April 2014

The Application of Information Integration Theory to Standard Setting: Setting Cut Scores Using Cognitive Theory

Christopher C. Foster
University of Massachusetts Amherst

THE APPLICATION OF INFORMATION INTEGRATION THEORY TO STANDARD SETTING: SETTING CUT SCORES USING COGNITIVE THEORY

A Dissertation Presented

By

CHRISTOPHER C FOSTER

Submitted to the Graduate School of the

University of Massachusetts Amherst in partial fulfillment

Of the requirements for the degree of

DOCTOR OF EDUCATION

February 2014

Education

THE APPLICATION OF INFORMATION INTEGRATION THEORY TO STANDARD SETTING: SETTING CUT SCORES USING COGNITIVE THEORY

A Dissertation Presented

By

CHRISTOPHER C FOSTER

Approved as to style and content by:

_____

Craig Wells, Committee Chairperson

_____

Stephen G. Sireci, Committee Member

_____

Aline Sayer, Committee Member

_____

Christine B. McCormick,

Dean of the School of Education

ACKNOWLEDGMENTS

I would like to thank all the faculty members in the department for being patient with me and working hard to help me improve and mature. Specifically I would like to thank Craig Wells, who helped me with most of my projects and always gave encouraging words. Finally, I would like to thank the people at both HP and Excelsior College for their contributions to this work. Without their generosity, the topic of my dissertation would have been quite different.

ABSTRACT

THE APPLICATION OF INFORMATION INTEGRATION THEORY TO STANDARD SETTING: SETTING CUT SCORES USING COGNITIVE THEORY

FEBRUARY 2014

CHRISTOPHER C FOSTER, B.A. WESLEYAN UNIVERSITY

Ed.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Craig Wells

Information integration theory (IIT) is a cognitive psychology theory that is primarily concerned with understanding rater judgments and deriving quantitative values from rater expertise. Since standard setting is a process by which subject matter experts are asked to make expert judgment about test content, it is an ideal context for the application of information integration theory.

Information integration theory (IIT) was proposed by Norman H. Anderson, a cognitive psychologist. It is a cognitive theory that is primarily concerned with how an individual integrates information from two or more stimuli to derive a quantitative value. The theory focuses on evaluating the unobservable psychological processes involved in making complex judgments. IIT is developed around four interlocking psychological concepts: stimulus integration, stimulus valuation, cognitive algebra, and functional measurement (Anderson, 1981).

The current study evaluates how IIT performs in an actual operational standard workshop across three different exams: HP storage solutions, Excelsior College nursing exam and the Trends for International Math and Science (TIMSS) exam. Each exam has cut scores set using both the modified Angoff method and the IIT method. Cut scores are evaluated based on Kane's (2001) framework for evaluating the validity of a cut score by evaluating procedural, internal and external sources of validity evidence.

The procedural validity for both methods was relatively comparable. Both methods took approximately about the same amount of time to complete. Raters for both methods felt comfortable with the rating systems and expressed confidence in their ratings. Internal validity evidence was evaluated through the calculation of reliability coefficients. The inter-rater reliabilities for both methods were similar. However, the IIT method provided data to calculate intra-rater reliability as well. Finally, external validity evidence was collected on the TIMSS exam by comparing cut score classifications based on the Angoff and IIT methods to other performance criteria such as teacher expectations of the student. In each case, the IIT method was either equal or outperformed the Angoff method.

Overall, the current study emphasizes the potential benefits IIT could produce by incorporating the theory into standard setting practice. It provided industry standard procedural, internal and external validity data as well provided additional information to evaluate raters. The study concludes that IIT should be investigated in future research as a potential improvement to current standard setting methods.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

Standard setting has grown from relative obscurity thirty years ago to a prominent topic in psychometrics today. Standard setting is the task of deriving levels of performance on education or professional assessments by which decisions or classification of persons can be made (Cizek, 1993). Methods of standard setting attempt to dichotomize a range of test performance into definable categories. These categories may be as simple as pass-fail or more elaborate as seen in the state of Massachusetts, which uses four categories: advanced, proficient needs improvement, and warning. Therefore, standard setting is the delineation of examinee performance to differentiate between degrees of performance on an assessment. Each of these performance categories are separated by a point on the score scale called a cut score. Cut scores are developed by following a system of rules defined by a particular standard setting method. Popular standard setting methods include the Angoff method (Angoff, 1971), the modified Angoff method (Angoff, 1971), the bookmark method (Lewis, Mitzel & Green, 1996), and many more. Standard setting varies widely in practice and is used in areas from educational settings to credentialing exams to licensure tests. However, some researchers have noted that different standard setting methods produce different cut scores on the same test (Jaeger, 1991).

One of the most important aspects of standard setting is its use in making decisions. Some of the earliest standard setting procedures appear in China as early as 2000 B.C. where it was used for military entrance. Kane (1994) cites a biblical record that recounts one of the earliest accounts of standard setting:

> Are you a member of the tribe of Ephraim?" they asked. If the man replied that he was not, then they demanded, "Say Shibboleth." But if he couldn't pronounce the H and said Shibboleth instead of Shibboleth he was dragged away and killed. So forty-two thousand people of Ephraim died there (Judges 12:5-6).

