middle School would like to study the feelings of students about the food served in the cafeteria. He plans to have college students, who are volunteering in the school, interview every 10th student who walks by the cafeteria between the hours of 11:00 am and 1:00 pm. Of the following statements on the strengths or weaknesses of this sampling plan, which would you recommend as **the most appropriate**?

    A. The survey is good, because every student has the same chance to be interviewed.

    B. The survey is not fair, because some groups of students might not have lunch in the cafeteria.

    C. The survey is good, because the students are interviewed randomly near the cafeteria.

    D. The survey is not fair, because most of the students who are interviewed are boys.

10. John is flipping a coin ten times. Tony is flipping a coin 100 times. Which one of the following options is appropriate to describe the possible outcomes that John and Tony get?

    A. It is impossible that John gets five heads and five tails and Tony gets 50 heads and 50 tails.

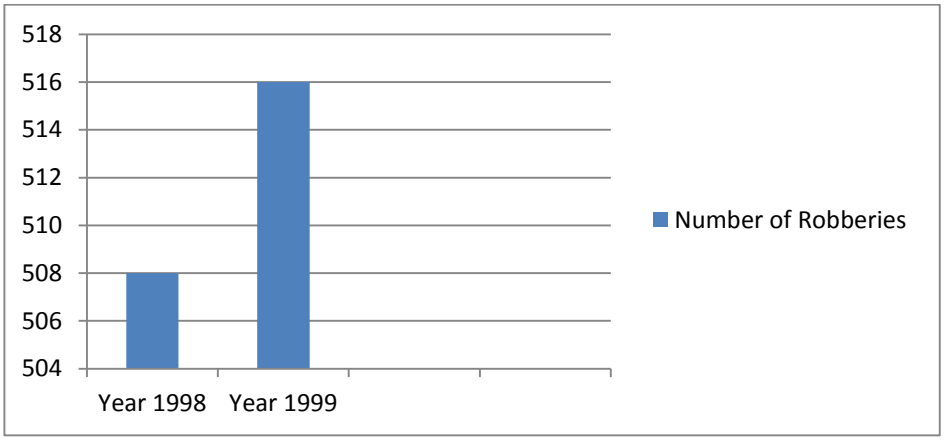
    B. It is less likely that Tony gets 50 heads and 50 tails rather than that John gets five heads and five tails.

    C. It is more likely that Tony gets 50 heads and 50 tails rather than that John gets five heads and five tails.

    D. It is equally likely that John gets 5 heads and 5 tails and Tony gets 50 heads and 50 tails.

11. A TV reporter showed this graph and said:

"The graph shows that there is a huge increase in the number of robberies from 1998 to 1999."



Of the following options, which one do you think to be **the most appropriate** answer for the following question? Do you consider the reporter's statement to be a reasonable interpretation of the graph? Why?

A. **Yes**, it is reasonable because the bar for Year 1999 is three times higher than the bar for 1998.

B. **No**, it is not reasonable because only a small part of the graph is shown; if the whole graph is shown, you would see that there is only a slight increase in robberies.

C. **No**, it is not reasonable because "huge" is not an appropriate term to describe the increasing number of robberies.

D. **Yes**, it is reasonable because robberies increases almost doubled from 1998 to 1999.

12. Mrs. Jones wants to buy a new car, either a **Honda** or a **Toyota**. She wants whichever car that will break down the least. She read in Consumer Reports that for 400 cars of each type, the **Toyota** had more breakdowns than the **Honda**. She talked to three friends. Two were **Toyota** owners, who had no major breakdowns. The other friend used to own a **Honda**, but it had lots of breakdowns, so he sold it. He said he'd never buy another **Honda**. Which car should Mrs. Jones buy?

    A. Mrs. Jones should buy the **Toyota**, because her friend had so much trouble with his Honda, while her other friends had no trouble with their Toyotas.

    B. She should buy the **Honda**, because the information about break-downs in Consumer Reports is based on many cases, not just one or two cases.

    C. It doesn't matter which car she buys. Whichever type she gets, she could still be unlucky and get stuck with a particular car that would need a lot of repairs.

    D. Mrs. Jones should **NOT** buy either the **Honda** or the **Toyota**, because both cars have major breakdowns history.

13. A bowl has 100 color candies in it. 20 are yellow, 50 are red, and 30 are blue. They are well mixed up in the bowl. Randy pulls out a handful of 10 candies, counts the number of reds, and records it on the board.

    Then, Randy puts the candies back into the bowl, and mixes them up again. Four of Randy's classmates, Renee, Ricky, Robby, and Rosie do the same thing. One at a time they pull ten candies, count the reds, and write down the number of reds, and put the candies back in the bowl and mix them up again.

    Which of the following lists for the number of reds is **most likely** to be?

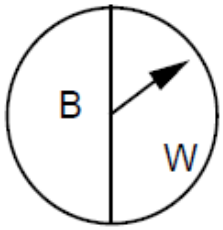
    A. 8, 9, 7, 10, 9

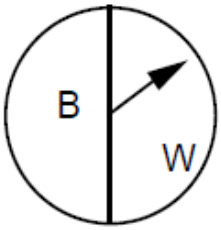
    B. 3, 7, 5, 8, 5

    C. 5, 5, 5, 5, 5

    D. 2, 4, 3, 4, 3

14. Two fair spinners (half black (B) and half white (W)) are part of a carnival game. A player wins a prize only when both arrows land on black (BB) after each spinner has been spun once.





Cody wanted to play the game; he thinks he has a 50-50 chance of winning. **Do you agree?**

Which of the following options is **the most appropriate** to answer the question above?

A. **No**, because if there were one spinner, the chance of winning it would be 50%, so it has to be less with two spinners.

B. **Yes**, because each spinner are half white and half black.

C. **No**, because there are four possible outcomes, BB, BW, WB, WW. So, the chance will be 25%.

D. **Yes**, because there are two spinners with the same areas of white and black.

15. Consider a situation involving two variables X and Y. What conditions would need to be satisfied in order to say that a change in the variable X causes a change in the variable Y?

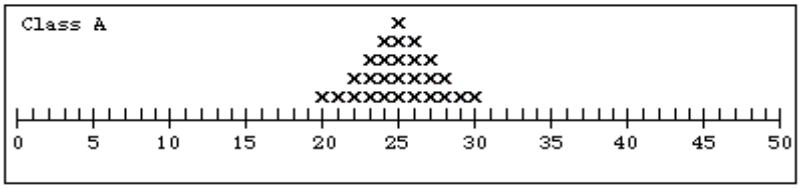
A. When the correlation between X and Y is close to 1 or -1.

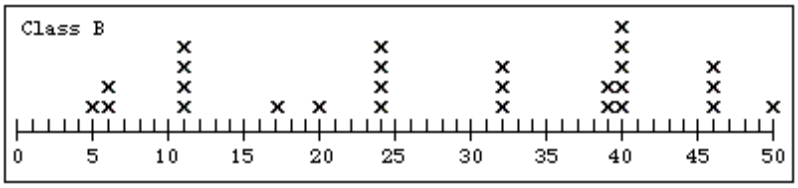
B. When an experiment reveals that a change in X causes a change in Y.

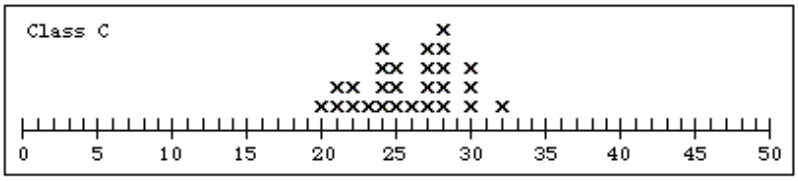
C. When possible confounding variables have been ruled out.

D. All of the above.

16. A class of students tossed 50 pennies and counted the number of heads. They repeated this many times. Imagine that two other classes produced graphs for the same experiment. In some cases, the results were just made up without actually doing the experiment.
The following dot plots show the results obtained by all classes.







Which of the results is more likely made up (not really from the experiment)?

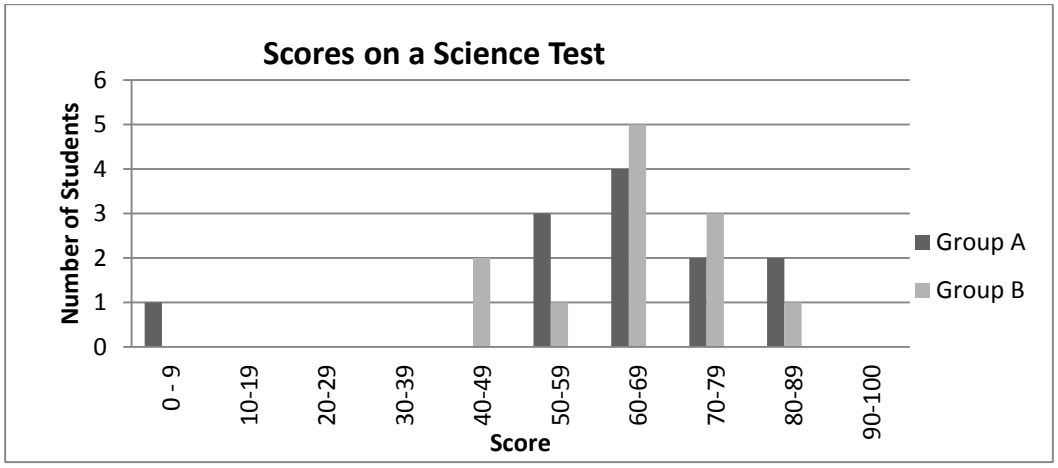
A. Class A's results

B. Class B's results

C. Class C's results

D. All results are more likely made up.

17. The diagram below shows the results on a science test for two groups, labeled as Group A and Group B. The mean score for group A is 62.0 and the mean for group B is 64.5. Students pass this test when their score is 50 or above.



Looking at the diagram, the teacher claims that group B did better than Group A in this test. The students in Group A don't agree with their teacher. They try to convince the teacher that Group B may not necessary have done better. Which of the following arguments is the **most appropriate** to be used by the students in Group A?

A. The scores of Group A have more variations than the scores of Group B.

B. More students in Group A than in Group B passed the test.

C. Group A has better score results in the 80-89 range and the 50-59 range.

D. The difference between the highest and lowest scores is smaller for Group B than for Group A.

18. The following information is from a survey about smoking and lung disease among 250 people.

	Lung disease	**No lung disease**	**Total**
Smoking	90	60	150
No smoking	60	40	100
Total	150	100	250

Using this information; of the following options, which do you think is **the most appropriate**?

A. Lung disease is associated with smoking, because the number of people who are smoking and have lung disease are bigger than the number of people who are smoking and do not have lung disease.

B. Lung disease is **NOT** associated with smoking, because the percentage of people who have lung disease and smoking and the people who have lung disease and not smoking are the same (0.6).

C. Lung disease is associated with smoking, because smoking is known to cause lung cancer.

D. Lung disease is **NOT** associated with smoking, because the number of people who are smoking and have no lung disease is the same as the number of people who are not smoking and have lung disease.

End of the survey. Thank you for taking the time to complete the survey.

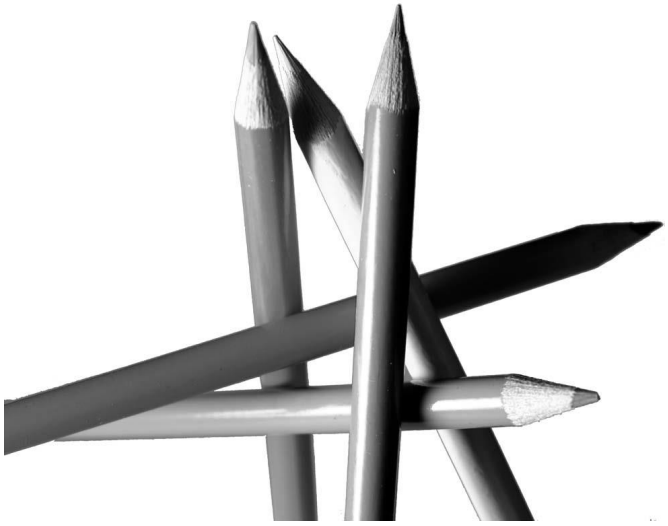
APPENDIX C

SURVEY INSTRUMENT FORM 2

TEXAS STATE UNIVERSITY
SAN MARCOS

Identifying Developmental Levels and Learning Trajectories of Statistics For Grade 6-12 Students: A survey

---



@ 2012 Developed by Rini Oktavia
Advisor: Dr. Maria Alejandra Sorto
Department of Mathematics
Texas State University-San Marcos

October 2012,

Dear Students.

This survey is a part of a dissertation project for pursuing a doctoral degree in mathematics education from Texas State University – San Marcos. Your response will help develop an instrument to identify students' learning trajectories in statistics.

The survey has 18 multiple-choice questions that assess students' knowledge and skills on several statistical ideas.

Thank you for taking the time to complete the test. Your participation is truly appreciated.

Sincerely,

Rini Oktavia

Doctoral student in Mathematics Education
Texas State University - San Marcos

Form 2

School Grade : ☐ Grade 6 ☐ Grade 7 ☐ Grade 8

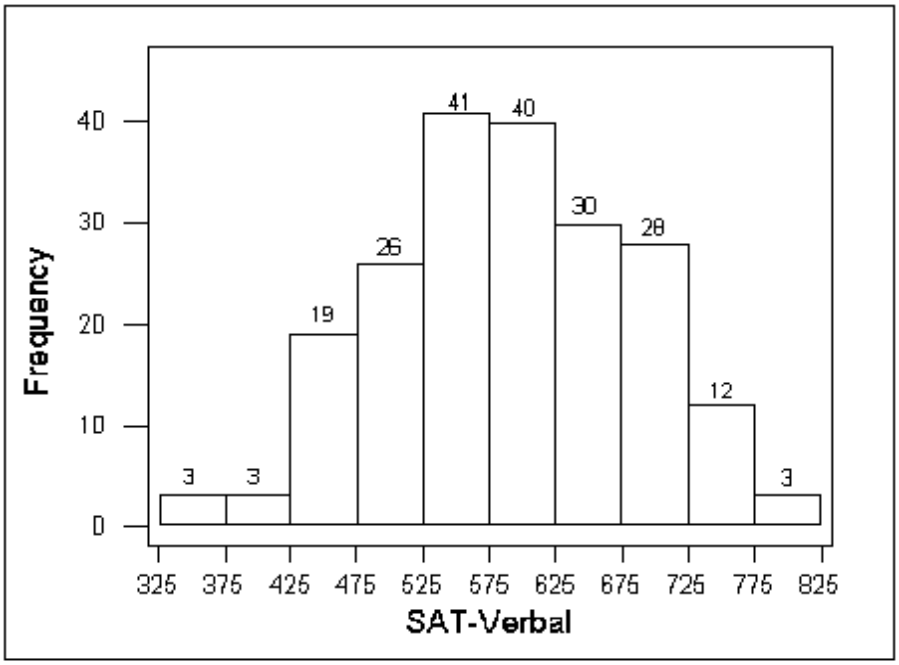
Age : ☐ 10 years old ☐ 11 years old ☐ 12 years old

☐ 13 years old ☐ 14 years old

Mathematics Course currently taken: _____

Please circle one answer from several answer choices given in the following questions.

1. The following histogram shows the Verbal SAT scores for 205 students entering a local college in the fall of 2002.



How many of the students had verbal SAT scores between 425 and 725?

A. 19

B. 28

C. 184

D. 139

2. A certain state lottery awards 18 $200 prizes, 120 $25 prizes and 270 $20 prizes, for every 10,000 tickets sold. Bob and Bill each bought one ticket each week for the past 100 weeks. Bill has not won a single prize yet. Bob just won a $20 prize last week. Who is more likely to win a prize this coming week? Select the best answer.

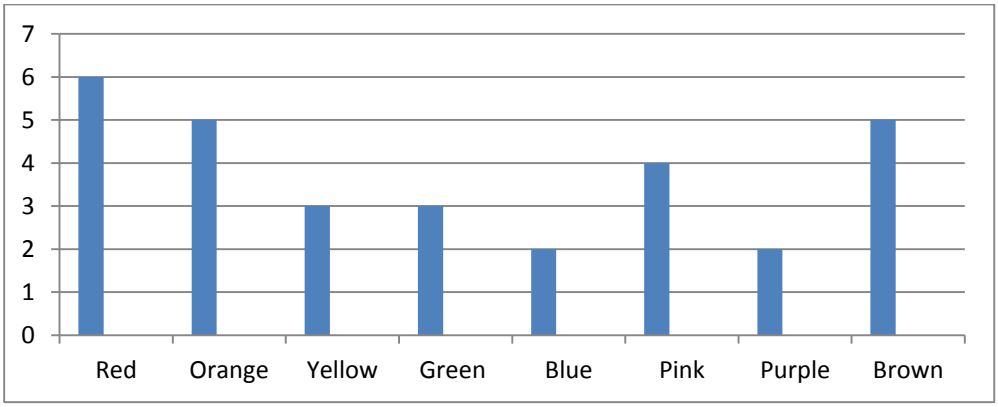
   A. Bill

   B. Bob

   C. They have an equal chance of winning

   D. Not enough information to tell

3. Robert's mother lets him pick one candy from a bag. He can't see the candies. The number of candies of each color in the bag is shown in the following graph.



What is the probability that Robert will pick a red candy?

   a. 10%

   b. 20%

   c. 25%

   d. 50%

4. If a fair coin is tossed, the probability that it will land heads up is 1/2. In four successive tosses, a fair coin lands heads up each time. What is likely to happen when the coin is tossed a fifth time?

    A. It is more likely to land tails up than heads up.

    B. It is more likely to land heads up than tails up.

    C. It is equally likely to land heads up or tails up.

    D. More information is needed to answer the question.

5. When three fair dice are simultaneously thrown, which of the following results is **MOST LIKELY** to be obtained?

    A. Result 1: A 5, a 3 and a 6 in any order

    B. Result 2: Three 5's

    C. Result 3: Two 5's and a 3

    D. All three results are equally likely.

6. A small object was weighed on the same scale separately by nine students in a science class. The weights (in grams) recorded by each student are shown below.

    6.3 6.0 6.0 15.3 6.1 6.3 6.2 6.15 6.3

    The students want to determine as accurately as they can the actual weight of this object. Of the following methods, which would you recommend they use?

    E. Use the most common value, which is 6.2.

    F. Use the 6.15 since it is the most accurate weighing.

    G. Add up the nine numbers and divide by 9.

    H. Throw out the 15.3, add up the other 8 numbers and divide by 8.

7. A class of students asks a question on who can jump farther, boys or girls. Of the following options which one do you think as the best way to collect data to answer the question?

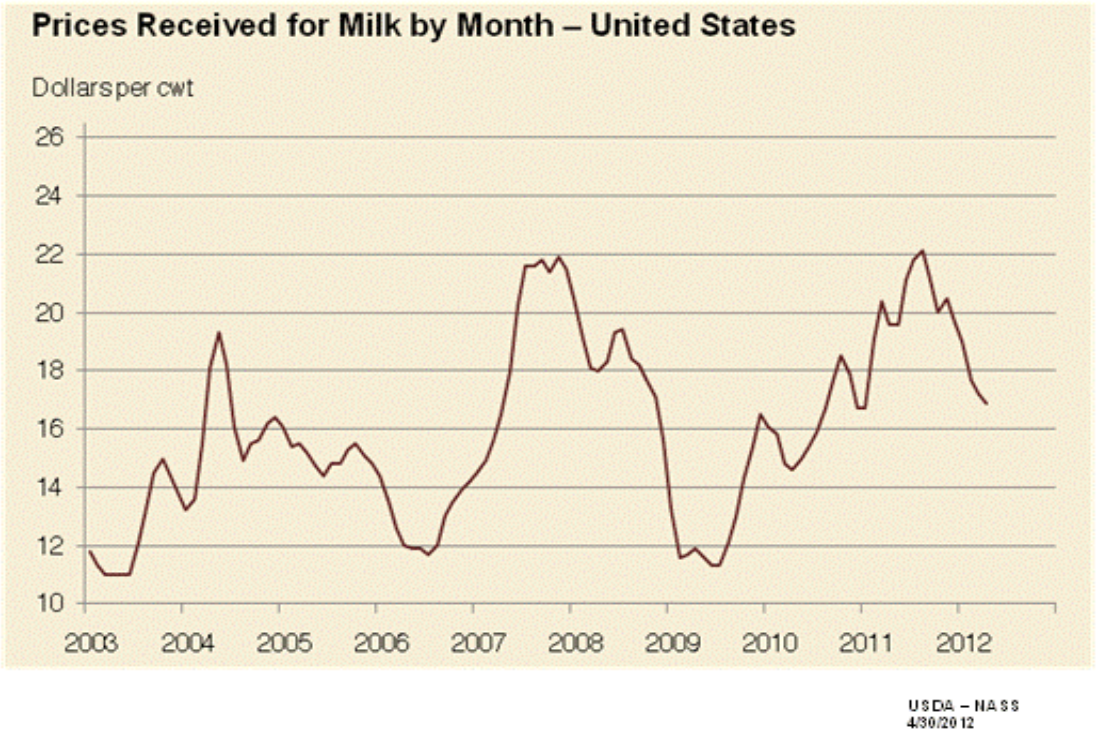
   A. Students measure the jumping distances for all of their classmates.

   B. Students measure the heights for all of their classmates.

   C. Students count how many boys and girls who can jump farther than four feet.

   D. Some students volunteer to jump and their jumping distances are measured.

8. The following graph provided by the U. S. Department of Agriculture shows the prices for milk by month in dollars per hundred pounds (cwt) in the U. S. since 2003 to 2012.



How much is the prices for milk in the middle of 2006?

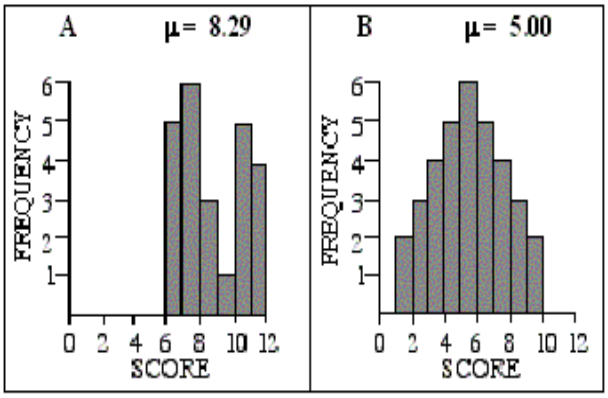
   A. 21.8 dollars per hundred pounds.