While standards set on tests today may not have stakes as high as those in this biblical passage, many tests are still considered high stakes assessments. High stakes assessments are tests that have important consequences for the examinee based on test score. For example, No Child Left Behind (NCLB, 2002) mandated high stakes assessments in educational programs across the nation. Often, a standard setting process is used to establish a pass/fail decision associated with high stakes testing. Since decisions associated with high stakes testing are frequently attached to a standard setting procedure, it is important that the procedure be accurate and well documented so decisions based on these standards are as fair and defensible as possible (Cizek, 2001).

1.1.1 Overview of Standard Setting

As previously defined, standard setting is the process by which cut scores are established that separate examinees into buckets based on definable performance categories. While the operational definition is simple and concise, the relationship between the operational definition of standard setting and the actual process in practice is much more difficult to define. Cizek (2001) stated that “psychometrics falls more along the lines of science, standard setting falls more into the social. Standard setting is perhaps the branch of psychometrics that blends more artistic, political, and cultural ingredients into the mix of its products than any other" (p. 5). This blend of science and art, politics and culture makes standard setting a very difficult and complex task that may results in inaccuracies.

Although there are many different standard setting methods, Hambleton and Pitoniak (2012) outlined nine essential steps to setting performance standards that are applicable to the majority of standard setting methods. While the authors proposed these steps as important criteria for defensible standards, they also provided a detailed summary of the standard setting process. The steps in order are described below.

1) Select a standard setting method and prepare for the first meeting of the panel.

In the first step of standard setting, it is important to select the type of standard setting method that will be used. Although some methods are more popular than others, each method serves a purpose and is applicable in certain situations. The majority of standard setting methods used today make judgments after reviewing assessment material and scoring rubrics (Hambleton et al., 2012). Hambleton et al. also mention that, in their personal experience, the method chosen is not as important as the implementation of the method because of various external biases that may influence cut scores such as training, panel, and administrator effects. The impact of these external sources of bias may come if an administrator controls the discussion in certain methods or a single panelist dominates the discussion during the standard setting workshop. If multiple panels are being used, then each panel facilitator needs to be trained so they manage their panels similarly. If panels are being facilitated in vastly different ways, there may be a large amount of variability across different panels due to a facilitator effect. The authors suggested that even the item presentation order may affect the outcome of the standards setting workshop.

2) Choose a large panel that is representative of stakeholders and a standard setting method for the study.

The second step is concerned with selecting an appropriate number of panelists that is representative of the stakeholders in the assessment. For example, the National

Assessment of Educational Progress (NAEP) has a diverse group of stakeholders, from educators to policymakers. For that reason, the panelists for the NAEP include 70% educators, further broken down into 55% classroom teachers and 15% other educators, and 30% non-educators (Loomis, 2012). The educators may come from teachers, school administrators, curriculum directors or many other educational professions. The non-educators include parents, policy makers, and employers (Loomis, 2012). As demonstrated by the diversity used for setting standards in the NAEP exam, it is important to select an appropriately diverse panel.

3) Prepare descriptions of the performance categories.

Many authors have noted that there is increased attention given to selecting and defining performance level descriptors (PLDs; Huff & Plake, 2010; Perie, 2008). The increased attention is a result of the increased attention received by performance standards as well as the important role that PLDs play in setting accurate and valid performance standards (Perie, 2008). In every standard setting process, PLDs convey information about performance categories and in some cases describe the candidate that is appropriate for the category. Raters in turn use this information to help anchor scale points in the psychological judgment process. The development of these standards may differ in length and specificity, but a performance standard will outline what an examinee needs to accomplish in order to obtain the standard.

4) Train panelists to use the method.

In order to obtain the most defensible and accurate standards possible, it is necessary to have an effective training for panelists. Panelists need to know about the standard setting methodology, the use of scoring rubrics, and the development of PLDs. Additionally, effective training may include practice rating sessions, taking practice tests,

reviewing the item pool, and even developing PLDs or descriptions of borderline candidates. It is not uncommon for training to take half a day or even more, depending on the complexity of the estimating process and description of the exam (Hambleton et al., 2012; Hein & Skaggs, 2009).

5) Collect ratings.

The fifth step described by Hambleton et al. (2012) is where many differences between standard setting methods are introduced. Raters review the information required by the standard setting method and provide the appropriate ratings. The process is relatively straight forward, if time intensive. This is often done privately at each panelist's discretion.

6) Provide panelists with feedback on their rating and facilitate a discussion.

During the sixth step, panelists review their ratings and receive feedback. The facilitator of the panel will often promote discussion among the panelists. This time is used for panelists to review and change their ratings if desired.

7) Compile panelist ratings again and obtain performance standards.

After each of the panelists has finalized his/her ratings, all of the ratings are compiled and used to obtain performance standards. This is done by whatever process is required by the standard setting method. While calculating the performance standards may be a relatively quick process, the amount of time and effort in collecting, compiling and discussing performance standards may be quite long. If panelist's judgments are paper based, then each panelist's ratings must be entered into a computer.

8) Conduct an evaluation of the standard-setting process and recommend performance standards.

In the penultimate step, raters are provided with feedback surveys and asked descriptive information on their feelings and experiences during the standard setting process. The recommended cut scores obtained through the standard setting process are forwarded to policy makers as recommended cut scores, which can either be accepted or changed by this group.

9) Compile technical documentation and validity evidence.