   B. 11.90 dollars per hundred pounds.

   C. 13.5 dollars per hundred pounds.

   D. 14.7 dollars per hundred pounds.

9. For each pair of graphs, determine which graph has the higher standard deviation (it is not necessary to do any calculations to answer these questions).



   A. A has a larger standard deviation than B

   B. B has a larger standard deviation than A

   C. Both graphs have the same standard deviation

   D. Cannot be determined

10. A sample of 50 students was taken from a large urban school with 1000 students and a sample of 20 students was taken from a small rural school with 300 students. Both schools have the same percentage of girls and boys. One of these samples was strange in that it had 80% boys. Which do you think is more likely?

    A. The sample is from the small school.

    B. The sample is from the large school.

    C. The sample could be from the large school or the small school.

    D. It is impossible to have a sample with 80% boys.

11. Four students at a local high school conducted surveys.

    Shannon got the names of all 800 children in the high school and put them in a hat, and then pulled out 60 of them.

    Jake asked 10 students at an after-school meeting of the computer games club.

    Adam asked all of the 200 children in Grade 10.

    Claire set up a booth outside of the school. Anyone who wanted to stop and fill out a survey could. She stopped collecting surveys when she got 60 students to complete them.

    Who do you think has the best sampling method? Why?

    A. **Adam**, because asking all Grade 10 students are a good way to get all possible opinions of all students in the school.

    B. **Jake**, because all the computer games club members are Jake's friends. Their answers are trustworthy.

    C. **Claire**, because every student has an opportunity to be interviewed before she gets 60 students.

    D. **Shannon**, because the participants are chosen randomly.

12. A town contains three elementary schools. School A has a mean class size of 30 pupils for its three fifth-grade classrooms. School B has a mean class size of 25 pupils in its two fifth-grade classrooms. School C has 20 pupils in its only fifth-grade classroom. What is the average class size for fifth-grade classrooms in this town?

    A. 12.5

    B. 25

    C. 26.7

    D. Cannot be determined

13. Grade 6 students in Goodnight Middle Schools will conduct an experiment to answer the question of whether beans grow faster in the dark or in the light. From the following options which one would be **the most appropriate** way to collect data to answer the question?

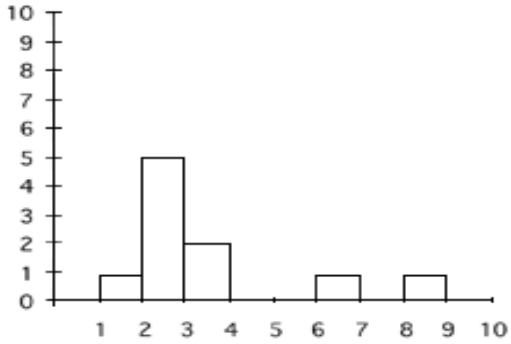
    A. Students plant the same number of dried beans in two large plant pots. They put one pot in the light and the other in the dark, and let the beans sprout. After two weeks, the number of growing plants in each pot is counted.

    B. Students plant the same number of dried beans in two large plant pots and put one pot in the light and the other in the dark, and let the beans sprout. After two weeks, the heights of the plants are measured.

    C. Students plant some dried beans in one large plant pot. They put the pot in the light and let the beans sprout. After two weeks, the heights of the plants are measured. The pot, then, is put in the dark, and after two weeks the plants are measured again.

    D. Students plant some dried beans in one large plant pot. They put the pot in the dark and let the beans sprout. After two weeks, the heights of the plants are measured. The pot, then, is put in the light, and after two weeks the plants are measured again.

14. Here is a histogram for a set of test scores from a 10-item makeup quiz given to a group of students who were absent on the day the quiz was given.



What do the numbers on the horizontal axis represent? Please select the best response from the list.

A. Scores on the test

B. Independent variable

C. Dependent variable

D. Number of Students

15. There are 20 students in a mathematics class; ten boys and ten girls. The teacher will choose three students randomly and ask them to show their work on the board. A student can only be chosen once. After choosing two boys, what is the chance that the teacher will choose another boy to work on the board?

    A. Ten out of twenty.

    B. Eight out of twenty.

    C. Ten out of eighteen.

    D. Eight out of eighteen.

16. A group of 649 men with lung cancer was identified from a certain population in England. A control group about the same size was established by matching these patients with other men from the same population who did not have lung cancer. The matching was on background variables such as ethnicity, age, and socioeconomic status. The summary of level of smoking and the number of lung cancer and control cases is given in the following table.

Cigarettes /Day	**Lung Cancer Cases**	**Control**	**Probability of Lung Cancer**
0	2	27	2/29 = 0.07
1 - 14	283	346	283/629 = 0.45
15 - 24	196	190	196/ 386 = 0.51
25 +	168	84	168/252 = 0.67

What is the association between the level of smoking and the number of lung cancer cases that can be inferred by the given data?

A. A **decrease** in the lung cancer rate is associated with an **increase** in cigarette smoking.

B. An **increase** in the lung cancer rate is associated with an **increase** in cigarette smoking.

C. An **increase** in the lung cancer rate is associated with a **decrease** in cigarette smoking.

D. There is **no association** between the level of smoking and the number of lung cancer cases.

17. The following table summarizes the data on a survey that ask the following questions. “Do you like rock music?” and “Do you like rap music?”

    The participants are randomly selected from all middle school students in San Marcos, TX.

		Like Rock Music?		
		Yes	*No*	***Row total***
Like Rap Music?	*Yes*	*25*	*4*	***29***
	No	*6*	*15*	***21***
Column total		***31***	***19***	***50***

    Of the following options, which one is **the most accurate** explanation of the data represented in the table?

    A. There may be a strong association between liking Rock music and liking Rap music. However this association could simply be a consequence of a random sampling.

    B. There may **NOT** be a strong association between liking Rock music and liking Rap music. However this association could simply be a consequence of a random sampling.

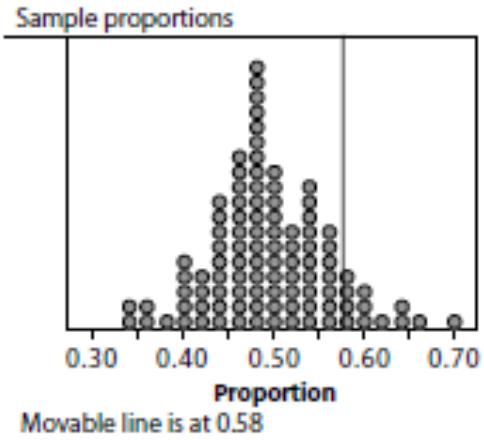
    C. More than 50% of San Marcos middle school students do **NOT** like Rap music. However this association could simply be a consequence of a random sampling.

    D. More than 50% of San Marcos middle school students do **NOT** like Rock music. However this association could simply be a consequence of a random sampling.

18. Among 50 students in a middle school who were randomly chosen to participate in a survey, fifty eight percent of the students like ice cream and 52% of the students like cakes for dessert.

    It is claimed that more than 50% of students in the middle school like ice cream. To simulate the situation, a computer generates a set of even and odd digits to represent students who like ice cream and who do not like ice cream respectively. Samples with size 50 are randomly chosen repeatedly from the set of digits.

    The number of even digits from each sample is counted and the proportion of even digits from each sample is recorded. After 100 simulations the sampling distribution is represented by the following graph.



    Based on this simulation, a sample proportion greater than or equal to the observed 0.58 occurred 12 times of 100 just by chance variation alone when the actual population proportion is 0.5. What is suggested by this result?

    E. The claim that more than 50% of students in the middle school like ice cream is **NOT** supported by the evidence.

    F. The claim that more than 50% of students in the middle school like ice cream is supported by the evidence.

    G. The result that 58% of the students like ice cream is **NOT** likely due to chance alone.

    H. The result that 58% of the students who were interviewed like ice cream is **NOT** supported by the evidence.

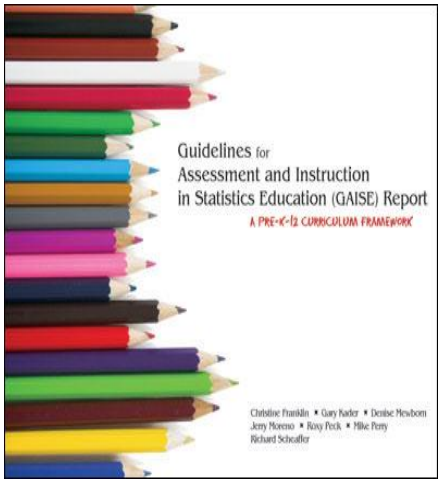
End of the survey. Thank you for taking the time to complete the survey.

APPENDIX D

EXPERT SURVEY INSTRUMENT FORM 1

TEXAS STATE UNIVERSITY
SAN MARCOS

Expert Survey to Measure the Alignment between Assessment Items and Pre-K-12 GAISE Levels



Generated by Rini Oktavia
Advisor: Dr. Maria Alejandra Sorto
Department of Mathematics
Texas State University-San Marcos

July 11, 2013

Dear Experts,

This survey is a part of a dissertation project for pursuing a doctoral degree in mathematics education at Texas State University – San Marcos. Your response will help develop an instrument to identify students' learning trajectory in statistics based on the Pre-K-12 Guidelines for Assessment and Instruction in Statistics Education (GAISE) Framework and the Common Core State Standards in Mathematics (CCSS-M).

The survey asks your judgment about how 11 content items align with the developmental level of statistics education suggested by the Pre-K-12 GAISE Framework (see page 3 and 4). Your participation in this study is completely voluntary and you can refuse to answer any questions. You may choose to discontinue completing it at any point. If you choose to discontinue, the information collected will not be used in this project.

There is no direct benefit to you for completing the survey. However, your response will help develop an important tool for identifying students' learning trajectories in Statistics. Hence, your response is critically important. There is no risk to you beyond that associated with the completion of a survey.

All responses will be treated with the utmost confidentiality and your privacy will be protected to the maximum extent allowable by law. To ensure confidentiality, your name will not be associated with your responses and no identifying information will be reported in any way unless you give permission to do so.

If you have questions about this survey or about your participation in this research project, you may contact Rini Oktavia by phone at 512-245-4747 or by e-mail at ro1088@txstate.edu.

Thank you for taking the time to complete the survey.

Sincerely,

Rini Oktavia

Doctoral candidate in Mathematics Education

DEPARTMENT OF MATHEMATICS

601 University Drive |San Marcos, Texas 78666-4616 | *phone:* 512.245.2551 | *fax:* 512.245.3425 |
WWW.MATH.TXSTATE.EDU/MATH

Table 1.
Pre-K-12 GAISE Framework (*Franklin et al., 2005)*

Process Component	Level A	Level B	Level C
Formulate Question	Beginning awareness of the *statistics question distinction* Teachers pose questions of interest Questions restricted to classroom	Increase awareness of the *statistics question distinction* Students begin to pose their own questions of interest Question not restricted to classroom	Students can make the *statistics question distinction* Students pose their own questions of interest Questions seek generalization
Collect Data	Do not yet design for differences Census of classroom Simple experiment	Beginning awareness of design for differences Sample surveys; begin to use random selection Comparative experiment; begin to use random allocation	Students make design for differences Sampling designs with random selection Experimental designs with randomization
Analyze Data	Use particular properties of distributions in the context of a specific example Display variability within a group Compare individual to individual Compare individual to group Beginning awareness of group to group Observe association between two variables	Learn to use particular properties of distributions as tools of analysis Quantify variability within a group Compare group to group in displays Acknowledge sampling error Some quantification of association; simple models for association	Understand and use distributions in analysis as a global concept Measure variability within a group; measure variability between groups Compare group to group using displays and measures of variability Describe and quantify sampling error Quantification of association; fitting of models for association

Table 1 continued

Process Component	Level A	Level B	Level C
Interpret Results	Students do not look beyond the data No generalization beyond the classroom Note difference between two individuals with different conditions Observe association in displays	Students acknowledge that looking beyond the data is feasible Acknowledge that a sample may or may not be representative of the larger population Note the difference between two groups with different conditions Aware of distinction between observational study and experiment Note differences in strength of association Basic interpretation of models for association Aware of the distinction between association and cause and effect	Students are able to look beyond the data in some contexts Generalize from sample to population Aware of the effect of randomization on the results of experiments Understand the difference between observational studies and experiments Interpret measures of strength of association Interpret models of association Distinguish between conclusions from association studies and experiments
Nature of Variability	Measurement variability Natural variability Induced variability	Sampling variability	Chance variability
Focus on variability	Variability within a group	Variability within a group and variability between groups Covariability	Variability in model fitting

Note. The Pre-K-12 GAISE Framework. Reprinted from "*Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A Pre-K-12 Curriculum Framework*," by C. Franklin, G. Kader, D. Mewborn, J. Moreno, R. Peck, M. Perry, & R. Scheaffer, 2005, Copyright 2005 by the Joint American Statistics Association/ National Council of Teachers of Mathematics. Adapted with permission.

1. Consider the following multiple-choice item.

> *There are 20 students in a mathematics class; ten boys and ten girls. The teacher will choose three students randomly and ask them to show their work on the board. A student can only be chosen once. After choosing two boys, what is the chance that the teacher will choose another boy to work on the board?*
>
> *A. Ten out of twenty.*
> *B. Eight out of twenty.*
> *C. Ten out of eighteen.*
> D. ***Eight out of eighteen***.

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

Comments:

_____

_____

2. Consider the following multiple-choice item.

> *When three fair dice are simultaneously thrown, which of the following results is MOST LIKELY to be obtained?*
> ***E. Result 1: A 5, a 3 and a 6 in any order***
> *F. Result 2: Three 5's*
> *G. Result 3: Two 5's and a 3*
> *H. All three results are equally likely.*
>
> (Adapted from the NSF-funded Web Assessment Resource Tools for Improving Statistical Thinking (ARTIST) Project).

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

Comments:

_____

_____

3. Consider the following multiple-choice item.

> *John is flipping a coin ten times. Tony is flipping a coin 100 times. Which one of the following options is appropriate to describe the possible outcomes that John and Tony get?*
>
> *A. It is impossible that John gets five heads and five tails and Tony gets 50 heads and 50 tails.*
>
> *B. It is less likely that Tony gets 50 heads and 50 tails rather than that John gets five heads and five tails.*
>
> ***C. It is more likely that Tony gets 50 heads and 50 tails rather than that John gets five heads and five tails.***
>
> *D. It is equally likely that John gets 5 heads and 5 tails and Tony gets 50 heads and 50 tails.*

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

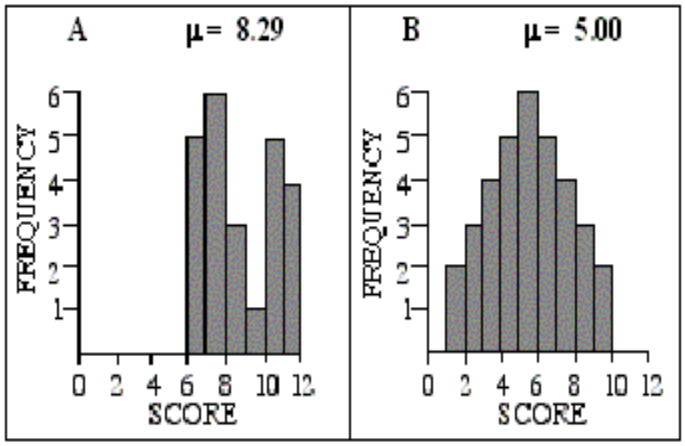
Comments:

_____

_____

4. Consider the following multiple-choice item.

For each pair of graphs, determine which graph has the higher standard deviation (it is not necessary to do any calculations to answer these questions).



E. A has a larger standard deviation than B.

F. B has a larger standard deviation than A.

G. Both graphs have the same standard deviation.

H. Cannot be determined.

(Adapted from the NSF-funded Web Assessment Resource Tools for Improving Statistical Thinking (ARTIST) Project).

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

Comments:

_____

_____

5. Consider the following multiple-choice item.

> *Of the following options, which one can be answered with a statistical investigation using Miller Middle School students' basic health information data?*
>
> ***A. What is the rate of obesity among students in the school?***
> *B. Who is the tallest student in the school?*
> *C. Is the overall health of middle school students declining in this country?*
> D. *All of the above questions can be answered using statistical investigation.*

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

Comments:

_____

_____

6. Consider the following multiple-choice item.

> *A farmer wants to know how many fish there are in his dam. He took out 200 fish and tagged each of them, with a colored sign. He put the tagged fish back in the dam and let them get mixed with the others. On the second day, he took out 250 fish randomly and found that 25 of them were tagged. Estimate how many fish are in the dam.*
>
> *A. 250*
> *B. 500*
> *C. 1000*
> ***D. 2000***
>
> (Adapted from Watson & Callingham, 2003).

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

Comments:

_____

_____

7. Consider the following multiple-choice item.

> *A small object was weighed on the same scale separately by nine students in a science class. The weights (in grams) recorded by each student are shown below.*
>
> *6.3 6.0 6.0 15.3 6.1 6.3 6.2 6.15 6.3*
>
> Of the following methods, which would you recommend they use?
>
> A. Use the most common value, which is 6.2.
> B. Use the 6.15 since it is the most accurate weighing.
> C. Add up the nine numbers and divide by 9.
> **D. Throw out the 15.3, add up the other 8 numbers and divide by 8.**
>
> (Adapted from Statistical Reasoning Assessment (SRA), Garfield 2003)

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

Comments:

_____

_____

8. Consider the following multiple-choice item.

Mrs. Jones wants to buy a new car, either a Honda or a Toyota. She wants whichever car will break down the least. She read in Consumer Reports that for 400 cars of each type, the Toyota had more breakdowns than the Honda.
She talked to three friends. Two were Toyota owners, who had no major breakdowns. The other friend used to own a Honda, but it had lots of breakdowns, so he sold it. He said he'd never buy another Honda.
Which car should Mrs. Jones buy?

A. *Mrs. Jones should buy the Toyota, because her friend had so much trouble with his Honda, while her other friends had no trouble with their Toyotas.*

B. ***She should buy the Honda, because the information about break-downs in consumer Reports is based on many cases, not just one or two cases.***

C. *It doesn't matter which car she buys. Whichever type she gets, she could still be unlucky and get stuck with a particular car that would need a lot of repairs.*

D. *Mrs. Jones should not buy either the Honda or the Toyota, because both cars have major breakdowns history.*

(Adapted from Callingham & Watson, 2005)

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

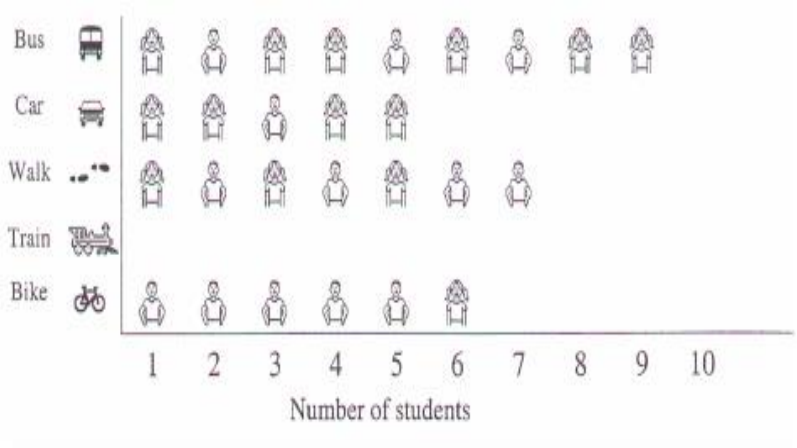
Comments:

_____

_____

9. Consider the following multiple-choice item.

How children came to school one day



How many children walk to school?

A. 9
B. 5
C. 7
D. 6

(Adapted from Callingham & Watson, 2005)

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

Comments:

_____

_____

10. Consider the following multiple-choice item.

A class of students asks a question on who can jump farther, boys or girls. Of the following options which one do you think as the best way to collect data to answer the question?

A. Students measure the jumping distances for all of their classmates.
B. Students measure the heights for all of their classmates.
C. Students count how many boys and girls who can jump farther than four feet.
D. *Some students volunteer to jump and their jumping distances are measured.*

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

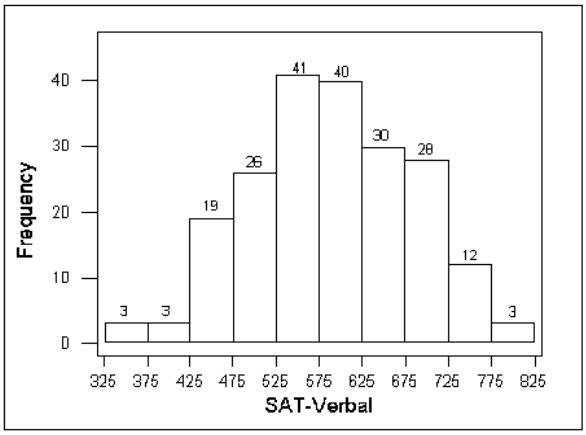
Comments:

_____

_____

11. Consider the following multiple-choice item.

The following histogram shows the Verbal SAT scores for 205 students entering a local college in the fall of 2002.



How many of the students had verbal SAT scores between 425 and 725?

E. 19

F. 28

G. 184

H. 139

(Adapted from the NSF-funded Web Assessment Resource Tools for Improving Statistical Thinking (ARTIST) Project).

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

Comments:

_____

_____

12. Consider the following multiple-choice item.

A city council wanted to estimate the proportion of residents of the city that would support an increase in taxes for education. A survey is conducted to ask residents whether they would support the increase tax or not. Of the following options, which data collection method will give the most accurate estimation?

A. ***A sample is chosen by randomly select residents from the list of all residents of the city.***

B. A sample is chosen by randomly select residents from a certain area in the city.

C. A sample is chosen randomly from government employees.

D. A sample is chosen randomly from residents who have kids that are still in school.

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

Comments:

_____

_____

References

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., and Scheaffer, R. (2007). *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A Pre-K-12 Curriculum Framework.* Alexandria, VA: American Statistical Association.

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

Watson, J.M. & Callingham, R.A. 2003. Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal*, 2(2), 3-46.

Watson, J.M., & Callingham, R.A. 2004. Statistical literacy: From idiosyncratic to critical thinking. Paper presented at the International Association for *Statistics Education Roundtable on "Curricular Development in Statistics Education,"* Lund, Sweden.