In the final stage of setting performance standards, the suggested cut scores have been submitted, but the standard setting process is still incomplete. It is still necessary to compile validity information on the standard setting process and the corresponding cut scores. While more detailed information will be provided in the literature review on validity issues in standard setting, there are several important sources of validity evidence that should be considered. Kane (2001) suggested three important sources of validity evidence that should be collected after a standard setting session is complete. The first is *procedural* evidence. Procedural evidence is the extent to which the implementation of a standard setting method is consistent and well documented. This includes documentation of the selection of candidates and the standard setting process. The second is *internal validity* evidence, which is the extent to which a method is consistent with itself. Internal validity includes the relevance of the chosen method, consistency within the method, inter-rater consistency, intra-rater consistency and across-panel consistency. Finally, *external validity* evidence is the comparison of cut scores to an external criterion. This form of evidence is important and includes comparing a new method with an established method, comparing final categories of students with external information about the examinees, and reviewing

the reasonableness of standards by investigating the proportion of examinees placed into each performance category.

Each of the nine steps provides an important function in standard setting, from selecting panel candidates to choosing a method. The defensibility of setting performance standards is greatly increased when each of these steps is implemented in the standard setting process. It should be noted that very few of the steps are actually collecting ratings and selecting a standard setting procedure. It is important that time is spent training panelists as well as collecting feedback on the procedure from the panelists. When developing new standard setting methodologies, it is important to investigate each type of validity evidence. Every standard setting process, including the method described in this paper, should adhere to these validity principles.

1.1.2 Information Integration Theory

Information integration theory (IIT) was proposed by Norman H. Anderson, a cognitive psychologist. It is a cognitive theory that is primarily concerned with how an individual integrates information from two or more stimuli to derive a quantitative value. The theory focuses on evaluating the unobservable psychological processes involved in making complex judgments. IIT is developed around four interlocking psychological concepts: stimulus integration, stimulus valuation, cognitive algebra, and functional measurement (Anderson, 1981). Each of these processes will be briefly described in this section and discussed in more depth in chapter II.

Stimulus Integration

How an individual internalizes and integrates information in thought is a core concept in IIT. It is rare for a thought or behavior to be predicted from a single predictor

variable or stimuli. The process of multiple sources causing a single behavior is called multiple causation (Anderson, 1981), and it is important to understanding how multiple variables are integrated to produce response. For example, when determining the loudness of a police siren, an individual might process the sound as two different stimuli: pitch and tone. Individuals may provide numerical judgments about the loudness of a sound differently based on changes in its tone and pitch, even if the decibel level remains constant. IIT studies how these variables are integrated and combined cognitively to form a final response.

Stimulus Valuation

Stimuli may either be physical or psychological. Physical stimuli can be observed and modified in experiments. Psychological stimuli are unobservable and it is difficult to assign a numerical value to these variables. IIT's dominant concern is with psychological variables and obtaining quantitative values from unobservable psychological processes. Valuation in IIT is the process by which an individual processes information and arrives at conclusions. Two different people may respond differently to the same colors or light patterns since the value the hue or color saturation differently. Different loudness can be interpreted from a sound for two people, even if the sound was the same pitch and intensity. Valuation underscores these individual differences to show that differences in opinion are present due to the psychological evaluation process.

Cognitive Algebra

Cognitive algebra is a byproduct of integration. Many studies on cognitive algebra have shown that information integration often follows very simple mathematical rules. In unobservable neural pathways, the human mind is multiplying, averaging, subtracting, or adding stimuli together to arrive at a final conclusion. Returning to the example of the

loudness of a siren, the perceived loudness of a police siren may be the tone of the siren multiplied by the pitch. In deciding how much an individual likes a president it may be as simple as adding all the approved platform agendas and subtracting all the bad platform agendas. When integrating information about motivation of workers, a manager may simply multiply the ability of an individual by their effort. Adding, subtracting, multiplication, and averaging are four simple algebraic models that have been used to demonstrate how individuals integrate multiple sources of information.

Functional Measurement

Functional measurement is the unification of several theories of psychological measurement. Inherent in the functional measurement theories are the psychophysical laws (valuation), psychological laws (integration), and psychomotor laws (responses) (Anderson, 1981). Each of these laws helps to evaluate how an initial physical stimulus is eventually converted into a numerical response. The psychophysical law investigates the relationship between physical stimuli and psychological qualities, like sensation and perception. The psychological laws employ cognitive algebra to combine the psychological qualities from the psychophysical law into a single, integrated judgment. The psychomotor laws apply to how the integrated psychological stimuli manifest in a physical or numerical judgment. A complete example will help solidify the concept of functional measurement and IIT. Suppose an individual wants to order a pizza. There are two factors that must be evaluated: the size of the pizza and the number of toppings. The person values information on the size of the pizza as fixed at $16 for a large. Similarly, the individual values a pepperoni topping at $2. This information is integrated using a cognitive algebra addition model. So the price of a large pepperoni pizza is equal to the price of a large pizza plus the price of a pepperoni topping. Therefore the final quantitative value for the price of a large

pepperoni pizza is $18. Although this example is simple, it provides information about a model that is currently used in decision theory and pizza pricing in the United States (Anderson, 1981).

IIT is a process whose purpose is to derive accurate quantitative values from the decision and judgmental process of raters. It uses statistical measures to validate equal interval scales that the judges are using and focuses on understanding the cognitive process of judges. Standard setting at its core is a judgmental task where raters are asked to provide quantitative values on a definable scale. The main focus and fundamental purpose of IIT appears as if it could be appropriately applied to standard setting.