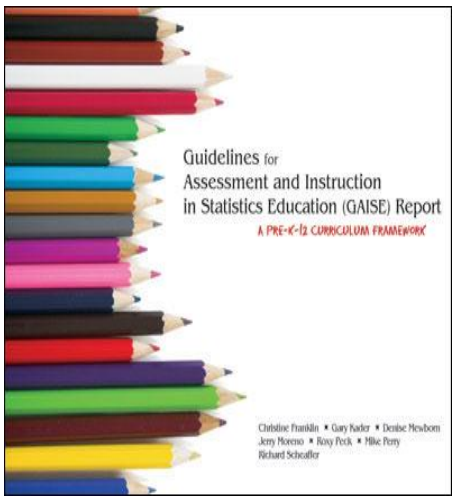
Watson, J.M. & Callingham, R.A. 2005. Measuring statistical literacy. *Journal of Applied Measurement*, 6(1), 19-47.

APPENDIX E

EXPERT SURVEY INSTRUMENT FORM 2

TEXAS STATE UNIVERSITY
SAN MARCOS

Expert Survey to Measure the Alignment between Assessment Items and Pre-K-12 GAISE Levels



@ 2012Generated by Rini Oktavia
Advisor: Dr. Maria Alejandra Sorto
Department of Mathematics
Texas State University-San Marcos

July 11, 2013

Dear Experts,

This survey is a part of a dissertation project for pursuing a doctoral degree in mathematics education at Texas State University – San Marcos. Your response will help develop an instrument to identify students' learning trajectory in statistics based on the Pre-K-12 Guidelines for Assessment and Instruction in Statistics Education (GAISE) Framework and the Common Core State Standards in Mathematics (CCSS-M).

The survey asks your judgment about how 11 content items align with the developmental level of statistics education suggested by the Pre-K-12 GAISE Framework (see page 3 and 4). Your participation in this study is completely voluntary and you can refuse to answer any questions. You may choose to discontinue completing it at any point. If you choose to discontinue, the information collected will not be used in this project.

There is no direct benefit to you for completing the survey. However, your response will help develop an important tool for identifying students' learning trajectories in Statistics. Hence, your response is critically important. There is no risk to you beyond that associated with the completion of a survey.

All responses will be treated with the utmost confidentiality and your privacy will be protected to the maximum extent allowable by law. To ensure confidentiality, your name will not be associated with your responses and no identifying information will be reported in any way unless you give permission to do so.

If you have questions about this survey or about your participation in this research project, you may contact Rini Oktavia by phone at 512-245-4747 or by e-mail at ro1088@txstate.edu.

Thank you for taking the time to complete the survey.

Sincerely,

Rini Oktavia

Doctoral candidate in Mathematics Education

DEPARTMENT OF MATHEMATICS

601 University Drive | San Marcos, Texas 78666-4616 | *phone:* 512.245.2551 | *fax:* 512.245.3425 |

WWW.MATH.TXSTATE.EDU/MATH

Table 1.
Pre-K-12 GAISE Framework (*Franklin et al., 2005)*

Process Component	Level A	Level B	Level C
Formulate Question	Beginning awareness of the *statistics question distinction* Teachers pose questions of interest Questions restricted to classroom	Increase awareness of the *statistics question distinction* Students begin to pose their own questions of interest Question not restricted to classroom	Students can make the *statistics question distinction* Students pose their own questions of interest Questions seek generalization
Collect Data	Do not yet design for differences Census of classroom Simple experiment	Beginning awareness of design for differences Sample surveys; begin to use random selection Comparative experiment; begin to use random allocation	Students make design for differences Sampling designs with random selection Experimental designs with randomization
Analyze Data	Use particular properties of distributions in the context of a specific example Display variability within a group Compare individual to individual Compare individual to group Beginning awareness of group to group Observe association between two variables	Learn to use particular properties of distributions as tools of analysis Quantify variability within a group Compare group to group in displays Acknowledge sampling error Some quantification of association; simple models for association	Understand and use distributions in analysis as a global concept Measure variability within a group; measure variability between groups Compare group to group using displays and measures of variability Describe and quantify sampling error Quantification of association; fitting of models for association

Table 1 continued

Process Component	Level A	Level B	Level C
Interpret Results	Students do not look beyond the data No generalization beyond the classroom Note difference between two individuals with different conditions Observe association in displays	Students acknowledge that looking beyond the data is feasible Acknowledge that a sample may or may not be representative of the larger population Note the difference between two groups with different conditions Aware of distinction between observational study and experiment Note differences in strength of association Basic interpretation of models for association Aware of the distinction between association and cause and effect	Students are able to look beyond the data in some contexts Generalize from sample to population Aware of the effect of randomization on the results of experiments Understand the difference between observational studies and experiments Interpret measures of strength of association Interpret models of association Distinguish between conclusions from association studies and experiments
Nature of Variability	Measurement variability Natural variability Induced variability	Sampling variability	Chance variability
Focus on variability	Variability within a group	Variability within a group and variability between groups Covariability	Variability in model fitting

Note. The Pre-K-12 GAISE Framework. Reprinted from "*Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A Pre-K-12 Curriculum Framework*," by C. Franklin, G. Kader, D. Mewborn, J. Moreno, R. Peck, M. Perry, & R. Scheaffer, 2005, Copyright 2005 by the Joint American Statistics Association/ National Council of Teachers of Mathematics. Adapted with permission.

1. Consider the following multiple-choice item.

> *A sample of 50 students was taken from a large urban school with 1000 students and a sample of 20 students was taken from a small rural school with 300 students. Both schools have the same percentage of girls (50 %) and boys (50%). One of these samples was strange in that it had 80% boys. Which do you think is more likely?*
>
> ***A. The sample is from the small school.***
> *B. The sample is from the large school.*
> *C. The sample could be from the large school or the small school.*
> D. *It is impossible to have a sample with 80% boys.*

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

Comments:

_____

_____

Consider the following multiple-choice item.

> *Of the following questions, which one is a statistical question?*
> ***Hint:*** *A statistical question is a question that anticipates an answer based on data that vary.*
>
> *A. How tall is the tallest building in the world?*
> ***B. How tall are adult men in the United States?***
> *C. How many students attend Miller Middle School in 2012?*
> D. *How many times a week do you practice soccer?*

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

Comments:

_____

_____

2. Consider the following multiple-choice item.

A group of 649 men with lung cancer was identified from a certain population in England. A control group about the same size was established by matching these patients with other men from the same population who did not have lung cancer. The matching was on background variables such as ethnicity, age, and socioeconomic status. The summary of level of smoking and the number of lung cancer and control cases is given in the following table.

Cigarettes/Day	***Lung Cancer Cases***	***Control***	***Probability of Lung Cancer***
0	*2*	*27*	*2/29 = 0.07*
1 - 14	*283*	*346*	*283/629 = 0.45*
15 - 24	*196*	*190*	*196/ 386 = 0.51*
25 +	*168*	*84*	*168/252 = 0.67*

What is the association between the level of smoking and the number of lung cancer cases that can be inferred by the given data?

A. A decrease in the lung cancer rate is associated with an increase in cigarette smoking.

B. An increase in the lung cancer rate is associated with an increase in cigarette smoking.

C. An increase in the lung cancer rate is associated with a decrease in cigarette smoking.

D. There is no association between the level of smoking and the number of lung cancer cases.

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

Comments:

_____

_____

3. Consider the following multiple-choice item.

> *The principal of Miller Middle School would like to study the feelings of students about the food served in the cafeteria. He plans to have college students, who are volunteering in the school, interview every 10th student who walks by the cafeteria between the hours of 11:00 am and 1:00 pm. Of the following statements on the strengths or weaknesses of this sampling plan, which would you recommend as the most appropriate?*
>
> *A. The survey is good, because every student has the same chance to be interviewed.*
>
> ***B. The survey is not fair, because some groups of students might not have lunch in the cafeteria.***
>
> *C. The survey is good, because the students are interviewed randomly near the cafeteria.*
>
> *D. The survey is not fair, because most of the students who are interviewed are boys.*

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

Comments:

_____

_____

4. Consider the following multiple-choice item.

Box A and Box B are filled with red and blue marbles as follows. Each box is shaken. You want to get a blue marble, but you are only allowed to pick out one marble without looking. Which box should you choose, and why?

Box A	*Box B*
6 red	*60 red*
4 blue	*40 blue*

A. *Box B (with 60 red and 40 blue), because it contains more blue marbles.*
B. *Box A (with 6 red and 4 blue), because the difference between the number of red and blue marbles is small.*
C. *It doesn't matter, because Box B has ten times the amount in Box A.*
D. *It doesn't matter, because both boxes have 40% blue marbles.*

(Adapted from Watson & Callingham, 2003).

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

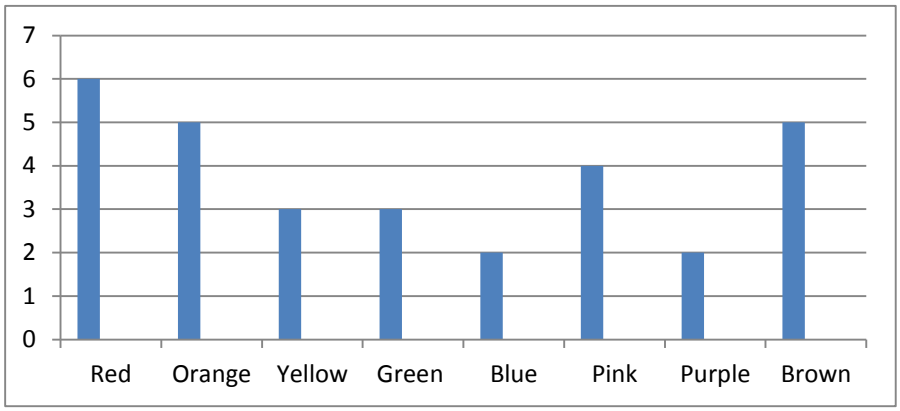
Comments:

_____

_____

5. Consider the following multiple-choice item.

Robert's mother lets him pick one candy from a bag. He can't see the candies. The number of candies of each color in the bag is shown in the following graph.



What is the probability that Robert will pick a red candy?

A. 10%
B. 20%
C. 25%
D. *50%*

(Adapted from PISA Assessment 2009, OECD, 2009).

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

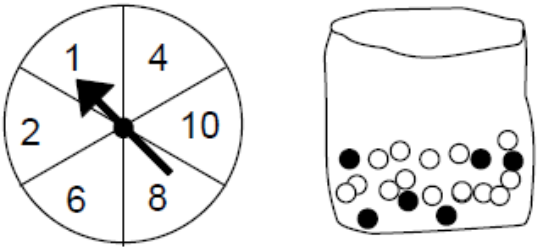
Comments:

_____

_____

6. Consider the following multiple-choice item.

A game in a booth at a spring fair involves using a spinner first. Then, if the spinner stops on an even number, the player is allowed to pick a marble from a bag. The spinner and the marbles in the bag are represented in the diagram below.



Prizes are given when a black marble is picked. Sue plays the game once. How likely it is that Sue will win a prize?

A. Impossible.
B. Not very likely.
C. About 50% likely.
D. Very likely.

(Adapted from PISA Assessment 2009, OECD, 2009).

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

Comments:

_____

_____

7. Consider the following multiple-choice item.

The following information is from a survey about smoking and lung disease among 250 people.

	Lung disease	***No lung disease***	***Total***
Smoking	*90*	*60*	*150*
No smoking	*60*	*40*	*100*
Total	*150*	*100*	*250*

Using this information; of the following options, which do you think is the most appropriate?

A. Yes, lung disease is associated with smoking, because the number of people who are smoking and have lung disease are bigger than the number of people who are smoking and do not have lung disease.

B. No, lung disease is not associated with smoking, because the percentage of people who have lung disease and smoking and the people who have lung disease and not smoking are the same (0.6).

C. Yes, lung disease is associated with smoking, because smoking is known to cause lung cancer.

D. No, lung disease is not associated with smoking, because the number of people who are smoking and have no lung disease is the same as the number of people who are not smoking and have lung disease.

(Adapted from Watson & Callingham, 2004).

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

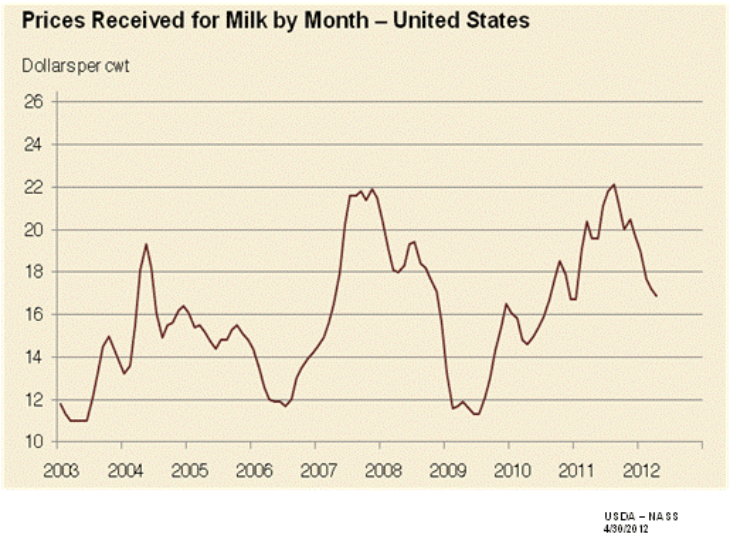
Comments:

_____

_____

8. Consider the following multiple-choice item.

The following graph shows the prices for milk by month in dollars per hundred pounds (dollars per cwt) in the United States since 2003 to 2012.



How much is the prices for milk in the middle of 2006?

A. *21.8 dollars per hundred pounds (dollars per cwt).*
B. *11.90 dollars per hundred pounds (dollars per cwt).*
C. *16.10 dollars per hundred pounds (dollars per cwt).*
D. *14.5 dollars per hundred pounds (dollars per cwt).*

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

Comments:

_____

_____

9. Consider the following multiple-choice item.

> *Four students at a local high school conducted surveys. Shannon got the names of all 800 students in the high school and put them in a hat, and then pulled out 60 of them. Jake asked 10 students at an after-school meeting of the computer games club. Adam asked all of the 200 students in Grade 10. Claire set up a booth outside of the school. Anyone who wanted to stop and fill out a survey could. She stopped collecting surveys when she got 60 students to complete them. Who do you think has the best sampling method? Why?*
>
> *A. Adam, because asking all Grade 10 students are a good way to get all possible opinions of all students in the school.*
>
> *B. Jake, because all the computer games club members are Jake's friends. Their answers are trustworthy.*
>
> *C. Claire, because every student has an opportunity to be interviewed before she gets 60 students.*
>
> ***D. Shannon, because the participants are chosen randomly.***
>
> (Adapted from the NSF-funded Web Assessment Resource Tools for Improving Statistical Thinking (ARTIST) Project).

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

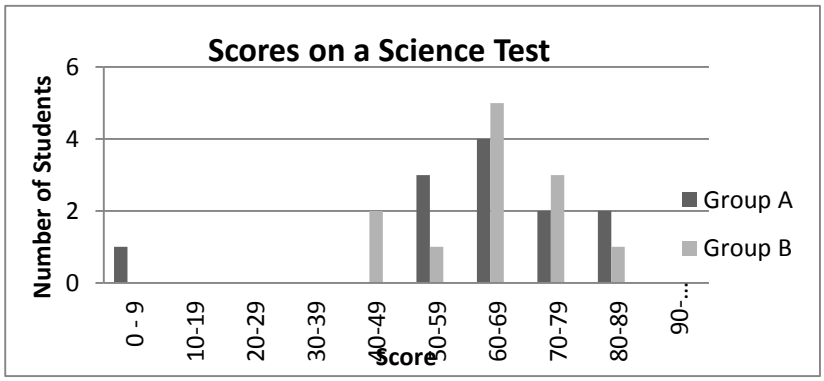
Comments:

_____

_____

10. Consider the following multiple-choice item.

The diagram below shows the results on a science test for two groups, labeled as Group A and Group B. The mean score for group A is 62.0 and the mean for group B is 64.5. Students pass this test when their score is 50 or above.



Looking at the diagram, the teacher claims that group B did better than Group A in this test. The students in Group A don't agree with their teacher. They try to convince the teacher that Group B may not necessary have done better. Which of the following arguments is the most appropriate to be used by the students in Group A?

A. The scores of Group A have more variations than the scores of Group B.

B. More students in Group A than in Group B passed the test.

C. Group A has better score results in the 80-89 range and the 50-59 range.

D. The difference between the highest and lowest scores is smaller for Group B than for Group A.

(Adapted from PISA Assessment 2009, OECD, 2009).

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

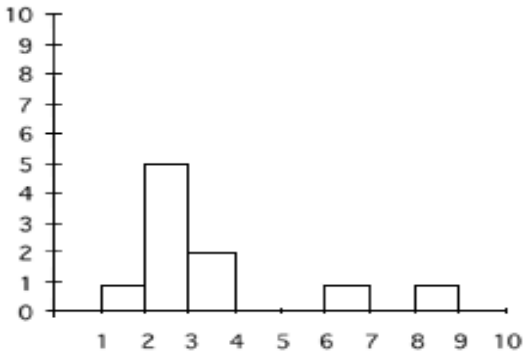
Comments:

_____

_____

11. Consider the following multiple-choice item.

Here is a histogram for a set of test scores from a 10-item makeup quiz given to a group of students who were absent on the day the quiz was given.



What do the numbers on the horizontal axis represent? Please select the best response from the list.

E. Scores on the test.
F. Independent variable.
G. Dependent variable.
H. Number of Students.

(Adapted from the NSF-funded Web Assessment Resource Tools for Improving Statistical Thinking (ARTIST) Project).

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

Comments:

_____

_____

References

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., and Scheaffer, R. (2007). *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A Pre-K-12 Curriculum Framework.* Alexandria, VA: American Statistical Association.

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

Organization for Economic Cooperation and Development (OECD). (2009). *Take The Test: Sample Questions from OECD's PISA Assessment*. Paris: Author.

Watson, J.M. & Callingham, R.A. 2003. Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal*, 2(2), 3-46.

Watson, J.M., & Callingham, R.A. 2004. Statistical literacy: From idiosyncratic to critical thinking. Paper presented at the International Association for *Statistics Education Roundtable on "Curricular Development in Statistics Education,"* Lund, Sweden.

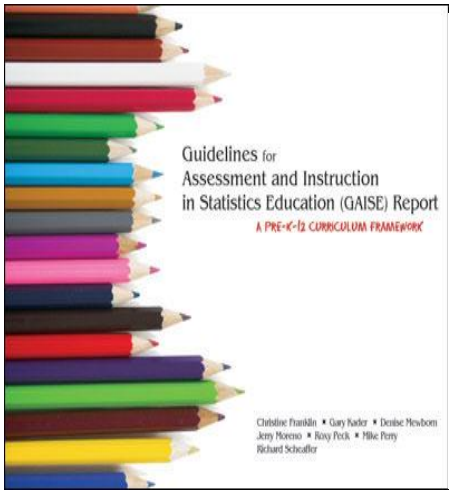
Watson, J.M. & Callingham, R.A. 2005. Measuring statistical literacy. *Journal of Applied Measurement*, 6(1), 19-47.

APPENDIX F

EXPERT SURVEY INSTRUMENT FORM 3

TEXAS STATE UNIVERSITY
SAN MARCOS

Expert Survey to Measure the Alignment between Assessment Items and Pre-K-12 GAISE Levels



@ 2012Generated by Rini Oktavia
Advisor: Dr. Maria Alejandra Sorto
Department of Mathematics
Texas State University-San Marcos

July 11, 2013

Dear Experts,

This survey is a part of a dissertation project for pursuing a doctoral degree in mathematics education at Texas State University – San Marcos. Your response will help develop an instrument to identify students' learning trajectory in statistics based on the Pre-K-12 Guidelines for Assessment and Instruction in Statistics Education (GAISE) Framework and the Common Core State Standards in Mathematics (CCSS-M).

The survey asks your judgment about how 11 content items align with the developmental level of statistics education suggested by the Pre-K-12 GAISE Framework (see page 3 and 4). Your participation in this study is completely voluntary and you can refuse to answer any questions. You may choose to discontinue completing it at any point. If you choose to discontinue, the information collected will not be used in this project.

There is no direct benefit to you for completing the survey. However, your response will help develop an important tool for identifying students' learning trajectories in Statistics. Hence, your response is critically important. There is no risk to you beyond that associated with the completion of a survey.

All responses will be treated with the utmost confidentiality and your privacy will be protected to the maximum extent allowable by law. To ensure confidentiality, your name will not be associated with your responses and no identifying information will be reported in any way unless you give permission to do so.

If you have questions about this survey or about your participation in this research project, you may contact Rini Oktavia by phone at 512-245-4747 or by e-mail at ro1088@txstate.edu.

Thank you for taking the time to complete the survey.

Sincerely,

Rini Oktavia

Doctoral candidate in Mathematics Education

DEPARTMENT OF MATHEMATICS

601 University Drive | San Marcos, Texas 78666-4616 | *phone:* 512.245.2551 | *fax:* 512.245.3425 |

WWW.MATH.TXSTATE.EDU/MATH

Table 1.
Pre-K-12 GAISE Framework (*Franklin et al., 2005)*

Process Component	Level A	Level B	Level C
Formulate Question	Beginning awareness of the *statistics question distinction*	Increase awareness of the *statistics question distinction*	Students can make the *statistics question distinction*
	Teachers pose questions of interest	Students begin to pose their own questions of interest	Students pose their own questions of interest
	Questions restricted to classroom	Question not restricted to classroom	Questions seek generalization
Collect Data	Do not yet design for differences	Beginning awareness of design for differences	Students make design for differences
	Census of classroom		
	Simple experiment	Sample surveys; begin to use random selection	Sampling designs with random selection
		Comparative experiment; begin to use random allocation	Experimental designs with randomization
Analyze Data	Use particular properties of distributions in the context of a specific example	Learn to use particular properties of distributions as tools of analysis	Understand and use distributions in analysis as a global concept
	Display variability within a group	Quantify variability within a group	Measure variability within a group; measure variability between groups
	Compare individual to individual	Compare group to group in displays	Compare group to group using displays and measures of variability
	Compare individual to group	Acknowledge sampling error	Describe and quantify sampling error
	Beginning awareness of group to group	Some quantification of association; simple models for association	Quantification of association; fitting of models for association
	Observe association between two variables		

Table 1 continued

Process Component	Level A	Level B	Level C
Interpret Results	Students do not look beyond the data No generalization beyond the classroom Note difference between two individuals with different conditions Observe association in displays	Students acknowledge that looking beyond the data is feasible Acknowledge that a sample may or may not be representative of the larger population Note the difference between two groups with different conditions Aware of distinction between observational study and experiment Note differences in strength of association Basic interpretation of models for association Aware of the distinction between association and cause and effect	Students are able to look beyond the data in some contexts Generalize from sample to population Aware of the effect of randomization on the results of experiments Understand the difference between observational studies and experiments Interpret measures of strength of association Interpret models of association Distinguish between conclusions from association studies and experiments
Nature of Variability	Measurement variability Natural variability Induced variability	Sampling variability	Chance variability
Focus on variability	Variability within a group	Variability within a group and variability between groups Covariability	Variability in model fitting

Note. The Pre-K-12 GAISE Framework. Reprinted from "*Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A Pre-K-12 Curriculum Framework*," by C. Franklin, G. Kader, D. Mewborn, J. Moreno, R. Peck, M. Perry, & R. Scheaffer, 2005, Copyright 2005 by the Joint American Statistics Association/ National Council of Teachers of Mathematics. Adapted with permission.

1. Consider the following multiple-choice item.

> *If a fair coin is tossed, the probability that it will land heads up is 1/2. In four successive tosses, a fair coin lands heads up each time. What is likely to happen when the coin is tossed a fifth time?*
>
> *A. It is more likely to land tails up than heads up.*
> *B. It is more likely to land heads up than tails up.*
> ***C. It is equally likely to land heads up or tails up.***
> D. *More information is needed to answer the question.*

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

Comments:

_____

_____

2. Consider the following multiple-choice item.

> *A certain state lottery awards 18 $200 prizes, 120 $25 prizes and 270 $20 prizes, for every 10,000 tickets sold. Bob and Bill each bought one ticket each week for the past 100 weeks. Bill has not won a single prize yet. Bob just won a $20 prize last week. Who is more likely to win a prize this coming week? Select the best answer.*
>
> *C. Bill*
> *D. Bob*
> ***E. They have an equal chance of winning***
> F. *Not enough information to tell*
>
> (Adapted from the NSF-funded Web Assessment Resource Tools for Improving Statistical Thinking (ARTIST) Project).

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

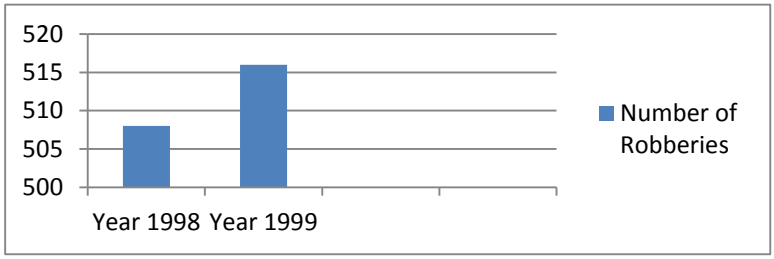
Comments:

_____

_____

3. Consider the following multiple-choice item.

A TV reporter showed this graph and said:
"The graph shows that there is a huge increase in the number of robberies from 1998 to 1999."



Of the following options, which one do you think to be the most appropriate answer for the following question? Do you consider the reporter's statement to be a reasonable interpretation of the graph? Why?

A. Yes, it is reasonable because the bar for Year 1999 is three times higher than the bar for 1998.
B. No, it is not reasonable because only a small part of the graph is shown; if the whole graph is shown, you would see that there is only a slight increase in robberies.
C. No, it is not reasonable because "huge" is not an appropriate term to describe the increasing number of robberies.
D. Yes, it is reasonable because robberies increases almost doubled from 1998 to 1999.

(Adapted from PISA Assessment 2009, OECD, 2009).

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

Comments:

_____

_____

4. Consider the following multiple-choice item.

> *Grade 6 students in Goodnight Middle Schools will conduct an experiment to answer the question of whether beans grow faster in the dark or in the light. From the following options which one would be the most appropriate way to collect data to answer the question?*
>
> *A. Students plant the same number of dried beans in two large plant pots. They put one pot in the light and the other in the dark, and let the beans sprout. After two weeks, the number of growing plants in each pot is counted.*
>
> ***B. Students plant the same number of dried beans in two large plant pots and put one pot in the light and the other in the dark, and let the beans sprout. After two weeks, the heights of the plants are measured.***
>
> *C. Students plant some dried beans in one large plant pot. They put the pot in the light and let the beans sprout. After two weeks, the heights of the plants are measured. The pot, then, is put in the dark, and after two weeks the plants are measured again.*
>
> *D. Students plant some dried beans in one large plant pot. They put the pot in the dark and let the beans sprout. After two weeks, the heights of the plants are measured. The pot, then, is put in the light, and after two weeks the plants are measured again.*

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

Comments:

_____

_____

5. Consider the following multiple-choice item.

> *A town contains three elementary schools. School A has a mean class size of 30 pupils for its three fifth-grade classrooms. School B has a mean class size of 25 pupils in its two fifth-grade classrooms. School C has 20 pupils in its only fifth-grade classroom. What is the average class size for fifth-grade classrooms in this town?*
>
> *A. 12.5*
> *B. 25*
> ***C. 26.7***
> D. *Cannot be determined.*

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

Comments:

_____

_____

6. Consider the following multiple-choice item.

A bowl has 100 color candies in it. 20 are yellow, 50 are red, and 30 are blue. They are well mixed up in the bowl. Randy pulls out a handful of 10 candies, counts the number of reds, and records it on the board.

Then, Randy puts the candies back into the bowl, and mixes them up again. Four of Randy's classmates, Renee, Ricky, Robby, and Rosie do the same thing. One at a time they pull ten candies, count the reds, and write down the number of reds, and put the candies back in the bowl and mix them up again.

Which of the following lists for the number of reds is most likely to be?

A. 8, 9, 7, 10, 9
B. 3, 7, 5, 8, 5
C. 5, 5, 5, 5, 5
D. 2, 4, 3, 4, 3

(Adapted from The Lollie Task, Shaughnessy, 2007).

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

Comments:

_____

_____

7. Consider the following multiple-choice item.

Two fair spinners (half black (B) and half white (W)) are part of a carnival game. A player wins a prize only when both arrows land on black (BB) after each spinner has been spun once.

Cody wanted to play the game; he thinks he has a 50-50 chance of winning. Do you agree?
Which of the following options are the most appropriate to answer the question above?

A. No, because if there were one spinner, the chance of winning it would be 50%, so it has to be less with two spinners.

B. Yes, because each spinner are half white and half black.

C. No, because there are four possible outcomes, BB, BW, WB, WW. So, the chance will be 25%.

D. Yes, because there are two spinners with the same areas of white and black.

(Adapted from the Spinner task from the 1996 NAEP, Shaughnessy, 2007)

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

Comments:

_____

_____

8. Consider the following multiple-choice item.

> *Consider a situation involving two variables X and Y. What conditions would need to be satisfied in order to say that a change in the variable X causes a change in the variable Y?*
>
> *A. When the correlation between X and Y is close to 1 or -1.*
> *B. When an experiment reveals that a change in X causes a change in Y.*
> *C. When possible confounding variables have been ruled out.*
> ***D. All of the above.***
>
> (Adapted from the NSF-funded Web Assessment Resource Tools for Improving Statistical Thinking (ARTIST) Project).

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

Comments:

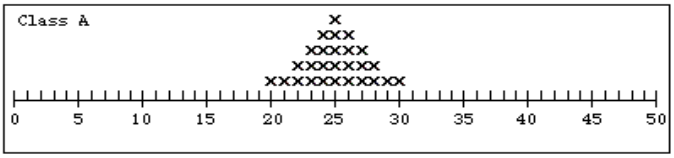
_____

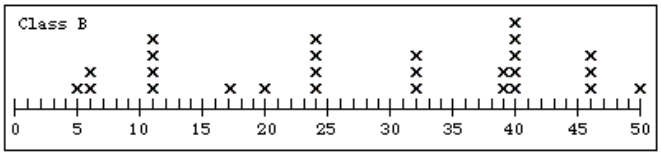
_____

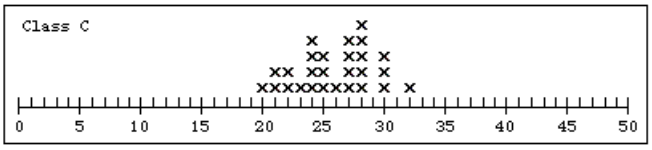
9. Consider the following multiple-choice item.

A class of students tossed 50 pennies and counted the number of heads. They repeated this many times. Imagine that two other classes produced graphs for the same experiment. In some cases, the results were just made up without actually doing the experiment.

The following dot plots show the results obtained by all classes.







Which of the results is more likely made up (not really from the experiment)?

A. Class A's results

B. Class B's results

C. Class C's results

D. All results are more likely made up.

(Adapted from the NSF-funded Web Assessment Resource Tools for Improving Statistical Thinking (ARTIST) Project).

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

Comments:

_____

_____

10. Consider the following multiple-choice item.

> *Which of the following sequences is most likely to result from flipping a fair coin five times?*
>
> *A. H H H T T*
>
> *B. T H H T H*
>
> *C. T H T T T*
>
> ***D. All three sequences are equally likely.***

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

Comments:

_____

_____

11. Consider the following multiple-choice item.

The following table summarizes the data on a survey that ask the following questions. "Do you like rock music?" and "Do you like rap music?" The participants are randomly selected from all middle school students in San Marcos, TX.

		Like Rock Music?		
		Yes	*No*	***Row total***
Like Rap Music?	*Yes*	*25*	*4*	***29***
	No	*6*	*15*	***21***
Column total		***31***	***19***	***50***

Of the following options, which one is the most accurate explanation of the data represented in the table?

A. There may be a strong association between liking Rock music and liking Rap music. However this association could simply be a consequence of a random sampling.

*B. There may **NOT** be a strong association between liking Rock music and liking Rap music. However this association could simply be a consequence of a random sampling.*

*C. More than 50% of San Marcos middle school students do **NOT** like Rap music. However this association could simply be a consequence of a random sampling.*

*D. More than 50% of San Marcos middle school students do **NOT** like Rock music. However this association could simply be a consequence of a random sampling.*

(Adapted from the Appendix of The Pre-K-12 GAISE Report, Franklin et al., 2007)

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

Comments:

_____

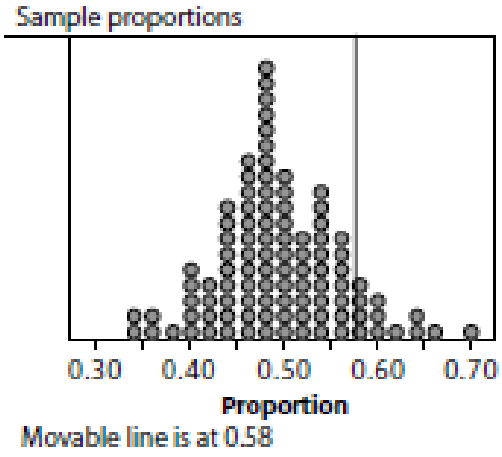
_____

12. Consider the following multiple-choice item.

Among 50 students in a middle school who were randomly chosen to participate in a survey, fifty eight percent of the students like ice cream and 52% of the students like cakes for dessert.

It is claimed that more than 50% of students in the middle school like ice cream.
To simulate the situation, a computer generates a set of even and odd digits to represent students who like ice cream and who do not like ice cream respectively. Samples with size 50 are randomly chosen repeatedly from the set of digits.

The number of even digits from each sample is counted and the proportion of even digits from each sample is recorded. After 100 simulations the sampling distribution is represented by the following graph.



Based on this simulation, a sample proportion greater than or equal to the observed 0.58 occurred 12 times of 100 just by chance variation alone when the actual population proportion is 0.5. What is suggested by this result?

A. ***The claim that more than 50% of students in the middle school like ice cream is NOT supported by the evidence.***

B. *The claim that more than 50% of students in the middle school like ice cream is supported by the evidence.*

C. *The result that 58% of the students like ice cream is **NOT** likely due to chance alone.*

D. *The result that 58% of the students who were interviewed like ice cream is **NOT** supported by the evidence.*

(Adapted from the Appendix of The Pre-K-12 GAISE Report, Franklin et al., 2007)

This item best aligns with:

1. Level A □
2. Level B □
3. Level C □
4. None □

Comments:

_____

_____

References

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., and Scheaffer, R. (2007). *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A Pre-K-12 Curriculum Framework.* Alexandria, VA: American Statistical Association.

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

Organization for Economic Cooperation and Development (OECD). (2009). *Take The Test: Sample Questions from OECD's PISA Assessment*. Paris: Author.

Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F. Lester, *Second Handbook of Research on Mathematics Teaching and Learning* (pp. 957-1009). Greenwich, CT: Information Age Publishing, Inc. and NCTM.

Watson, J.M. & Callingham, R.A. 2003. Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal*, 2(2), 3-46.

Watson, J.M., & Callingham, R.A. 2004. Statistical literacy: From idiosyncratic to critical thinking. Paper presented at the International Association for *Statistics Education Roundtable on "Curricular Development in Statistics Education,"* Lund, Sweden.

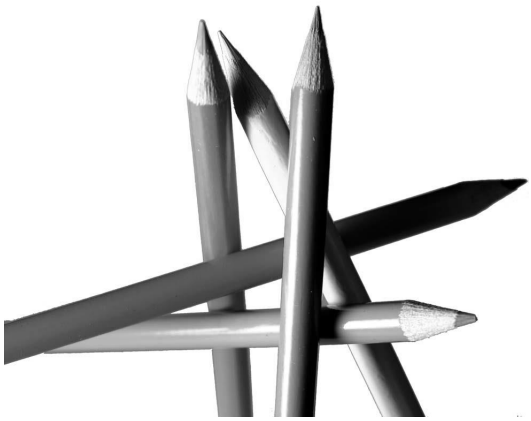
Watson, J.M. & Callingham, R.A. 2005. Measuring statistical literacy. *Journal of Applied Measurement*, 6(1), 19-47.

APPENDIX G

PILOT SURVEY INSTRUMENT

TEXAS STATE UNIVERSITY
SAN MARCOS

Diagnostic Assessment to Measure Students' Developmental Levels in Learning Statistics



@ 2012 Developed by Rini Oktavia
Department of Mathematics
Texas State University-San Marcos

August 2012,

Dear Students.

We need your help!

This assessment is part of a dissertation project for pursuing a doctoral degree in mathematics education from Texas State University – San Marcos. Your response will help us developing a valid and reliable diagnostic assessment to identify students' developmental levels in learning statistics based on Pre-K-12 GAISE framework.

The assessment consists of 13 multiple-choice items that assess students' knowledge and skills on several statistics ideas.

Your participation in this study is completely voluntary. You can refuse to answer any question. You may choose to discontinue completing it at any point. If you choose to discontinue, the information collected will not be used in this project.

There is no direct benefit to you for completing the test. However, your response will help us developing an important tool for teachers in assessing students' developmental level of statistical education. Hence, your response is critically important. There is no risk to you beyond that associated with the completion of the test.

All responses will be treated with the utmost confidentiality and your privacy will be protected to the maximum extent allowable by law. To ensure confidentiality, your name will not be associated with your responses and no identifying information will be reported in any way unless you give us permission to do so.

If you have questions about this test or your participation in this research project, you may contact Rini Oktavia by phone at 512-245-4747 or by e-mail at ro1088@txstate.edu.

Thank you for taking the time to complete the test.

Sincerely,

_____
Rini Oktavia

Doctoral student in Mathematics Education
Texas State University - San Marcos

Form 1

Name :_____

School Grade :_____

Gender : F / M

Please choose one answer from several answer choices given in the following questions.

1. There are 20 students in a mathematics class; ten boys and ten girls. The teacher will choose three students randomly and ask them to show their work on the board. A student can only be chosen once. After choosing two boys, what is the chance that the teacher will choose another boy to work on the board?

    A. Ten out of twenty.

    B. Eight out of twenty.

    C. Ten out of eighteen.

    D. Eight out of eighteen.

2. A teacher found that the median of her students' grades on a mathematics test is 69. There are 8 students took the test. Which of the following data more likely represent her students' grades on the test?

    A. 63, 65, 67, 69, 70, 72, 73, 75.

    B. 75, 69, 70, 69, 69, 78, 69, 100.

    C. 20, 67, 65, 69, 79, 69, 70, 90.

    D. 25, 27, 29, 67, 71, 80, 95, 100.

3. A bottle of medicine has printed on it:

> **WARNING: For applications to skin areas there is a 15% chance of getting a rash.**
>
> **If you get a rash, consult your doctor.**

What is the meaning of the warning?

A. Don't use the medicine on your skin – there's a good chance of getting rash.

B. For application to the skin, apply only 15% of the recommended dose.

C. If you get a rash, it will probably involve only 15% of the skin.

D. About 15 out of every 100 people who use this medicine get a rash.

4. When three fair dice are simultaneously thrown, which of the following results is MOST LIKELY to be obtained?

A. Result 1: A 5, a 3 and a 6 in any order

B. Result 2: Three 5's

C. Result 3: Two 5's and a 3

D. All three results are equally likely.

5. John is flipping a coin ten times. Tony is flipping a coin 100 times. Which one of the following options is appropriate to describe the possible outcomes that John and Tony get?

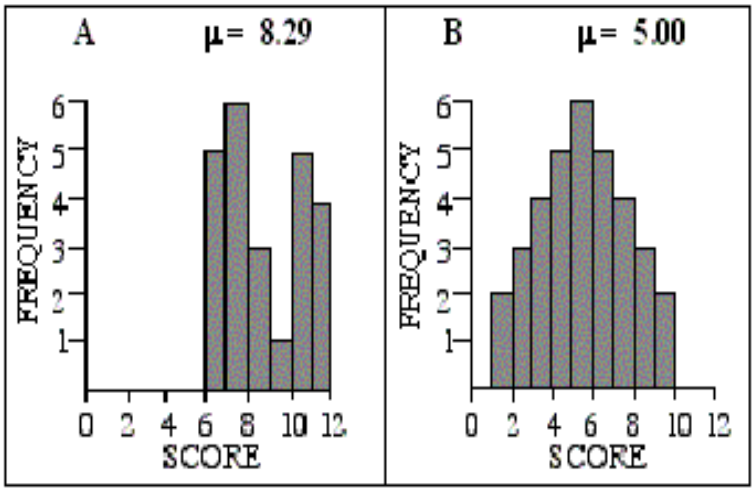
    A. It is impossible that John gets five heads and five tails and Tony gets 50 heads and 50 tails.

    B. It is less likely that Tony gets 50 heads and 50 tails rather than that John gets five heads and five tails.

    C. It is more likely that Tony gets 50 heads and 50 tails rather than that John gets five heads and five tails.

    D. It is equally likely that John gets 5 heads and 5 tails and Tony gets 50 heads and 50 tails.

6. For each pair of graphs, determine which graph has the higher standard deviation (it is not necessary to do any calculations to answer these questions).



    A. A has a larger standard deviation than B

    B. B has a larger standard deviation than A

    C. Both graphs have the same standard deviation

    D. Cannot be determined

7. Of the following options, which one can be answered with a statistical investigation using Miller Middle School students' basic health information data?

    A. What is the rate of obesity among students in the school?

    B. Who is the tallest student in the school?

    C. Is the overall health of middle school students declining in this country?

    D. All of the above questions can be answered using statistical investigation.

8. A farmer wants to know how many fish there are in his dam. He took out 200 fish and tagged each of them, with a colored sign. He put the tagged fish back in the dam and let them get mixed with the others. On the second day, he took out 250 fish randomly and found that 25 of them were tagged. Estimate how many fish are in the dam.

    A. 250

    B. 500

    C. 1000

    D. 2000

9. A small object was weighed on the same scale separately by nine students in a science class. The weights (in grams) recorded by each student are shown below.

    6.3 6.0 6.0 15.3 6.1 6.3 6.2 6.15 6.3

    The students want to determine as accurately as they can the actual weight of this object. Of the following methods, which would you recommend they use?

    A. Use the most common value, which is 6.2.

    B. Use the 6.15 since it is the most accurate weighing.

    C. Add up the nine numbers and divide by 9.

    D. Throw out the 15.3, add up the other 8 numbers and divide by 8.

10. Mrs. Jones wants to buy a new car, either a Honda or a Toyota. She wants whichever car that will break down the least. She read in Consumer Reports that for 400 cars of each type, the Toyota had more breakdowns than the Honda. She talked to three friends. Two were Toyota owners, who had no major breakdowns. The other friend used to own a Honda, but it had lots of breakdowns, so he sold it. He said he'd never buy another Honda. Which car should Mrs. Jones buy?

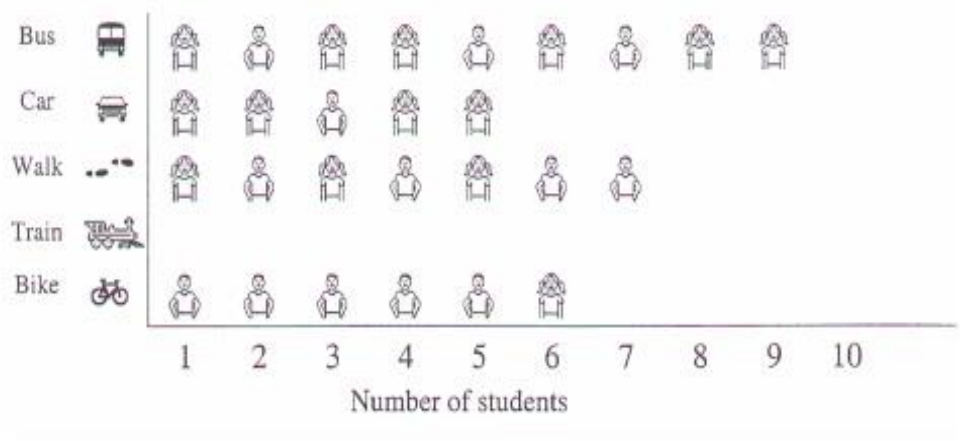
    A. Mrs. Jones should buy the Toyota, because her friend had so much trouble with his Honda, while her other friends had no trouble with their Toyotas.

    B. She should buy the Honda, because the information about break-downs in Consumer Reports is based on many cases, not just one or two cases.

    C. It doesn't matter which car she buys. Whichever type she gets, she could still be unlucky and get stuck with a particular car that would need a lot of repairs.

    D. Mrs. Jones should not buy either the Honda or the Toyota, because both cars have major breakdowns history.

11. The following graph represents how children came to school one day.



How many children walk to school?

A. 9

B. 5

C. 7

D. 6

12. A class of students asks a question on who can jump farther, boys or girls. Of the following options which one do you think as the best way to collect data to answer the question?