1.4 Statement of the Problem

Mehrens and Lehmann (1991) highlighted the importance of standard setting by saying:

> Decision making is a daily task. Many people make hundreds of decisions daily; and to make wise decisions, one needs information. The role of measurement is to provide decision makers with accurate and relevant information… The most basic principle of this text is that measurement and evaluation are essential to sound education decision making." (p. 3)

On the same note, Hambleton (1978) stated "I cannot see how instructional decisions can be made without the use of cut-off scores" (p. 281). Hambleton's statement emphasized that for policy makers to make a decision on criterion-referenced test, cut-off scores must be established. Since then, many psychometricians have stated the importance of standards in the decision making process (Cizek, 2001; Jaeger, 1991; Kane, 2001). At the same time, millions of examinees are affected by standard setting on high stakes testing

each year, and cut scores may be the most salient feature on these tests. Because of the effect that standards have on decisions in high stakes testing, it is important that standards be accurate, well developed, and reliable.

However, Kane (2001) pointed out that cut scores are relatively arbitrary, depending on the method used, the quality of rater training, and several other reasons. He is not the only psychometrician to criticize standard setting methods (see Block, 1978; Camilli, Cizek, & Lugg, 2002; Hambleton, 1978; Linn, 1978). Jaeger (1991) provided a compelling argument that cut scores are used to dichotomize continuous data, but who is to say that any give cut score should not be a bit higher or lower. Policy makers can change suggested cut scores because of political or policy decisions, often to something with no statistical justification. Standard setting has been criticized for a lack of statistical justification (Jaeger, 1991) and policy assumptions by decision makers (Kane, 2001).

Due to its mixture of politics, measurement, and psychology (Cizek, 2002), standard setting is a frequently criticized feature of modern measurement. Despite the problems with standard setting methods, it is important to continue diligent research and to develop new, researchable methods that are grounded solidly in theory.

1.5 Purpose of Current Study

One weakness of modern standard setting methods is the lack of cross-discipline research in the area. Standard setting is primarily a psychological judgmental process (Jaeger, 1990), but psychological theory has never been utilized in a major standard setting method. The purpose of this study is to investigate the effectiveness of applying IIT, a method developed by a cognitive psychologist to help interpret individual judgments, to setting performance standards. In addition the study will evaluate the strengths and weaknesses of applying such an approach through the use of an experimental design where

rater responses and their corresponding cut scores are analyzed using Kane's (2001) approach to constructing a validity argument to support or discourage the use of IIT in standard setting practice. Such an argument would be potentially invaluable and inform test publishers, developers, and researchers to a new method of standard setting based in a cognitive theory.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter reviews the literature on the standard setting procedures, their applications, and their limitations. Additionally, this chapter addresses the literature on IIT, including its practical applications, and methodology. Specifically, this chapter can be outlined into the following four sections:

1. Information Integration Theory
2. Standard Setting Practice
3. Standard Setting Methods
4. Issues in Standard Setting

2.2 Information Integration Theory

The goal of information integration theory is to provide a unified, general theory of everyday life (Anderson, 2004). The generality of IIT spans from person cognition, cognitive development, decision theory, language processing and has been applied to an even wider variety of fields because IIT methods can adapt to each setting. One of the most important aspect of IIT is that it is founded in and reliant upon empirical evidence (Anderson, 2004; Weiss, 2006).

IIT is primarily concerned with how multiple sources of stimuli are internalized and combined, resulting in a single quantifiable response. However, to arrive at a final response, multiple sources of observable variables must be cognitively analyzed in three unobservable stages. In the first stage stimuli are interpreted, in the second stage stimuli are integrated, and in the third stage a response is constructed. These stages are collectively

known as the *problem of three unobservables* (Anderson, 2008). IIT hinges on understanding the underlying unobservable psychological processes that produce a response.

A solution does exist to understand what is occurring cognitively during each unobservable portion of IIT (valuation, integration, and response development). The discovery of cognitive algebra (Anderson, 1978) provided a key to quantitatively estimate these different unobservable variables. While cognitive algebra will be described in more detail later, its application to IIT has been shown in a wide variety of circumstances. The basic IIT process, as well as the problem of three unobservables, is highlighted in Figure 1. Three unobservable functions are indicated in the diagram: the valuation function, the integration function and the response function. In the basic flow of IIT, stimuli are first interpreted in the valuation stage, then the different sources of stimuli are combined during the integration stage and then a quantitative judgment is developed and expressed during the response stage.

2.2.1 Valuation

Defined simply, valuation is the process of extracting information from a physical stimulus and turning it into a psychologically derived value (Anderson, 1981). Multiple causation states that no reaction, thought or behavior is simply a function of a single stimulus but multiple coacting factors. Depth is a mixture of color, triangulation, size, and shadows (Howard, 2012). Perceived sound intensity is affected by both pitch and tone as well as other factors (Plack, 2005). It is helpful to think of valuation as a numerical weighting system of different stimuli in order to come to a final conclusion. For example, two people see the same light. However, both individuals weigh the hue and saturation of the light differently, therefore when asked about the intensity of the light respond with different answers. Valuation is the internal weighting of the different stimuli components.

The valuation function obviously involves a long chain of neural networks and cognitive processing and is therefore the *first unobservable*. However, the direction and magnitude of these neural networks are not the subject of the current investigation. It is important, however, to investigate certain aspects of the valuation function in order to obtain a better understanding of IIT.