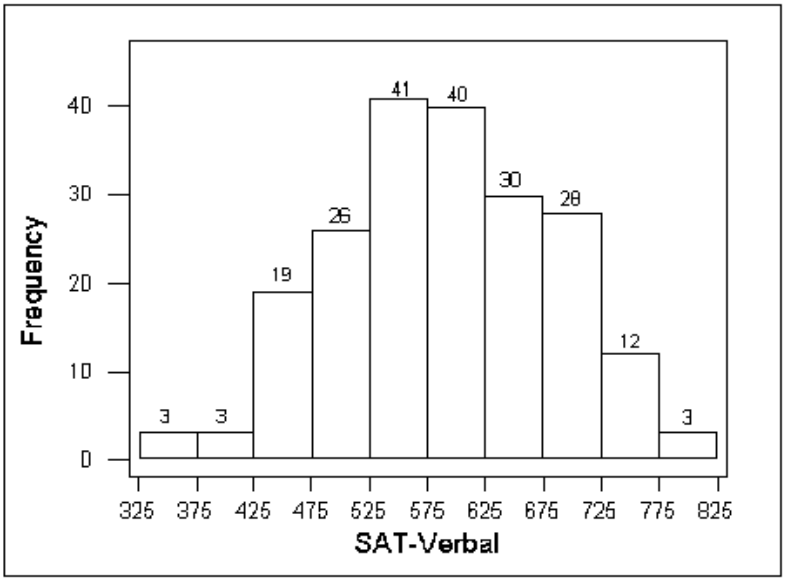
    A. Students measure the jumping distances for all of their classmates.

    B. Students measure the heights for all of their classmates.

    C. Students count how many boys and girls who can jump farther than four feet.

    D. Some students volunteer to jump and their jumping distances are measured.

13. The following histogram shows the Verbal SAT scores for 205 students entering a local college in the fall of 2002.



How many of the students had verbal SAT scores between 425 and 725?

    A. 19

    B. 28

    C. 184

    D. 139

14. A sample of 50 students was taken from a large urban school with 1000 students and a sample of 20 students was taken from a small rural school with 300 students. Both schools have the same percentage of girls and boys. One of these samples was strange in that it had 80% boys. Which do you think is more likely?

    A. The sample is from the small school.

    B. The sample is from the large school.

    C. The sample could be from the large school or the small school.

    D. It is impossible to have a sample with 80% boys.

15. Of the following questions, which one is a statistical question?
    Hint: A statistics question is a question that anticipates an answer based on data that vary.

    A. How tall is the tallest building in the world?

    B. How tall are adult men in the United States?

    C. How many students attend Miller Middle School in 2012?

    D. How many times a week do you practice soccer?

16. A group of 649 men with lung cancer was identified from a certain population in England. A control group about the same size was established by matching these patients with other men from the same population who did not have lung cancer. The matching was on background variables such as ethnicity, age, and socioeconomic status. The summary of level of smoking and the number of lung cancer and control cases is given in the following table.

Cigarettes /Day	**Lung Cancer Cases**	**Control**	**Probability of Lung Cancer**
0	2	27	2/29 = 0.07
1 - 14	283	346	283/629 = 0.45
15 - 24	196	190	196/ 386 = 0.51
25 +	168	84	168/252 = 0.67

What is the association between the level of smoking and the number of lung cancer cases that can be inferred by the given data?

A. A decrease in the lung cancer rate is associated with an increase in cigarette smoking.

B. An increase in the lung cancer rate is associated with an increase in cigarette smoking.

C. An increase in the lung cancer rate is associated with a decrease in cigarette smoking.

D. There is no association between the level of smoking and the number of lung cancer cases.

17. The principal of Miller Middle School would like to study the feelings of students about the food served in the cafeteria. He plans to have college students, who are volunteering in the school, interview every 10th student who walks by the cafeteria between the hours of 11:00 am and 1:00 pm. Of the following statements on the strengths or weaknesses of this sampling plan, which would you recommend as the most appropriate?

    A. The survey is good, because every student has the same chance to be interviewed.

    B. The survey is not fair, because some groups of students might not have lunch in the cafeteria.

    C. The survey is good, because the students are interviewed randomly near the cafeteria.

    D. The survey is not fair, because most of the students who are interviewed are boys.

18. Box A and Box B are filled with red and blue marbles as follows. Each box is shaken. You want to get a blue marble, but you are only allowed to pick out one marble without looking. Which box should you choose, and why?

Box A	Box B
6 red 4 blue	60 red 40 blue

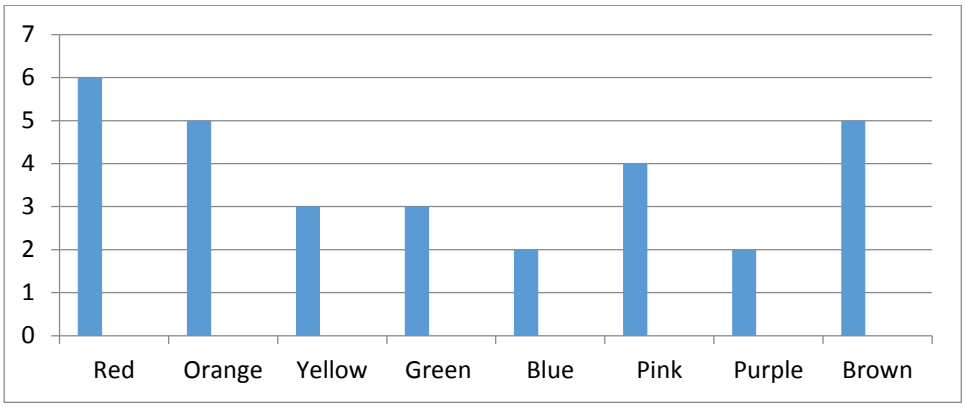
    A. Box B (with 60 red and 40 blue), because it contains more blue marbles.

    B. Box A (with 6 red and 4 blue), because the difference between the number of red and blue marbles is small.

    C. It doesn't matter, because Box B has ten times the amount in Box A.

    D. It doesn't matter, because both boxes have 40% blue marbles.

19. Robert's mother lets him pick one candy from a bag. He can't see the candies. The number of candies of each color in the bag is shown in the following graph.



What is the probability that Robert will pick a red candy?

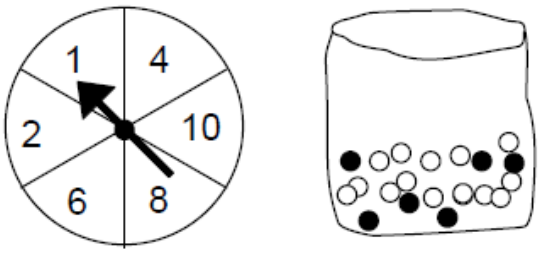
A. 10%

B. 20%

C. 25%

D. 50%

20. A game in a booth at a spring fair involves using a spinner first. Then, if the spinner stops on an even number, the player is allowed to pick a marble from a bag. The spinner and the marbles in the bag are represented in the diagram below.



Prizes are given when a black marble is picked. Sue plays the game once. How likely is it that Sue will win a prize?

A. Impossible.

B. Not very likely.

C. About 50% likely.

D. Very likely.

21. The following information is from a survey about smoking and lung disease among 250 people.

	Lung disease	**No lung disease**	**Total**
Smoking	90	60	150
No smoking	60	40	100
Total	150	100	250

Using this information; of the following options, which do you think is the most appropriate?

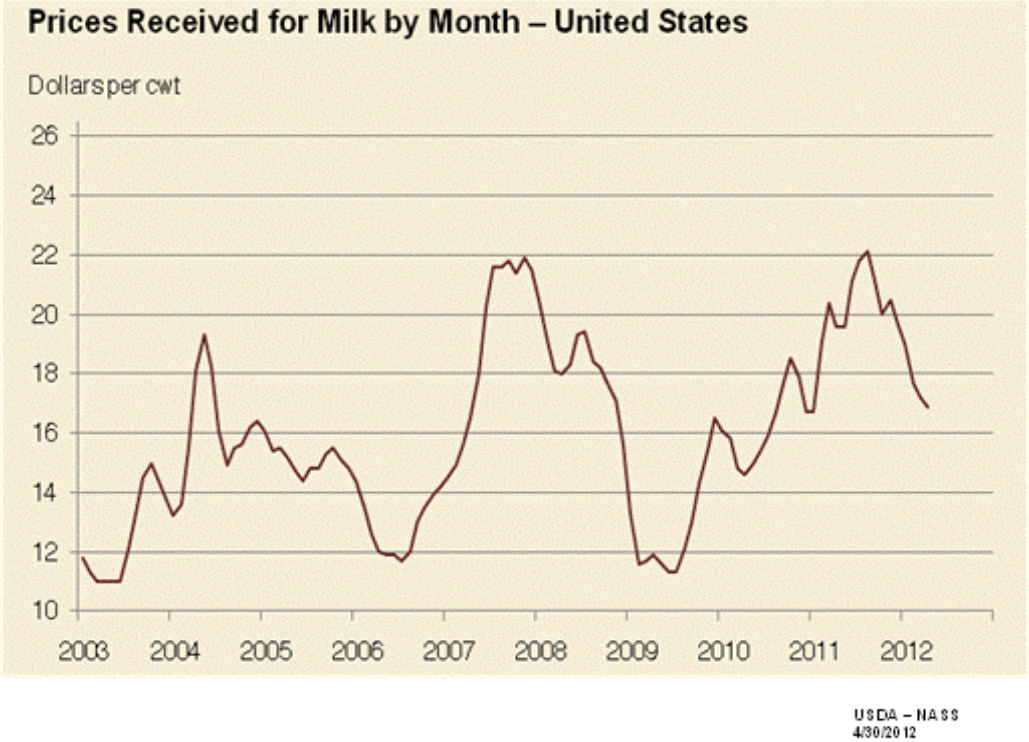
A. Yes, lung disease is associated with smoking, because the number of people who are smoking and have lung disease are bigger than the number of people who are smoking and do not have lung disease.

B. No, lung disease is not associated with smoking, because the percentage of people who have lung disease and smoking and the people who have lung disease and not smoking are the same (0.6).

C. Yes, lung disease is associated with smoking, because smoking is known to cause lung cancer.

D. No, lung disease is not associated with smoking, because the number of people who are smoking and have no lung disease is the same as the number of people who are not smoking and have lung disease.

22. The following graph provided by the U. S. Department of Agriculture shows the prices for milk by month in dollars per hundred pounds (cwt) in the U. S. since 2003 to 2012.



How much is the prices for milk in the middle of 2006?

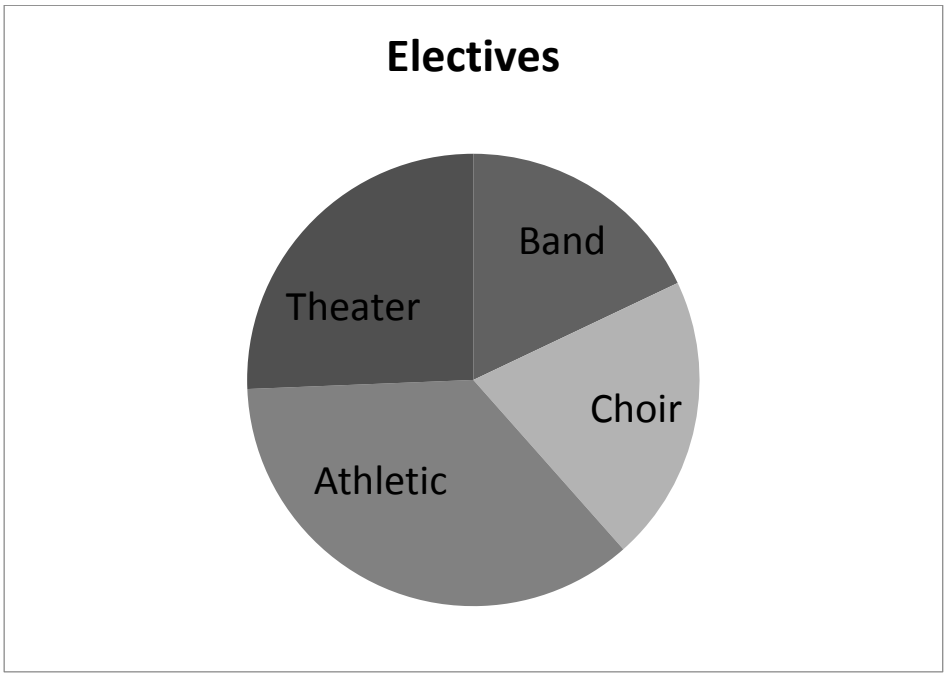
A. 21.8 dollars per hundred pounds.

B. 11.90 dollars per hundred pounds.

C. 13.5 dollars per hundred pounds.

D. 14.5 dollars per hundred pounds.

23. The graph shows the distribution of students among elective activities in a certain school.



According to the information in the graph, which of this statement is true?

A. More students join Band than Choir.

B. Half of students join Athletic.

C. More than one third of students join Band.

D. More students join Theater or Choir than Band.

24. Four students at a local high school conducted surveys. Shannon got the names of all 800 children in the high school and put them in a hat, and then pulled out 60 of them. Jake asked 10 students at an after-school meeting of the computer games club. Adam asked all of the 200 children in Grade 10. Claire set up a booth outside of the school. Anyone who wanted to stop and fill out a survey could. She stopped collecting surveys when she got 60 students to complete them. Who do you think has the best sampling method? Why?

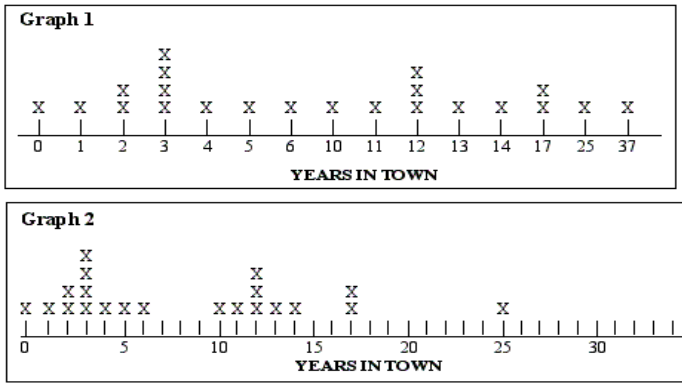
    A. Adam, because asking all Grade 10 students are a good way to get all possible opinions of all students in the school.

    B. Jake, because all the computer games club members are Jake's friends. Their answers are trustworthy.

    C. Claire, because every student has an opportunity to be interviewed before she gets 60 students.

    D. Shannon, because the participants are chosen randomly.

25. A class of students recorded the number of years their families had lived in their town. Here are two graphs that students drew to summarize the data.



Which graph gives a more accurate representation of the data? Why?

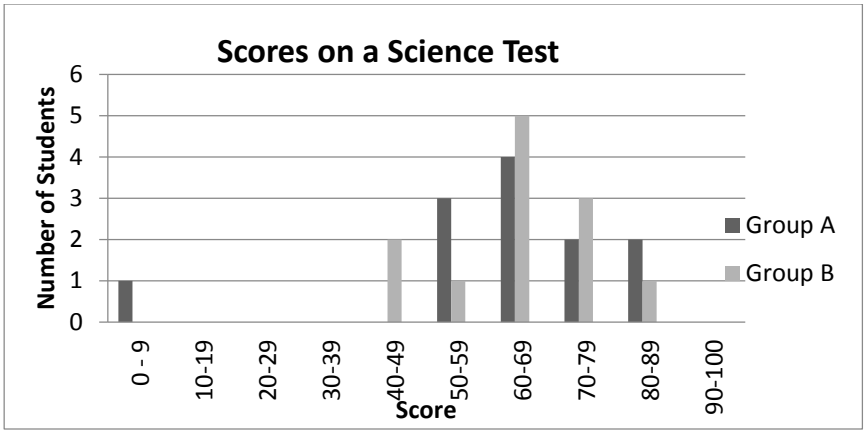
A. Graph 1, because it shows the number of years each student's family lived in their town.

B. Graph 2, because it shows all possible data values.

C. Both graphs give the same accurate representation of the data.

D. Both graphs give the same inaccurate representation of the data

26. The diagram below shows the results on a science test for two groups, labeled as Group A and Group B. The mean score for group A is 62.0 and the mean for group B is 64.5. Students pass this test when their score is 50 or above.



Looking at the diagram, the teacher claims that group B did better than Group A in this test. The students in Group A don't agree with their teacher. They try to convince the teacher that Group B may not necessary have done better. Which of the following arguments is the most appropriate to be used by the students in Group A?

A. The scores of Group A have more variations than the scores of Group B.

B. More students in Group A than in Group B passed the test.

C. Group A has better score results in the 80-89 range and the 50-59 range.

D. The difference between the highest and lowest scores is smaller for Group B than for Group A.

27. If a fair coin is tossed, the probability that it will land heads up is 1/2. In four successive tosses, a fair coin lands heads up each time. What is likely to happen when the coin is tossed a fifth time?

A. It is more likely to land tails up than heads up.

B. It is more likely to land heads up than tails up.

C. It is equally likely to land heads up or tails up.

D. More information is needed to answer the question.

28. A certain state lottery awards 18 $200 prizes, 120 $25 prizes and 270 $20 prizes, for every 10,000 tickets sold. Bob and Bill each bought one ticket each week for the past 100 weeks. Bill has not won a single prize yet. Bob just won a $20 prize last week. Who is more likely to win a prize this coming week? Select the best answer.

    A. Bill

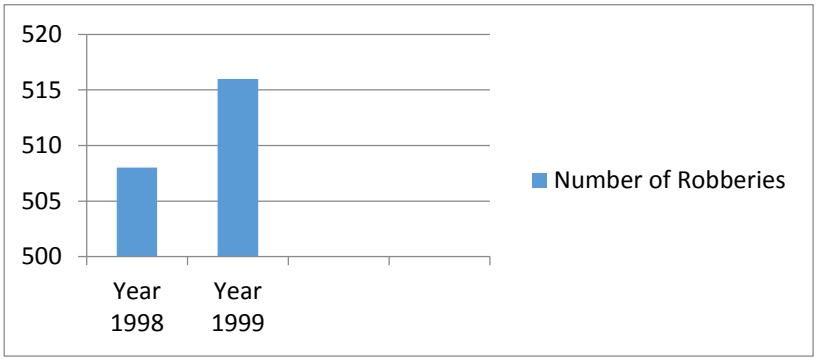
    B. Bob

    C. They have an equal chance of winning

    D. Not enough information to tell

29. A TV reporter showed this graph and said:

    "The graph shows that there is a huge increase in the number of robberies from 1998 to 1999."



    Of the following options, which one do you think to be the most appropriate answer for the following question? Do you consider the reporter's statement to be a reasonable interpretation of the graph? Why?

    A. Yes, it is reasonable because the bar for Year 1999 is three times higher than the bar for 1998.

    B. No, it is not reasonable because only a small part of the graph is shown; if the whole graph is shown, you would see that there is only a slight increase in robberies.

    C. No, it is not reasonable because "huge" is not an appropriate term to describe the increasing number of robberies.

    D. Yes, it is reasonable because robberies increases almost doubled from 1998 to 1999.

30. Grade 6 students in Goodnight Middle Schools will conduct an experiment to answer the question of whether beans grow faster in the dark or in the light. From the following options which one would be the most appropriate way to collect data to answer the question?

    A. Students plant the same number of dried beans in two large plant pots. They put one pot in the light and the other in the dark, and let the beans sprout. After two weeks, the number of growing plants in each pot is counted.

    B. Students plant the same number of dried beans in two large plant pots and put one pot in the light and the other in the dark, and let the beans sprout. After two weeks, the heights of the plants are measured.

    C. Students plant some dried beans in one large plant pot. They put the pot in the light and let the beans sprout. After two weeks, the heights of the plants are measured. The pot, then, is put in the dark, and after two weeks the plants are measured again.

    D. Students plant some dried beans in one large plant pot. They put the pot in the dark and let the beans sprout. After two weeks, the heights of the plants are measured. The pot, then, is put in the light, and after two weeks the plants are measured again.

31. A town contains three elementary schools. School A has a mean class size of 30 pupils for its three fifth-grade classrooms. School B has a mean class size of 25 pupils in its two fifth-grade classrooms. School C has 20 pupils in its only fifth-grade classroom. What is the average class size for fifth-grade classrooms in this town?

    A. 12.5

    B. 25

    C. 26.7

    D. Cannot be determined

32. A bowl has 100 color candies in it. 20 are yellow, 50 are red, and 30 are blue. They are well mixed up in the bowl. Randy pulls out a handful of 10 candies, counts the number of reds, and records it on the board.

    Then, Randy puts the candies back into the bowl, and mixes them up again. Four of Randy's classmates, Renee, Ricky, Robby, and Rosie do the same thing. One at a time they pull ten candies, count the reds, and write down the number of reds, and put the candies back in the bowl and mix them up again.

    Which of the following lists for the number of reds is most likely to be?

    A. 8, 9, 7, 10, 9

    B. 3, 7, 5, 8, 5

    C. 5, 5, 5, 5, 5

    D. 2, 4, 3, 4, 3

33. Two fair spinners (half black (B) and half white (W)) are part of a carnival game. A player wins a prize only when both arrows land on black (BB) after each spinner has been spun once.

    Cody wanted to play the game; he thinks he has a 50-50 chance of winning. Do you agree?

    Which of the following options are the most appropriate to answer the question above?

    A. No, because if there were one spinner, the chance of winning it would be 50%, so it has to be less with two spinners.

    B. Yes, because each spinner are half white and half black.

    C. No, because there are four possible outcomes, BB, BW, WB, WW. So, the chance will be 25%.

    D. Yes, because there are two spinners with the same areas of white and black.

34. Consider a situation involving two variables X and Y. What conditions would need to be satisfied in order to say that a change in the variable X causes a change in the variable Y?

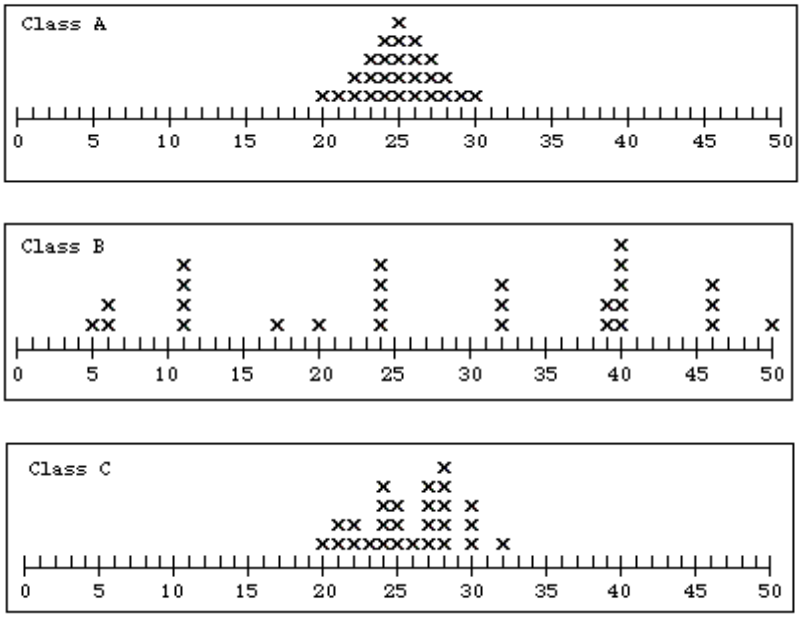
    A. When the correlation between X and Y is close to 1 or -1.

    B. When an experiment reveals that a change in X causes a change in Y.

    C. When possible confounding variables have been ruled out.

    D. When the values for X are always less than the values for Y.

35. A class of students tossed 50 pennies and counted the number of heads. They repeated this many times. Imagine that two other classes produced graphs for the same experiment. In some cases, the results were just made up without actually doing the experiment.
    The following dot plots show the results obtained by all classes.



Which of the results is more likely made up (not really from the experiment)?

    A. Class A's results

    B. Class B's results

    C. Class C's results

    D. All results are more likely made up.

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

    A. H H H T T

    B. T H H T H

    C. T H T T T

    D. All three sequences are equally likely.

37. The following table summarizes the data on a survey that ask the following questions. "Do you like rock music?" and "Do you like rap music?" The participants are randomly selected from all middle school students in San Marcos, TX.

Like Rock Music?

		Yes	No	**Row total**
Like Rap Music?	Yes	25	4	**29**
	No	6	15	**21**
Column total		**31**	**19**	**50**

Of the following options, which one is the most accurate explanation of the data represented in the table?

    A. There may be a strong association between liking Rock music and liking Rap music. However this association could simply be a consequence of a random sampling.

    B. There may **NOT** be a strong association between liking Rock music and liking Rap music. However this association could simply be a consequence of a random sampling.

    C. More than 50% of San Marcos middle school students do **NOT** like Rap music. However this association could simply be a consequence of a random sampling.

    D. More than 50% of San Marcos middle school students do **NOT** like Rock music. However this association could simply be a consequence of a random sampling.

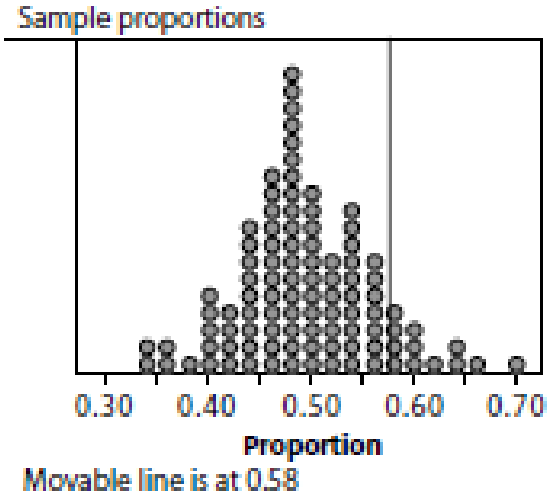
38. Fifty eight percent of 50 students in a middle school, who were randomly chosen to participate in a survey, like ice cream and 52% of the students like cakes for dessert.

    It is claimed that more than 50% of students in the middle school like ice cream.

    A computer generates a set of even and odd digits, and samples with size 50 are randomly chosen repeatedly from the set of digits.

    The number of even digits from each sample is counted and the proportion of even digits from each sample is recorded.

    After 100 simulations the sampling distribution is represented by the following graph.



    Based on this simulation, a sample proportion greater than or equal to the observed 0.58 occurred 12 times of 100 just by chance variation alone when the actual population proportion is 0.5. What is suggested by this result?

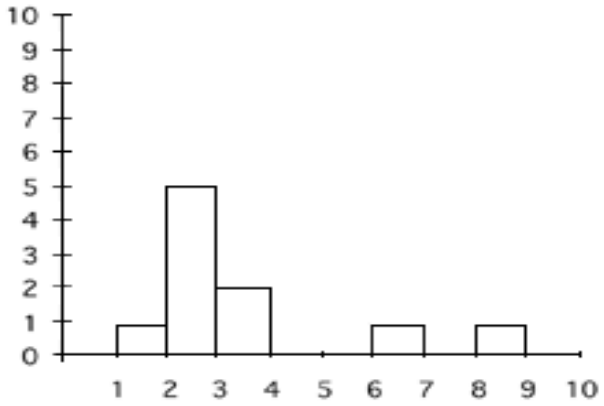
    A. The claim that more than 50% of students in the middle school like ice cream is **NOT** supported by the evidence.

    B. The claim that more than 50% of students in the middle school like ice cream is supported by the evidence.

    C. The result that 58% of the students like ice cream is **NOT** likely due to chance alone.

    D. The result that 58% of the students like ice cream is **NOT** supported by the evidence.

39. Here is a histogram for a set of test scores from a 10-item makeup quiz given to a group of students who were absent on the day the quiz was given.



What do the numbers on the horizontal axis represent? Please select the best response from the list.

A. Scores on the test

B. Independent variable

C. Dependent variable

D. Number of Students

40. A city council wanted to estimate the proportion of residents of the city that would support an increase in taxes for education. A survey is conducted to ask residents whether they would support the increase tax or not. Of the following options, which data collection method will give the most accurate estimation?

A. A sample is chosen by randomly select residents from the list of all residents of the city.

B. A sample is chosen by randomly select residents from a certain area in the city.

C. A sample is chosen randomly from government employees.

D. A sample is chosen randomly from residents who have kids that are still in school.

APPENDIX H

IRB CERTIFICATE

TEXAS STATE UNIVERSITY
SAN MARCOS
The rising STAR of Texas

Institutional Review Board

Request For Exemption

Certificate of Approval

Applicant: Rini Oktavia

Request Number : EXP2012B6438

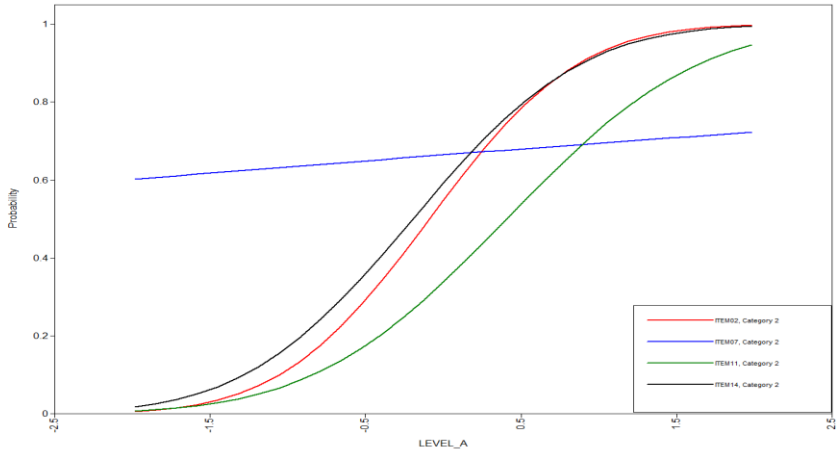
Date of Approval: 02/28/12

Assistant Vice President for Research and Federal Relations

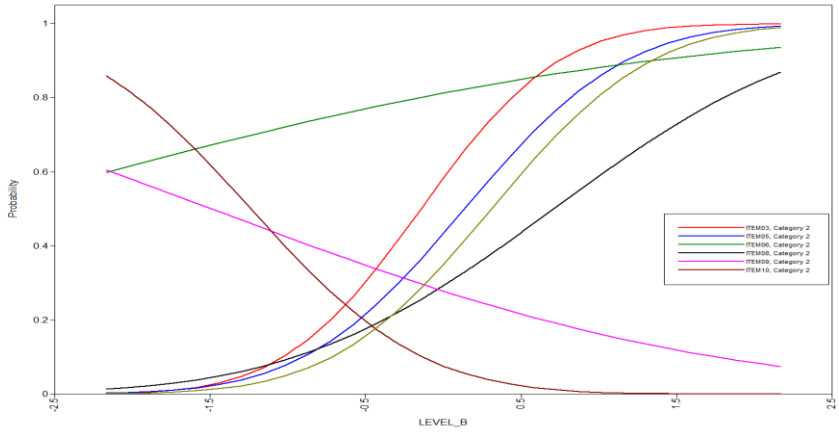
Chair, Institutional Review Board

APPENDIX I

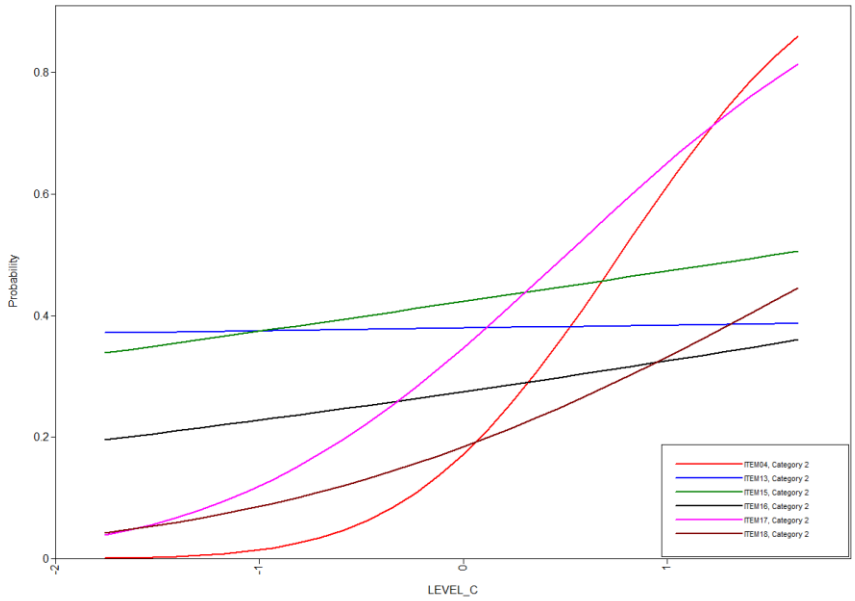
ITEM CHARACTERISTIC CURVES



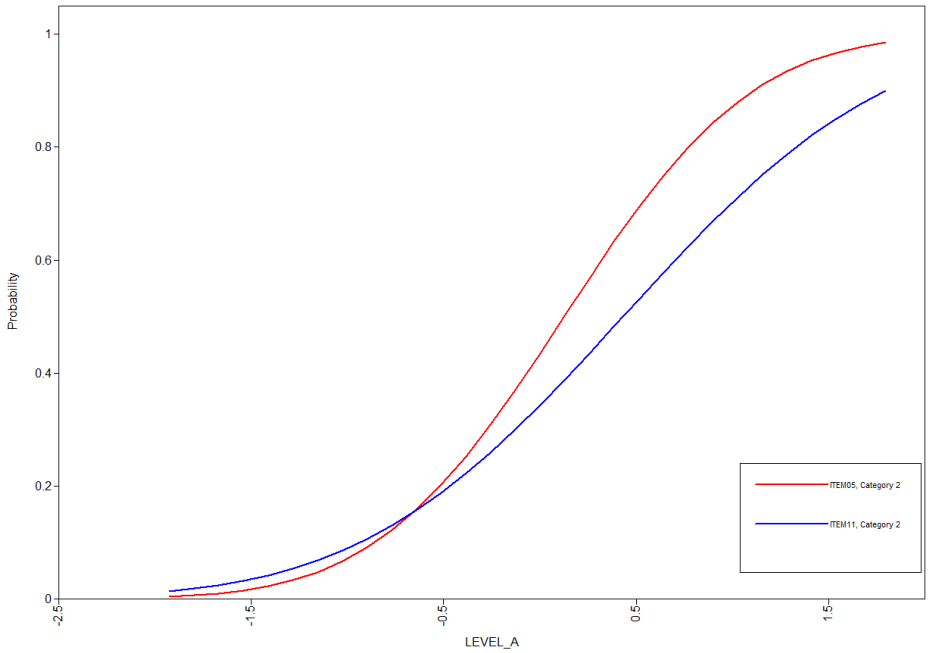
Curves 1: Item Characteristic Curves of Level A Items in Initial F1 Model



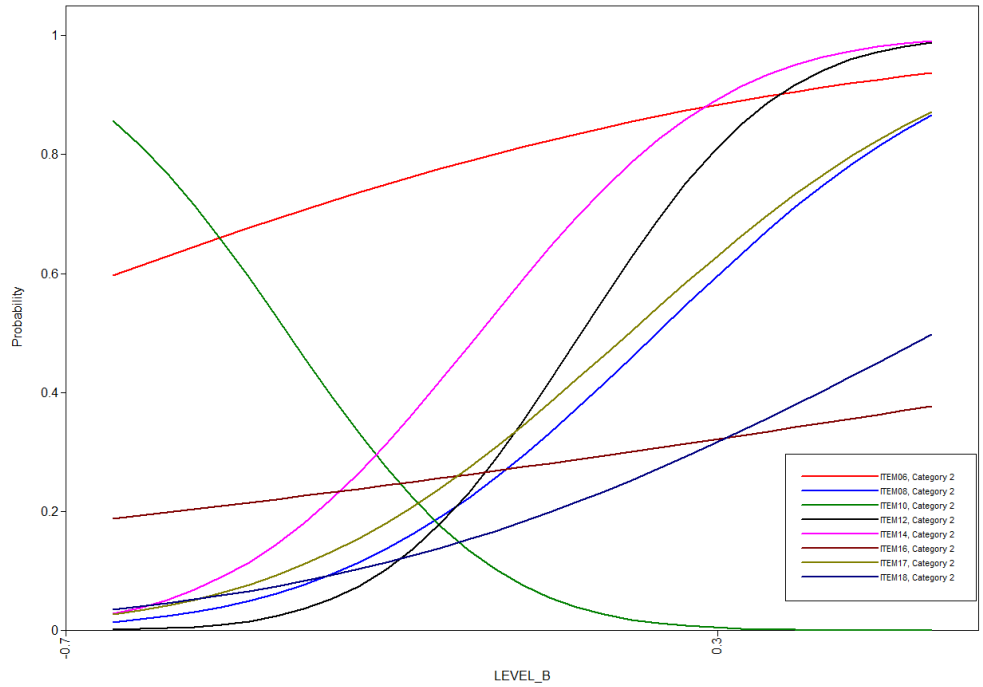
Curves 2: Item Characteristic Curves of Level B Items in Initial F1 Model



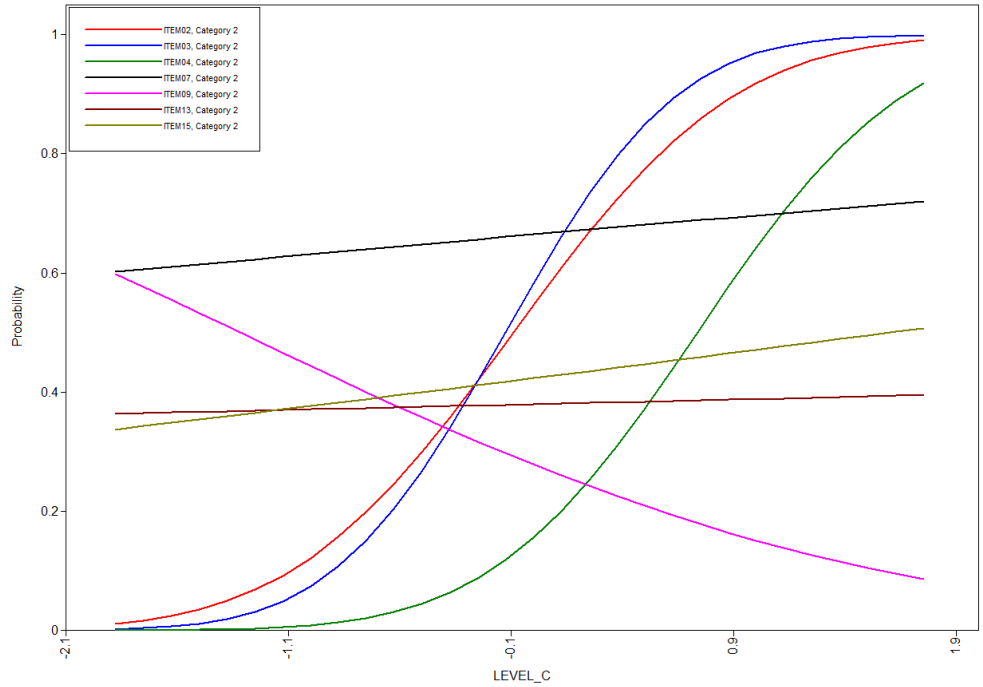
Curves 3: Item Characteristic Curves of Level C Items in Initial F1 Model



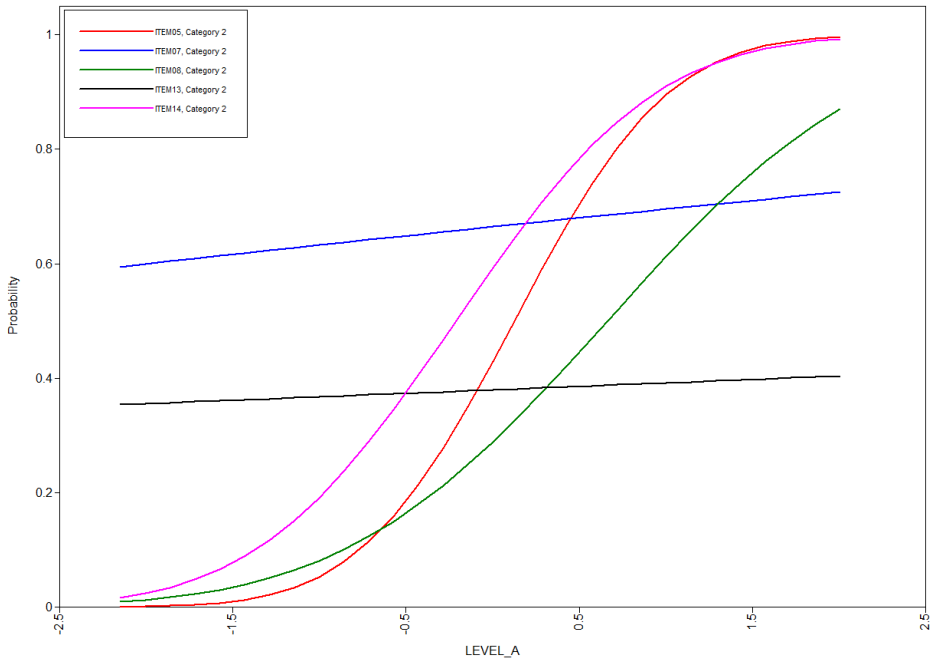
Curves 4: Item Characteristic Curves of Level A Items in Expert 1F1 Model



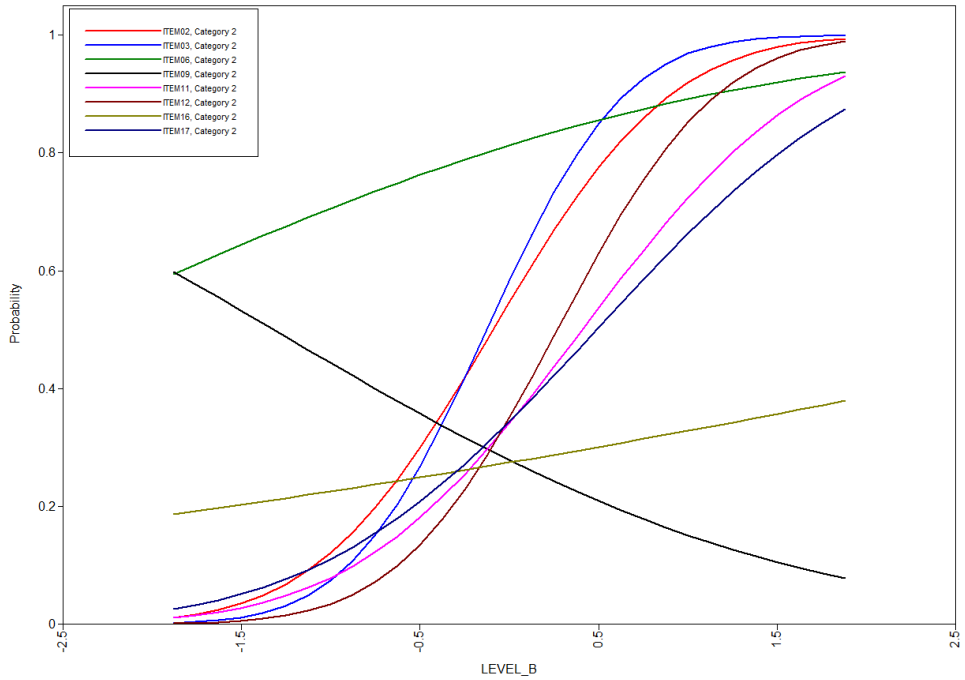
Curves 5: Item Characteristic Curves of Level B Items in Expert 1F1 Model



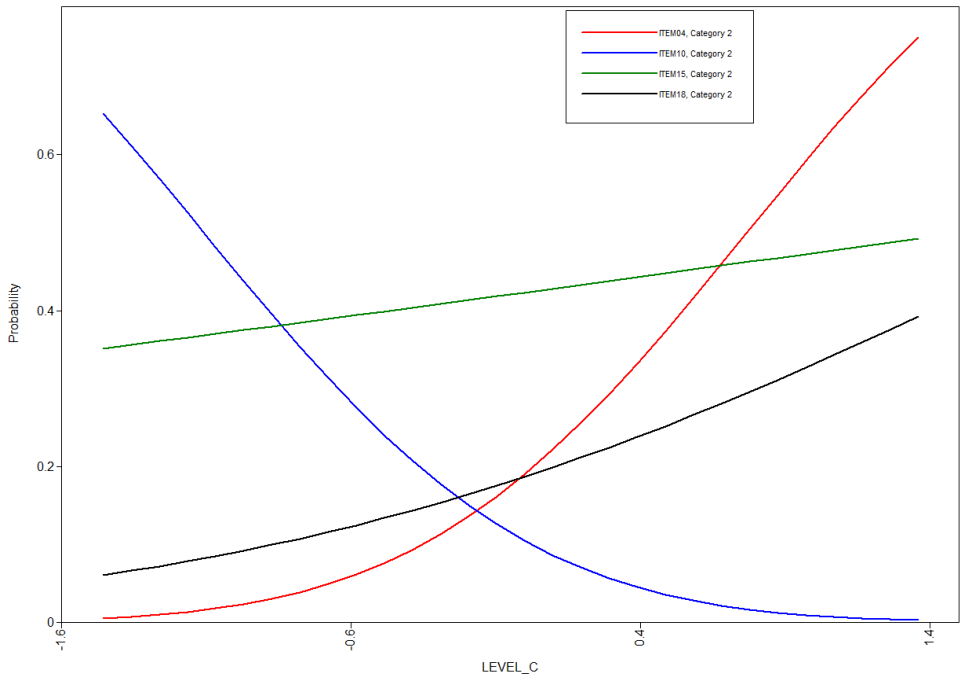
Curves 6: Item Characteristic Curves of Level C Items in Expert 1F1 Model