2.2.2 Integration

As mentioned in the previous section, most responses are based on multiple interacting factors. It is rare to find one perfect predictor of behavior. Depth perception is an example that is studied frequently in cognitive psychology. Depth is a perception that involves perspective, size, texture, color, triangulation, and several other co-acting factors. Without the integration of all these complex variables, determining depth would be impossible. IIT attempts to analyze how these factors are integrated psychologically. Since integration, like valuation, is psychological, it is the *second unobservable*. It is physically impossible to observe the exact psychological processes of integration. However, it is possible to infer what is occurring using cognitive algebra and the use of quantitative methods of analysis.

The *third unobservable* is the response function and is directly linked to the integration of multiple stimuli. The response function refers to the psychological process of imposing numerical values on the newly combined information. During the third stage, after information is weighted and integrated, it is formulated into a response that can be expressed in an observable form. A response may be a sound, action, writing or any other observable response variable.

2.2.3 Cognitive Algebra

Cognitive algebra is a mental step nested within integration phase of IIT. Cognitive algebra is the process by which individuals combine multiple sources of stimuli into a single judgment using algebraic rules (Anderson, 1981, 2004, 2008). When combined with factorial design, cognitive algebra can be used to infer what is occurring psychologically with each of the three unobservables stages (valuation, integration and response processing). Using cognitive algebra and several well defined and empirically researched models, one can interpret how things are weighted during valuation and combined during integration (Anderson, 1996; Anderson, 2004; Weis, 2006). Norman Anderson (1978) identified and described many cognitive algebra models that can be interpreted from empirical evidence. However, the three most popular cognitive algebra models are the adding, averaging, and multiplication models. During the valuation stage, the individual places weights on each of the presented stimuli. During the integration stage, stimuli are either added, multiplied, or averaged together using the stimuli's weights to form an integrated response. For example, when valuing different ice-creams and toppings, a chocolate lover may place a high weight on chocolate ice cream and fudge topping. If the individual is asked to rate their preference of an ice-cream by topping combination on a scale of 1-20, they may give a weight of 5 to the chocolate ice-cream and a weight of 4 to the fudge topping. If the cognitive algebra process involved in this situation is a multiplication model, then the two values for the stimuli are combined multiplicatively. Using this process, a total value of 5 x 4 = 20, a maximum value on the 1-20 scale, is produced.

While seemingly simple, these cognitive algebra models have been shown to work in a wide variety of empirical settings. Butzin (1978) has shown that children use an adding

model when determining if someone deserves gifts. The equation used in this cognitive algebra task was Deservingness of gift = Achievement + Need of the individual receiving the gift. Graesser (1974) showed when rating a coworker's performance, the cognitive algebra performed was a multiplication of motivation and ability. When coworkers were asked to rate each other's performance, the resulting numerical judgments exhibited a pattern of a motivation score multiplied by an ability score. In both cases, information was combined in a predictable mathematical way.

The specific cognitive algebra models, as well as methods to detect each, will be discussed in more detail later. In addition, the benefits of detecting the cognitive algebra models will be discussed.

To conclude, when stimuli are integrated using cognitive algebra, information is combined in a predictable way. Therefore, detecting predictable integration patterns is a reliable way to determine which cognitive model is being employed. Most of the cognitive algebra detection methods are done through a visual analysis of the factorial graph through the use and inspection of a factorial design.

2.2.4 Factorial Design

The basic analysis and design tool for IIT is the factorial design (Anderson, 2004), which is widely used throughout psychology and other disciplines as a way to manipulate two or more variables. For cognitive algebra, specific cognitive algebra models are detected by the patterns they produce in a factorial design. In order to detect these patterns, it is important to analyze the patterns in the factorial graph.

The simplest factorial designs involve two different factors (or stimuli using the terminology of IIT), which can be arranged easily in a Row x Column matrix as shown in

Figure 2. Each cell in this matrix corresponds to a combination of factor A and factor B. A graph called the factorial graph can be constructed from a factorial design. An example factorial graph is displayed in Figure 3. The graph is constructed by placing the columns of the factorial table on the horizontal axis and the rows on the vertical axis of a Euclidian plane and graphing individual cell means. The row data points are then connected to form a curve. This factorial graphs is the main form of data presentation and analysis in IIT. Discovering patterns in these graphs helps diagnose the cognitive algebra rule, if it exists, that is being used to integrate different sources of information.

2.2.5 Functional Measurement

Functional measurement is the combination of the weighting factors in valuation, the integration of information using cognitive algebra, and finally outputting the result as a numerical response. This process is shown in Figure 1. In the diagram, S is a physical stimulus, ψ is the psychological value interpreted through valuation, I is the integration function, ρ is the integrated psychological stimuli, and R is the physical response from the produced from the integrated information. The figure reveals the three important functions integral to functional measurement:

$$V\{S\} = \psi \tag{1}$$

$$I\{\psi\} = \rho \tag{2}$$

$$I\{\rho\} = R \tag{3}$$

Equation 1, the valuation function, shows how the psychological valuation converts S, a physical stimulus, into ψ, a psychological variable. Equation 2 is the integration

function and takes each psychological value ψ from the valuation function and integrates them into a single response ρ. Finally, equation 3, the response or action function, converts the physiological ρ into an observable or quantitative response R.

One problem with validating this process is that the majority occurs psychologically and is therefore unobservable. While the true rationale for functional measurement lies in substantive theory, the final principal of functional measurement requires an empirical analysis. Information integration theory derives its name from the integration function in functional measurement where cognitive algebra is the key component. Anderson (1971, 1979, & 1991) asserts that IIT can only be valid if the algebraic models of stimulus integration are validated empirically. The essence of functional measurement lies in the empirical testing of the algebraic laws of cognitive algebra.

2.2.5.1 Adding Type Models

Adding type models occur when the values of observed stimuli are added together to produce the final response. For example, Anderson (1968) showed that when participants were asked to rate the overall impression of a random individual based on two adjectives, they simply added the value for both variables. While integrating the adjectives into an overall impression is complicated, it obeyed a simple adding process. This algebraic rule is inferred based on a parallelism analysis of graphical data. An example of observed parallelism is shown in Figure 3.

The concept of parallelism is simple. To test the hypothesis that two variables are being integrated additively, it is necessary to manipulate the stimuli into a factorial design. If the addition model is being used to integrate information, then the adding-type operation will produce a pattern of parallelism in the response data. Take the example given in Figure

3, where raters were asked to rate the impression of an individual based on a combination of two adjectives. The first adjective was gloomy, proud or courteous. The second adjective was worrier, thrifty or considerate. This 3 x 3 factorial design required each rater to make 9 distinct ratings based on every combination of adjectives. Figure 3 shows two factorial graphs for two different subjects. This graph helps reveal the nature of the integration procedure. As shown, the distance between each adjective's starting point and end point in comparison to the other adjectives remains constant, and all the lines are parallel to each other. This is a visual inspection of observed parallelism. While initially it seems that testing functional measurement is impossible because the three functions are unobservable, an analysis of the matrix of responses in a factorial design can help reveal and validate the true nature of the integration function.

There is an important proof for the parallelism theory that provides support for the use and existence of additive models. The proof focuses on the factorial design, where *i* and *j* are rows and columns, respectively.

$$P_{ij} = \psi_{Ai} + \psi_{Bj} \tag{4}$$

$$R_{ij} = C_0 + C_1 P_{ij} \tag{5}$$

Equation 4 shows an additive cognitive algebra model where ψ_{Ai} and ψ_{Bj} are being combined using simple addition. The equation also shows the addition integration function. Equation 5 shows the response function for linearity. Response linearity is important, as the factorial graph will reveal if the underlying cognition pattern is linear (Anderson, 2004). There are two premises, that if proven, show the algebraic adding rule to function correctly. The first premise is that the factorial graph will show observed parallelism. The second is that the marginal means of the rows will be a linear scale of ψ_{Ai}, and the column marginal

means will be a linear scale of ψ_{Bj} . The proof as given by Anderson for the first premise begins with equation 4 and continues:

$$R_{ij} = C_0 + C_1(\psi_{Ai} + \psi_{Bj}) \tag{6}$$

Now consider rows 1 and 2 of the factorial design:

$$R_{1j} = C_0 + C_1(\psi_{1i} + \psi_{Bj}) \tag{7}$$

$$R_{2j} = C_0 + C_1(\psi_{2i} + \psi_{Bj}) \tag{8}$$

Subtraction yields:

$$R_{1j} - R_{2j} = C_1(\psi_{1i} - \psi_{2i}) \tag{9}$$

The entire expression on the right of equation 9 is a constant, and this algebraic constancy is equal to graphical parallelism. Given this proof, if the graphical displays of the factorial data are parallel, then the graph displays *parallelism* and supports an additive model displayed in equation 4. Parallelism can also be supported statistically by the lack of a significant interaction in a repeated measures ANOVA

The second premise can also be proved algebraically beginning with equation 5 and continuing:

$$\bar{R}_{\bullet j} = \frac{1}{I}\sum_{i=1}^{I} R_{ij} \tag{10}$$

$$\bar{R}_{\bullet j} = \frac{1}{I}\sum_{i=1}^{I}[C_0 + C_1(\psi_{Ai} + \psi_{Bj})] \tag{11}$$

$$\bar{R}_{\bullet j} = \frac{1}{I}\sum_{i} C_0 + C_1(\frac{1}{I})\sum_{i}\psi_{Ai} + C_1(\frac{1}{I})\sum_{i}\psi_{Bj} \quad (12)$$

$$\bar{R}_{\bullet j} = C_0 + C_1\psi_{Ai} + C_1\psi_{Bj} \quad (13)$$

Since the first part is a constant, equation 13 reduces to:

$$\bar{R}_{\bullet j} = C_0^{'} + C_1\psi_{Bj} \quad (14)$$

Since $C_0^{'} + C_1\psi_{Bj}$ is a constant, $\bar{R}_{\bullet j}$, or the column mean, is equal to the column value on the right of the equation and shows linearity in the column means. The same logic holds true for the row means.

These two proofs provide valuable information about adding-type models. If the first proof is true, than the result will be a factorial table similar to Figure 2, and since the difference between levels is always a constant separates the resulting graph will exhibit observed parallelism. If the first proof is true then the second proof can also be proved and the scale raters are working with can be shown as equal interval. Thus, observed parallelism helps prove both equation 4 and equation 5 true. Additionally, if observed parallelism exists and the equations are true, there is a whole host of benefits:

1) support for the addition rule;

2) support for linearity (equal interval) of the response measure;

3) linear (equal interval) scales of each stimulus variable;

4) support for meaning invariance in the stimulus variables;

5) support for independence of valuation and integration (Anderson, 2004).