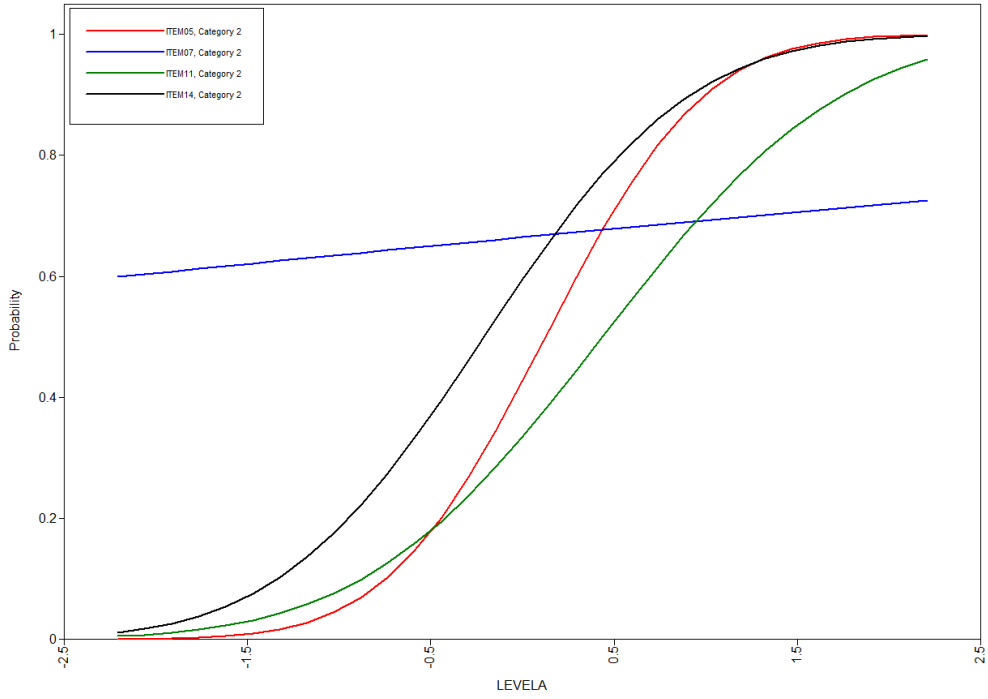
Curves 7: Item Characteristic Curves of Level A Items in Expert 2F1 Model



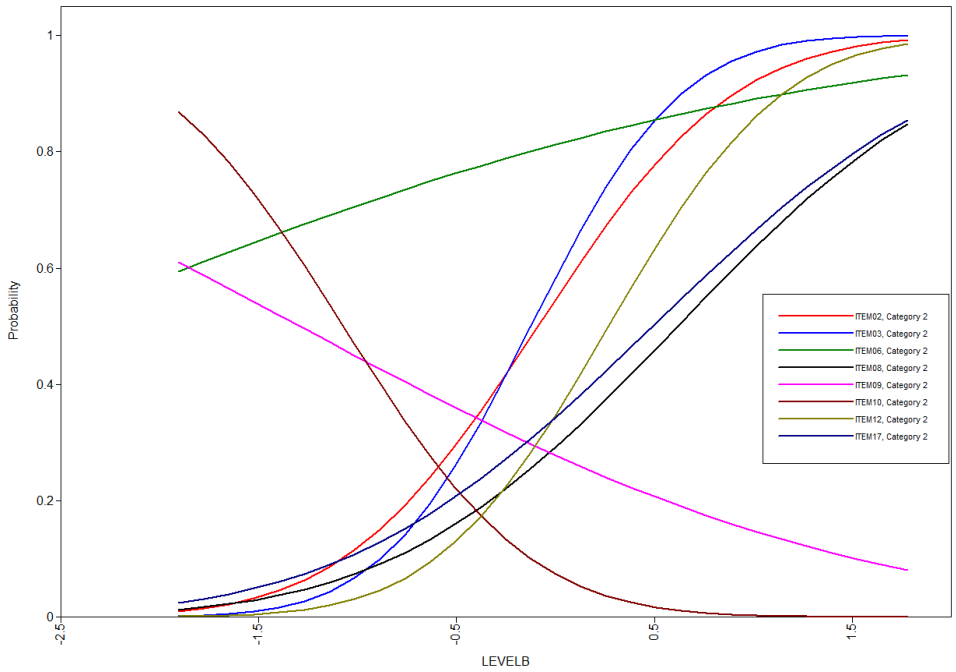
Curves 8: Item Characteristic Curves of Level B Items in Expert 2F1 Model



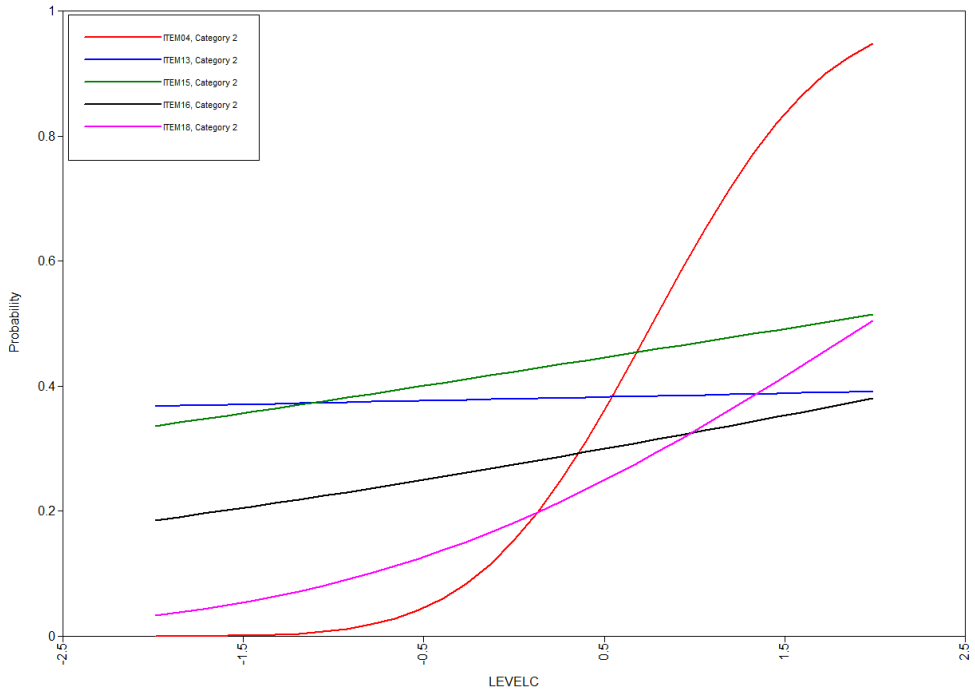
Curves 9: Item Characteristic Curves of Level C Items in Expert 2F1 Model



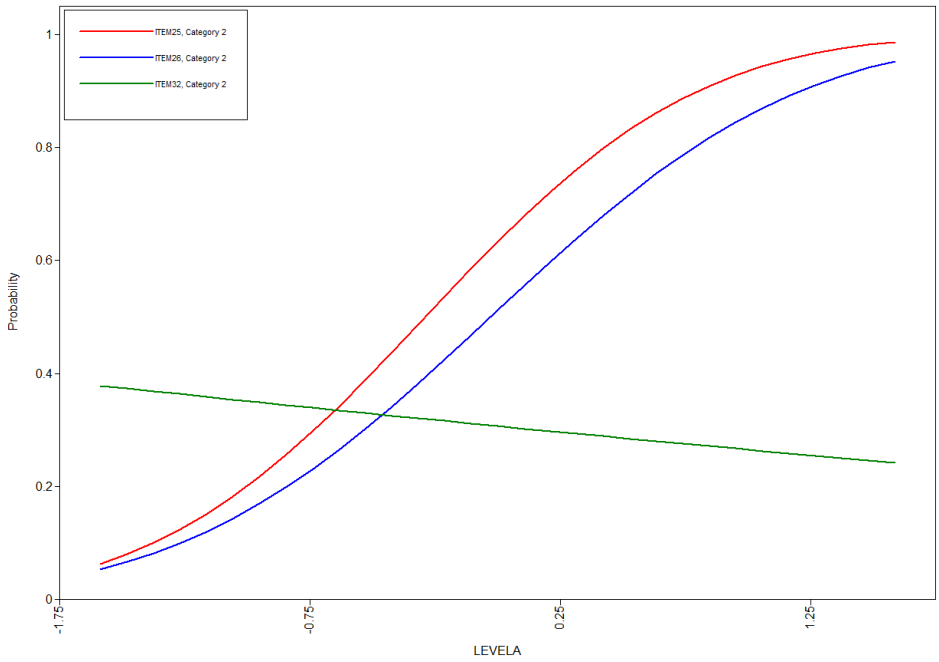
Curve 10: Item Characteristic Curve of Level A Items in Combination F1 Model



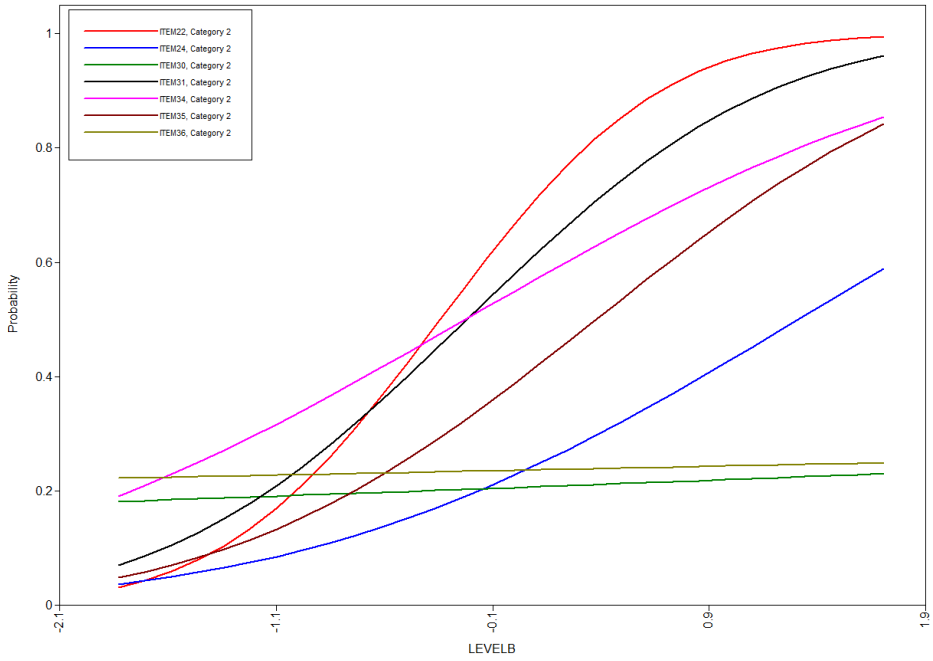
Curve 11: Item Characteristic Curve of Level B Items in Combination F1 Model



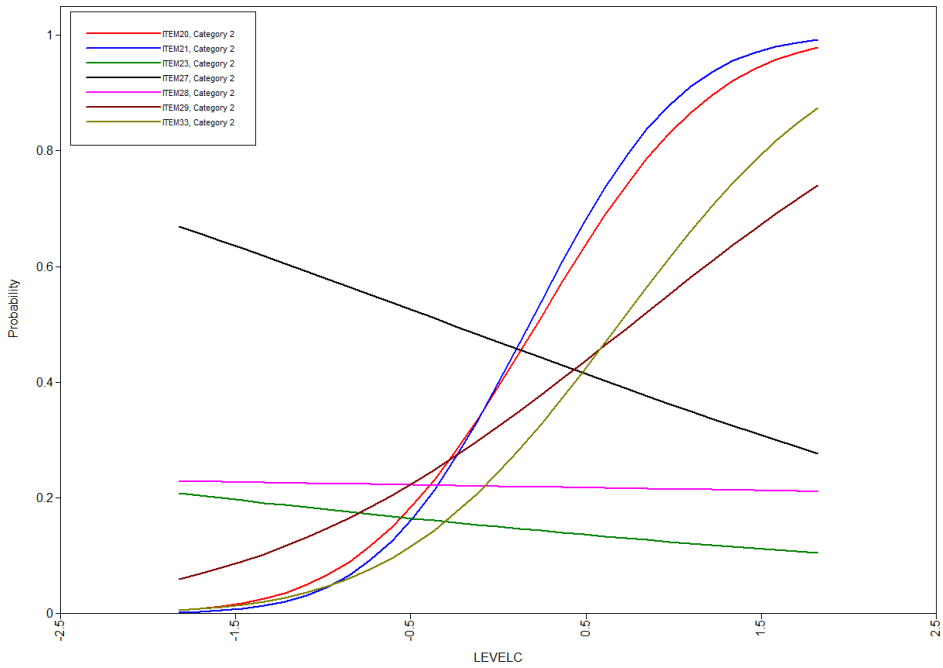
Curve 12: Item Characteristic Curve of Level C Items in Combination F1 Model



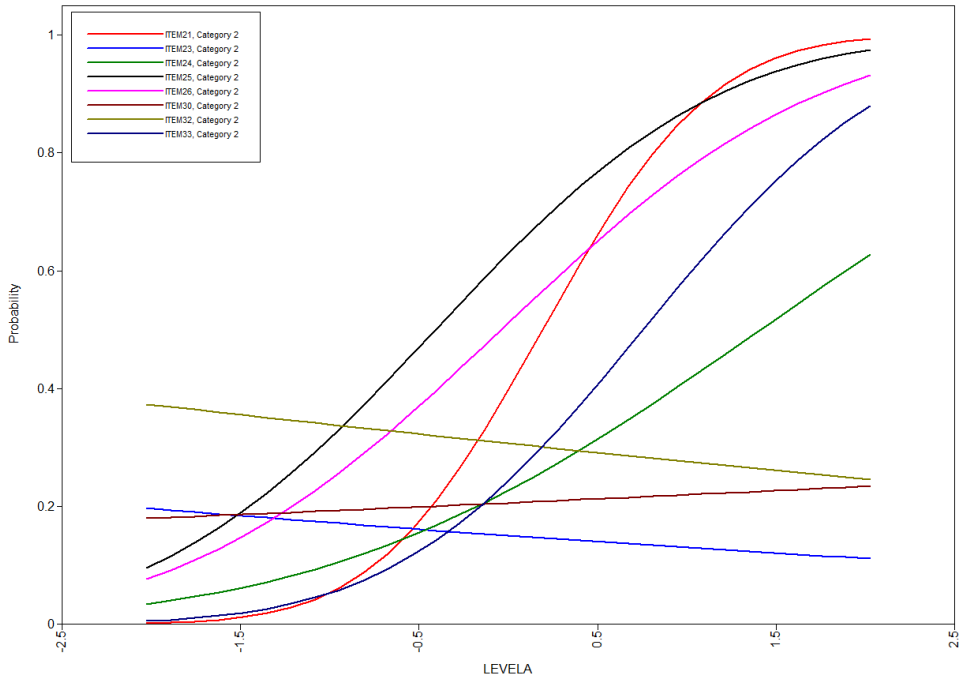
Curve 13: Item Characteristic Curves of Level A Items in Expert 1F2 Model



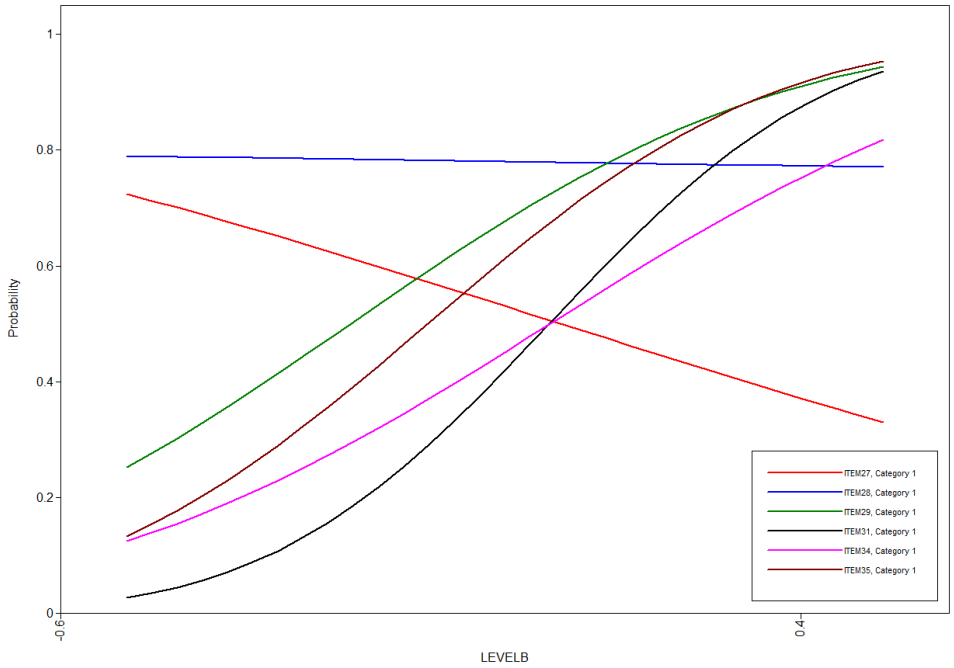
Curve 14: Item Characteristic Curves of Level B Items in Expert 1F2 Model



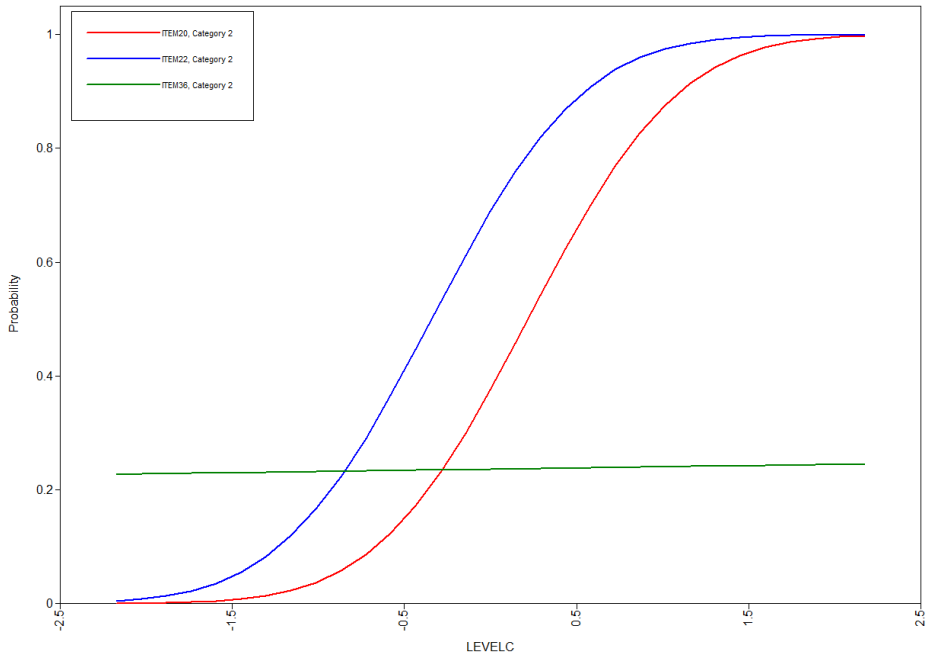
Curve 15: Item Characteristic Curves of Level C Items in Expert 1F2 Model



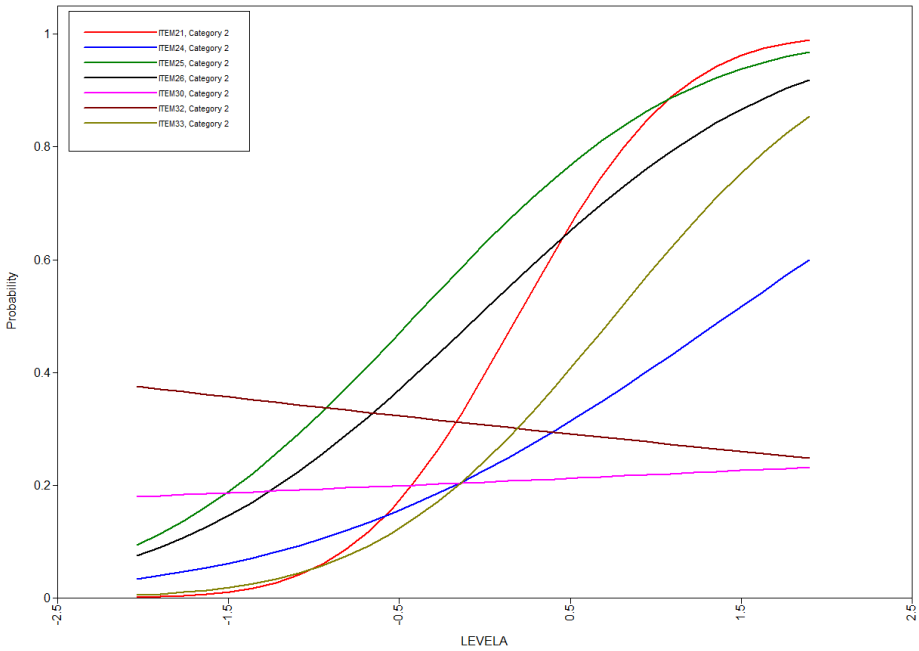
Curve 16: Item Characteristic Curves of Level A Items in Expert 2F2 Model



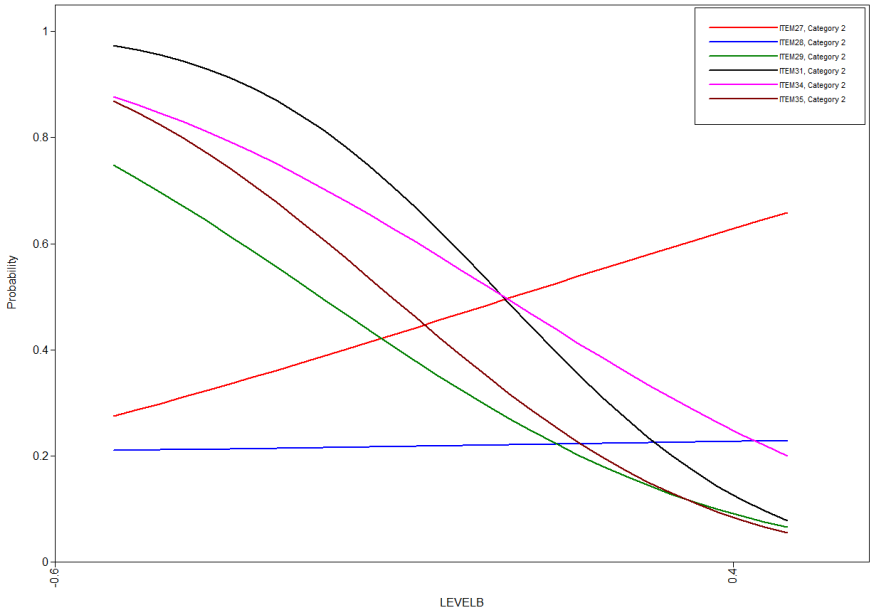
Curve 17: Item Characteristic Curves of Level B Items in Expert 2F2 Model



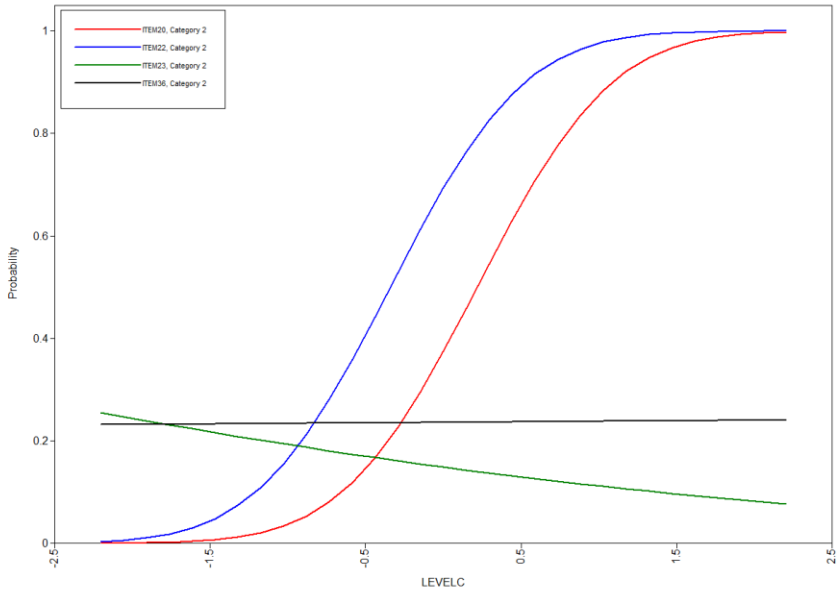
Curve 18: Item Characteristic Curves of Level C Items in Expert 2F2 Model



Curve 19: Item Characteristic Curves of Level A Items in Combination F2 Model



Curve 20: Item Characteristic Curves of Level B Items in Combination F2 Model



Curve 21: Item Characteristic Curves of Level C Items in Combination F2 Model

REFERENCES

Baker, F. (2001). *The Basics of Item Response Theory, (2nd ed.).* College Park, MD: ERIC Clearinghouse on Assessment and Evaluation (ERIC ED458219).