As previously discussed, observed parallelism offers strong support for an additive model. However, in fringe cases this may not always be true. If both assertions in equations 4 and 5 are true, then there will be observed parallelism. Similarly, If only one is true, then there will be no observed parallelism. However, if neither is true, then on the rare occasion, observed parallelism may occur due to chance in composite results across multiple raters. Results in this case should be validated or invalidated in other empirical studies and through an analysis of individual judgments.

It would be difficult to overemphasize the importance that observed parallelism shows support for a linear response scale. The pattern shown in the observed cells of the factorial design is a picture of an unobservable cognition pattern. Similarly, the scale values which guided the response processes are cognitively conceptualized by the rater as a linear, equal interval scale. Thus, the scale values used in the factorial design are a simple linear transformation from any other scale and changes in the scale have equal meaning. Linearity allows the response scale to be linear transformed to any other scale values.

Finally, observed parallelism shows that each stimulus is independent of other stimuli and has meaning invariance. For example, in Figure 3, the adjective *considerate* has the same scale value despite its combination with a variety of other adjectives. Considerate is meaning invariant, meaning its scale value has a fixed meaning within rater cognition.

The adding model, shown by observed parallelism in the factorial graph, provides important characteristics to the response scale. Equal interval scales and independence of stimuli are desirable in the majority of disciplines. It is important to note that observed parallelism and the adding model have been proven empirically in a wide domain of content areas. Anderson (1962) showed that human judgments of adjective traits follow this pattern. The additive model has been shown to function in decision theory (Anderson,

1991), self-estimation attribute evaluations (Zalinski, 1991), attitude (Anderson, 1971), inequity evaluations (Farkas, 1971), fairness evaluations (Farkas, 1991), and poker evaluations of risk and reward (Lopes, 1987). While dozens more cases of observed parallelism in empirical research could be cited, adding models are applicable in a variety of situations.

2.2.5.2 Multiplication Models

The multiplication cognitive algebra model, like the addition model, appears to be natural in many cognitive integration processes (Anderson, 1996). For example, a simple multiplying model that is used frequently in economics and statistics is that of expected value (EV). The basic equation in economics is: EV = Probability x Value. However, a study of the multiplicative rules requires methods for testing these cognitive algebra steps.

The basic tool in analyzing multiplication rules is the *linear fan* (see Figure 4). Just as observed parallelism is indicative of an additive model, a linear fan indicates a multiplication model. The basic multiplication model rests on two premises:

1) $P_{ij} = \psi_{Ai} \times \psi_{Bj}$ (Multiplication)

2) $R_{ij} = C_0 + C_1 P_{ij}$ (Linearity)

Both of these equations are proven in a similar way to the parallelism premises seen in equations 4 and 5. From these premises come two conclusions. The first conclusion is that the factorial graph will appear as a linear fan. The second conclusion is that the marginal means of the factorial table will be a linear (equal interval) scale.

Anderson (1981, 1996) mentions that in order for the linear fan to be visible, the factorial graph must be constructed appropriately. The graph must be constructed in such a

way that the spacing on the horizontal axis is equal to their subjective values. It is necessary to arrange the stimuli according to the column marginal means and place them on the horizontal axis in this order. If the multiplication rule is true, then linear fan pattern will appear, as shown in Figure 4. However, if the multiplication rule is false, then the factorial graph will not be a linear fan.

The linear fan theorem provides a simple test for the multiplication rule. An observed linear fan provides strong support for both premises of the multiplication theorem. Similar to the additive model, Anderson (1996) described several benefits to an observed linear fan:

1) support for the multiplication rule;
2) support for linearity in the response scale;
3) linear scales of each stimulus variable;
4) support for meaning invariance;
5) support for independence of valuation and integration.

Each of these benefits have been discussed previously section 2.2.5.1. However, the second and third benefits, those of linearity, should be re-emphasized. When there is an observable linear fan, the response measure is conceptualized cognitively as a linear scale. Differences in the scale have true meanings, and the scale itself has established validity evidence. Therefore, the detection of a linear fan provides validity evidence of the rater scale responses.

Similar to the additive model, it is unlikely but possible that a linear fan appears in the data when a multiplicative rule does not exist. If a linear fan appears in the aggregated data across participants, then the factorial graphs for each individual should be investigated. Rare combinations of non-linear fan data on the individual may produce a

linear fan occasionally by chance. A significant interaction from repeated measures ANOVA will also support the observable linear fan.

Figure 4 provides a near perfect example of a linear fan. Shanteau and Nagy (1976) asked females to rate the attractiveness of going on a date with a simulated individual by combining the physical attractiveness of the date and the probability of going on a date with them. Each subject was presented with a picture of a person and given the probability ranging from low (.05) and high (.95) that the person would ask the subject on a date. The subject then gave a numerical judgment about the relative attractiveness of going on a date with the presented individual. The integration of these two stimuli resulted in a multiplicative pattern. The date attractiveness was equal to the probability of being asked on a date multiplied by the attractiveness of the person in the picture. When this information was graphed it produced an observable linear fan.

2.3 Standard Setting Practice

2.3.1 Performance Levels

Performance level descriptors (PLDs) are frequently used in standard setting procedures. While performance standard is generally used to define the pass/fail categorical data applied to a standard setting procedure, performance levels provide multiple evaluative categories (Haertel, 1999). Egan, Schneider, and Ferrara (2012) describe PLDs as "the knowledge, skills and processes (KSPs) of students at specified levels of achievement and often include input from policy makers, stakeholders and SMEs" (p. 79). Kane (2001) explains that the purpose of a standard setting method is to convert PLDs to appropriate cut scores.