Bakker, A. G., & Hoffmann, M. (2005). Diagrammatic Reasoning as the Basis for Developing Concepts: A Semiotic Analysis of Students' Learning about Statistical Distribution. *Educational Studies in Mathematics, 60 (3),* 333-358.

Bentler, P. M. (2005). Latent growth curves. In J. Werner, *Zeitrehenanlysen* (pp. 13-36). Berlin: Logos.

Ben-Zvi, D., & Amir, Y. (2005). How do primary school students begin to reason about distributions? In K. Makar (Ed.), *Reasoning about Distribution: A collection of studies. Proceedings of the Fourth International Research Forum on Statistical Reasoning, Thinking, and Literacy (SRTL-4), Auckland, 2-7 July 2005.* Retrieved from: https://sites.google.com/a/edtech.haifa.ac.il/dani-ben-zvi-s-home-page/publications/articles-in-conference-proceedings.

Ben-Zvi, D., & Garfield, J. (2004). Statistical literacy, reasoning, and thinking: goals, definitions, and challenges. In D. Ben-Zvi, & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 3-15). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Bond, L. A. (1996). Norm- and Criterion-Referenced Testing. *Practical Assessment, Research & Evaluation, 5(2)*, Retrieved from http://PAREonline.net/getvn.asp?v=5&n=2 .

Bornstein, R. F. (2003). Face validity. In M. Lewis-Beck, A. E. Bryman, & T. F. Liao (Eds.), *The SAGE Encyclopedia of Social Science Research Methods* (pp. 367-368). Thousand Oaks, CA: SAGE.

Bornstein, R. F., Rossner, S. C., Hill, E. L., & Stepanian, M. L. (1994). Face Validity and Fakability of Objective and Projective Measures of Dependency. *Journal of Personality Assessment, 63(1)*, 363-386.

Bright, G. W., & Friel, S. N. (1998). *Interpretation of data in a bar graph by students in grades 6 and 8.* San Diego, CA: the American Education Research Association.

Bryne, B. M. (2010). *Structural Equation Modeling with AMOS: Basic Concepts, Application, and Programming. Second Edition.* New York, NY: Routledge.

Cai, J. (1995). Beyond the computational algorithm: Students' understanding of the arithmetic average concept. In L. Meira, & D. Carraher (Eds.), *Proceedings of the Nineteenth Annual Meeting of the International Group for the Psychology of Mathematics Education, Vol 3* (pp. 144-151). Recife, Brazil: Universidad Federal de Pernambuco.

Callingham, R. A., & Watson, J. (2005). Measuring statistical literacy. *Journal of Applied Measurement, 6(1)*, 19-47.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and Validity Assessment.* London: SAGE Publications, Inc. .

Chou, C. P., Bentler, P. M., & Satorra, A. (1991). Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: a Monte Carlo study. *British Journal of Mathematical and Statistical Psychology, 44*, 347-357.

Clark, J., Mathews, D., Kraut, G., & Wimbish, J. (1997). The Fundamental Theorem of Statistics: Classifying Student Understanding of Basic Statistical Concepts. *Unpublished manuscript*, http://www1.hollins.edu/faculty/clarkjm/papers.htm.

Clements, D., & Sarama, J. (2004). Learning trajectories in mathematics education. *Mathematics Thinking and Learning, 6(2)*, 81-89.

Cobb, G., & Moore, D. (1997). Mathematics, Statistics, and teaching. *American Mathematical Monthly*, 104, 801–823.

Conference Board of the Mathematical Sciences. (2001). *The Mathematical Education of Teachers.* Providence, RI; Washington D. C.: American Mathematical Society; Mathematical Association of America.

Confrey, J., Maloney, A., & Nguyen, K. (2010). *Learning Trajectory Display of the Common Core State Standards for Statistics.* New York: Wireless Generation, Inc.

Corcoran, T., Mosher, F. A., & Rogat, A. (2009). *Learning Progressions in Science: An Evidence-based Approach to Reform.* Philadephia, PA: Consortium for Policy Research in Education.

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

Curran, P. J., West, S. G., & Finch, G. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1,* 16-29.

Daro, P., Mosher, F. A., & Corcoran, T. (2011). *Learning trajectories in mathematics: A foundation for standards curriculum, assessment, and instruction.* Philadelphia, PA: Consortium for Policy Research in Education (CPRE).

delMas, R., Garfield, J., & Ooms, A. (2005). Using assessment items to study students' difficulty reading and interpreting graphical representations of distributions. In K. Makar (Ed.), *Proceedings of the Fourth International Research Forum on Statistical Reasoning, Literacy, and Reasoning (on CD).* Auckland, New Zealand: Retrieved from https://www.causeweb.org/artist/articles/SRTL4_ARTIST.pdf.

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

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Sceaffer, R. (2007). *Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report: A Pre-K-12 Curriculum Framework.* Alexandria, VA: American Statistical Association.

Friel, S. N., Curcio, F. R., & Bright, G. W. (2001). Making Sense of Graphs: Critical Factors Influencing Comprehension and Instructional Implications. *Journal for Research in Mathematics Education, 32 (2),* 124-158.

Gal, I. (2002). Adult's statistical literacy: Meanings, components, responsibilities. *International Statistical Review, 70(1),* 1-25.

Garfield, J. (1991). Evaluating Students' Understanding of Statistics: Development of the Statistical Reasoning Assessment. In R. G. Underhill (Ed.), *North American Chapter of the International Group for the Psychology of Mathematics Education, Proceeding of the 13th Annual Meeting, Blacksburg, Virginia, October 16-19, 1991* (pp. 1-7). Blacksburg, VA: North American Chapter of the International Group for the Psychology of Mathematics Education.

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

Garfield, J., & Ben-Zvi, D. (2007). How students learn statistics revisited: A current review research on teaching and learning statistics. *International Statistical Review, 75(3),* 372–396.

Garfield, J., & Gal, I. (1999). Teaching and Assessing Statistical Reasoning. In L. Stiff (Ed.), *Developing Mathematical Reasoning in Grades K-12* (pp. 207-219). Reston, VA: National Council Teachers of Mathematics.

Gould, R. (2004). Variability: One statistician's view. *Statistics Education Research Journal, 3(2),* 7-16.

Groth, R. (2003). High school students' levels of thinking in regard to statistical study design. *Mathematics Education Research Journal, 15*, 252-269.

Groth, R. E., & Bargagliotti, A. E. (2012). GAISE(ing) into the Common Core of Statistics. *Mathematics Teaching in the Middle School, 18(1),* 38 - 45.

Groth, R., & Bergner, J. (2006). Preservice elementary teachers' conceptual and procedural knowledge of mean, median, and mode. *Mathematical Thinking and Learning, 8*, 37-63.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their application to test development. *Educational Measurement: Issues and Practice, 12(3),* 38-47.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

IBM Corp. Released . (2012). *IBM SPSS Statistics for Windows, Version 21.0.* Armonk, NY: IBM Corp.

Jacobs, V. R. (1997). *Children's understanding of sampling in surveys.* Chicago: Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).

Jacobs, V. R. (1999). How do students think about statistical sampling before instruction? *Mathematics Teaching in the Middle School, 5*, 240-263.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: User's guide.* Chicago: Scientific Software.

Kim, J., & Mueller, C. W. (1981). *Factor Analysis: Statistical Methods and Practical Issues.* Newbury Park, CA: Sage Publication.

Kline, T. J. (2005). *Psychological Testing: A Practical Approach to Design and Evaluation.* Thousand Oaks, CA: Sage Publication, Inc.

Konold, C. (1989). Informal Conceptions of Probability. *Cognition and Instruction, 6*, 59-98.

Konold, C., & Higgens, T. (2003). Reasoning about data. In J. Kilpatrick, W. G. Martin, & D. Schifter (Eds.), *A research companion to principles and standards for school mathematics* (pp. 193-215). Reston, VA: National Council for Teachers of Mathematics.

Konold, C., & Pollatsek, A. (2002). Data analysis as a search for signals in noisy processes. *Journal for Research in Mathematics Education, 24*, 392-414.

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

Lecoutre, M. P. (1992). Cognitive Models and Problem spaces in "Purely Random" Situations. *Educational Studies in Mathematics, 23*, 557-568.

Mathews, D., & Clark, J. (2003). Successful Students' Conceptions of Mean, Standard Deviation and the Central Limit Theorem. *Unpublished manuscript*, http://www1.hollins.edu/faculty/clarkjm/papers.htm. .

McAlpine, M. (2002). *Principles of assessment. Bluepaper Number 1. February 2002. Robert Clark Centre for Technological Education.* Glasgow, Scotland: The CAA Centre.

Meletiou, M., & Lee, C. (2002). Student understanding of histograms: A stumbling stone to the development of intuitions about variation. *Proceedings of the Sixth International Conference on Teaching Statistics.* Durban, South Africa: Retrieved from https://www.stat.auckland.ac.nz/~iase/publications/1/10_19_me.pdf.

Mevarech, Z. (1983). A deep structure model of students' statistical misconceptions. *Educational Studies in Mathematics, 14*, 415–429.

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

Moore, D. S. (1990). Uncertainty. In L. A. Steen (Ed.), *On the shoulders of giants: new approaches to numeracy* (pp. 95-137). Washington D. C. : Mathematical Association of America.

Moore, D. S. (1992). Teaching statistics as a respectable subject. In F. S. Gordon, & S. P. Gordon (Eds.), *Statistics For the Twenty-First Century* (pp. 14-25). Washington, D.C.: Mathematical Association of America.

Moore, D. S. (1997). New pedagogy and new content: the case of statistics. *International Statistical Review, 65(2)*, 123-165.

Muthen, B. O. (2002). Beyond SEM: General Latent Variable Modeling. *Behaviormetrika, 29(1)*, 81-117.

Muthen, L. K., & Muthen, B. (2012). *Mplus Statistical Analysis With Latent Variables User's Guide.* Los Angeles, CA: Muthen & Muthen.

National Commission on Excellence in Education (NCEE). (1983). *A nation at risk: The imperative for educational reform.* Washington, DC: U.S. Government Printing Office.

National Council of Supervisors of Mathematics (NCSM). (1977). Position paper on basic skills. *Arithmetic Teacher, 25(1),* 19-22.

National Council of Teachers of Mathematics (NCTM). (1980). *An Agenda for action: Recommendations for school mathematics of the 1980s.* Reston, VA: NCTM.

National Council of Teachers of Mathematics (NCTM). (1989). *Principles and Standards.* Reston, VA: NCTM.

National Council of Teachers of Mathematics (NCTM). (2000). *Principles and Standards for School Mathematics.* Reston, VA: NCTM.

National Council of Teachers of Mathematics (NCTM). (2006). *Curriculum Focal Points for Kindergarten throgh Grade 8 mathematics: A Quest for Coherence.* Reston, VA: NCTM.

National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). *Common Core State Standards-Mathematics.* Washington D. C. : National Governors Association Center for Best Practices, Council of Chief State School Officers.

Nemirovsky, R. (1996). A functional approach to algebra: Two issues that emerge. In N. Dedrarg, C. Kieran, & L. Lee (Eds.), *Approaches to algebra: Perspectives for research and teaching* (pp. 295-313). Boston, MA: Kluwer Academic Publishers.

Nessom, J. T. (2012, May 12). *Jason Newsom's SEM Class (USP 655).* Retrieved from Portland State University: http://www.upa.pdx.edu/IOA/newsom/semclass/default.htm

Nevo, B. (1985). Face Validity Revisited. *Journal of Educational Measurement, 22 (4)* , 287-293.

Norusis, M. (2011). *IBM SPSS Statistics 19 Advanced Statistical Procedure Companion.* Englewood Cliffs: Prentice Hall.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory (3rd ed.).* New York, NY: McGraw-Hill.

Organization for Economic Cooperation and Development (OECD). (2009). *Take The Test: Sample Questions from OECD's PISA Assessment.* Paris: OECD.

Peck, R., Kader, G., & Franklin, C. (2008). Shaping K-12 statistics education in the United States. In C. Batanero, G. Burrill, R. C., & A. Rossman (Eds.), *Proceedings of the ICMI Study 18 and 2008 IASE Round Table Conference: Joint ICMI/IASE Study.* Voorburg, The Netherlands: ICMI, IASE, & ISI.

Pfannkuch, M. (1997). Statistical thinking: One statistician's perspective. In F. Bidduch, & K. Carr (Eds.), *Proceedings of the 20th Annual Conference of the Mathematics Education Research Group of Australasia* (pp. 406-413). Rotorua, NZ: MERGA.

Pfannkuch, M. (2006). Comparing Box Plot Distributions: A Teacher's Reasoning. *Statistics Education Research Journal, 5(2),* 27- 45.

Pfannkuch, M., & Reading, C. (2006). Reasoning about Distribution: A Complex Process. *Statistics Education Research Journal, 5(2),* 4-9.

Pfannkuch, M., & Wild, C. (2004). Towards an understanding of statistical thinking. In D. Ben-Zvi, & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 17-46). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Pollatsek, A., Konold, C., Well, A., & Lima, S. (1984). Beliefs underlying random sampling. *Cognition and Instruction, 12*, 395-401.

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

Price, L. R. (Preprint). Invesigating Perfect Model Fit in Higher-Order Confirmatory Factor Analysis: Exploratory/Confirmatory and Bayesian Approach. *The Proceeding of the American Psychological Association Annual Meeting July 31 - August 3, 2013.* Honolulu, HI: American Psychological Association.

Purpura, D. J., & Lonigan, C. J. (2013). Informal numeracy skills: The structure and relations among numbering, relations, and arithmetic operations skills in preschool. *American Education Research Journal, 50*, 178-209.

Reading, C., & Reid, J. (2006). An emerging hierarchy of reasoning about distribution: From a variation perspective. *Statistics Education Research Journal, 5(2),* 46-68.

Reading, C., & Shaughnessy, J. (2004). Reasoning about variation. In D. Ben-Zvi, & J. Garfield, *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (pp. 201-226). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Reading, C., & Shaughnessy, J. M. (2000). Student perceptions of variation in a sampling situation. *The Proceedings of the 24th annual meeting of the Intenational Group for Psychology and Mathematics Education.* Hiroshima, Japan: Retrieved from http://www.une.edu.au/ehps/resources/pdfs/creading/a.pdf.

Reys, B. J., Dingman, S., Sutter, A., & Teuscher, D. (2005). *Development of State-Level Mathematics Curriculum Documents: Report of a Survey.* Columbia, MO: Center for the Study of Mathematics Curriculum, University of Missouri.

Reys, B., & Lappan, G. (2007, May). Consensus or Confusion? The Intended Math Curriculum in State-Level Standards. *Phi Delta Kappan, 88 (9),* pp. 676-680.

Rossman, A., Chance, B., & Medina, E. (2006). Some important comparisons between statistics and mathematics, and why teachers should care. In G. Burrill (Ed.), *Thinking and Reasoning with Data and Chance; 68th NCTM Yearbook* (pp. 323-334). Reston, VA: National Council of Teachers of Mathematics.

Satorra, A., & Bentler, P. M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. *Proceedings of the Business and Economic Statistics Section of the American Statistical Association, 36,* 308-313.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye, & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399-419). Thousand Oaks, CA: Sage.

Shaughnessy, J. M. (1992). Research in Probability and Statistics: Reflections and Directions. In D. A. Grouws (Ed.), *Handbook of Research on Mathematics Teaching and Learning* (pp. 465-494). New York, NY: MacMillan.

Shaughnessy, J. M. (2003). Research on students' understandings of probability. In J. Kilpatrick, W. G. Martin, & D. Schifter (Eds.), *A research companion to Principles and Standards for School Mathematics* (pp. 216-226). Reston, VA: National Council of Teachers of Mathematics.

Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F. Lester (Ed.), *Second Handbook of Research on Mathematics Teaching and Learning* (pp. 957-1009). Greenwich, CT: Information Age Publishing, Inc. and NCTM.

Shu, L., & Schwarz, R. (2010). IRT Estimated reliability for tests containing mixed item formats. *Presented at the National Council on Measurement in Education May 2010.* Denver, CO: Retrieved from http://www.ctb.com/ctb.com/control/openFileShowAction?mediaId=17005.0.

Sorto, M. A. (2004). *Prospective Middle School Teachers' knowledge about data analysis and its application to teaching.* Ann Arbor, MI: ProQuest Information and Learning Company.

Strauss, S., & Bichler, E. (1988). The Development of Children's Concepts of the Arithmetic Mean. *Journal for Research in Mathematics Education, 19(1)*, 64-80.

Tavakol, M., & Dennick, R. (2011). Making sense of Crohnbach's alpha. *International Journal of medical Education, 2*, 53-55.

Texas Education Agency. (2011). *Texas School Directory.* Austin, TX: Texas Education Agency.

Torok, R., & Watson, J. (2000). Development of the concept of statistical variation: An exploratory study. *Mathematics Education Research Journal, 12*, 147-169.

Utts, J. (2003). What Educated Citizens Should Know about Statistics and Probability. *The American Statistician*, 57, 74–79.

Utts, J. (2010). Unintentional lies in the media: don't blame journalists for what we don't teach. *Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010)* (pp. 1-6). Voorburg, The Netherlands: Retrieved from https://www.stat.auckland.ac.nz/~iase/publications/icots8/ICOTS8_1G2_UTTS.pdf.

Vogt, W. P. (2007). *Quantitative Research Methods for Professionals.* Boston, MA: Allyn & Bacon.

Wallman, K. (1993). Enhancing statistical literacy: Enriching our society. *Journal of the American Statistical Association, 88(421)*, 1-8.

Watson, J. M. (1997). Assessing statistical literacy using the media. In I. al, & J. B. Garfield (Eds.), *The Assessment Challenge in Statistics Education* (pp. 107–121). Amsterdam, The Netherlands: IOS Press and The International Statistical Institute.

Watson, J. M., & Callingham, R. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal, 2(2)*, 3-46.

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

Watson, J., & Moritz, J. (1999b). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics, 37*, 145-168.

Watson, J., & Moritz, J. (2000a). Developing concepts of sampling. *Journal for Research in Mathematics Education, 31*, 44-70.

Watson, J., & Moritz, J. (2000b). Development of understanding of sampling for statistical literacy. *Journal of Mathematical Behavior, 19*, 109-136.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*, 223-265.

Zawojewski, J. S., & Shaughnessy, J. M. (2000). Data and chance. In E. A. Silver, & P. A. Kenney (Eds.), *Results from the seventh mathematics assessment of the National Assessment of Educational Progress* (pp. 235-268). Reston, VA: National Council of Teachers of Mathematics.

VITA

Rini Oktavia was born in Banda Aceh, Indonesia, the daughter of Farida Boerhan and Ramli Ibrahim. After completing high school at SMAN 3 Banda Aceh, Indonesia, she entered the Bandung Institute of Technology in Bandung, Indonesia. She received the degree of Sarjana Sains from the Bandung Institute of Technology in October, 1993. In August, 1995, she attended the graduate program in Mathematics at the Bandung Institute of Technology and graduated with a Master Sains degree in October, 1998. During her graduate studies, in December, 1995, she was chosen as a lecturer at the Education University of Indonesia in Bandung and worked there until June, 2003. In July, 2003 she moved to Syiah Kuala University in Banda Aceh, Indonesia. In August, 2007, she received a fellowship from the Ford Foundation International Fellowship Program to attend the graduate program in mathematics at The University of Texas at Austin where she earned her Master of Arts degree in mathematics in 2009. In August 2009, she started her doctoral program in mathematics education at Texas State University-San Marcos.

Permanent E-mail Address: roktavia@gmail.com

This dissertation was typed by Rini Oktavia.