The literature surrounding PLDs greatly increased throughout the 1990s (Egan et al., 2012). This was in part because of the first well-known use of PLDs with the 1992 NAEP standard setting. In 2002, NCLB required states to develop PLDs to use in standard setting and score reporting. One concern about using PLDs in standard setting was the difficulty in setting multiple cut scores (one for each PLD) using current standard setting methods (Egan et al., 2012).

PLDs usually define categories that describe examinee performance. In turn, examinee performance is frequently reported as a PLD. Practitioners, educators, parents and examinees may all interpret these performance categories differently (Hambleton & Slater, 1997). Recent research (Burt & Stapleton, 2010) showed that even SMEs working on the same standard setting panel interpret different performance categories differently. This indicates that PLDs deserve validation research and should be thoroughly addressed during the standard setting workshop.

2.3.2 Cognitive Process of Standard Setting

Many standard setting procedures incorporates raters' judgments into the computation of cut scores. The collective contribution of experience and intelligence of a group of SMEs is usually the most influential factor on the setting of performance standards. Because of the importance of rater's cognitive decisions in standard setting, many authors have focused on the difficulty of the cognitive task required by panelists (Impara and Plake, 1998; Impara, 1998). However, since rater judgments require a cognitive task, it is very difficult to monitor what is occurring in the neural pathways of the brain. Despite this difficulty, understanding the cognitive process of SMEs is a growing body of literature in standard setting (Brandon, 2004; Hurtz & Auerbach, 2003; Dawber, Lewis, & Rogers, 2002; Egan & Green, 2003).

The cognitive process for every SME can be a very difficult task in many standard setting procedures. SMEs must begin by internalizing performance level descriptors (PLD), which can include long lists of what candidates in this performance level can or cannot accomplish. Next, the SMEs must conceptualize not only a student that conforms to each category, but the borderline or minimally competent candidate (MCC) for each category as well. Imagining the MCC is again a complex task that requires candidates to be placed in performance categories within each PLD. For example, raters may conceptualize the minimally competent candidate in comparison to, the competent examinee, and the excellent examinee in the same PLD. Conceptualizing the MCC has been shown to be a difficult task for SMEs (Hein & Skaggs, 2010; Mills, Melican, & Ahluwalia, 1991). Hein and Skaggs (2010) showed that SMEs had a very difficult time envisioning these hypothetical MCCs. Skorupski (2012) points out that even when candidates are comfortable with PLDs, they still must define borderline performance level descriptors as well. SMEs have a difficult time imagining the combination of *minimally competent* with performance categories. Plake (2008) reported that there is little to no research on how the complexity of the cognitive task increases when multiple PLDs and cut scores are being used. However, Skorupski (2012) indicated that it is reasonable to assume that the task does increase in complexity when multiple cut scores are being suggested.

Not only must SMEs struggle with the conceptual task of imagining MCCs, but the understanding of MCCs interacts with the chosen standard setting method. The majority of the research focuses on how SMEs have difficulties understanding specific tasks related to standard setting methods such as the Angoff or Bookmark. The Angoff method(1971) requires SMEs to estimate p-values for a MCC. A p-value is an estimate of item difficulty and describes the proportion of examinees who answered an item correctly. While a seemingly simple task, research has shown (Impara & Plake, 1998) that panelists have a very difficult

time estimating the probability groups of examinees will get the item correct. This task is even more problematic when estimating item difficulties for MCCs and PLD. Since the cognitive task associated with the commonly used Angoff method was so difficult, many other popular methods were developed, such as the Bookmark. These new methods claim to be less cognitively complex (Lewis, Mitzel & Green, 1996). However, even the bookmark suffers from difficulties in conceptualizing the cognitive task (Plake, 2008).

While work has been done to evaluate the difficulty of the cognitive standard setting task, no research has been conducted to actually analyze the cognitive processes at work in the SME. The research does show that panelists have a very difficult time understanding the concept of the MCC, especially when pairing the MCC with multiple performance levels. Such difficulties call into question the use of MCCs in the standard setting process (Skorupski, 2012).

2.3.3 Subject Matter Expert Training

While cut scores set from different standard setting methods may differ (Jaeger, 1989), training for different methods may be relatively similar. Raymond and Reid (2001) outlined three important steps for effective standard setting training:

1) delineation of the task required of the panelist,

2) identification of the knowledge and skills underlying the panelist's task,

3) development of instructions so the panelist can acquire these knowledge and skills.

To establish these goals of effective training, it is necessary to describe the standard setting process, establish the context, develop a definition of the reference group, and teach panelists the skills required to make accurate judgments (Mills, 1995).

While each individual standard setting practice will differ based on panelists' personalities and test content, several training operations remain constant. First, the context of the exam should be explained (Raymond and Reid, 2001). Participants should understand the purpose and scope of the exam. The authors also noted that access to information about the test construction may also benefit ratings. The panelists should also be encouraged to talk about the consequences of passing or failing the exam, or ending up in each performance category.

Before panelists can begin the standard setting task, it is necessary to have definitions of the different performance levels. Defining the performance levels during training may help panelists internalize them. These descriptions may be range from very general to very specific (Cohen, Kane & Crooks, 1999). Kane (1998) suggested that it is possible to define the performance levels outside the standard setting operation, but it is still beneficial to discuss these performance levels with panelists.

The next step in the training process is practicing the standard setting task in a similar way to what will be done during operation standard setting. The materials in the practice should be the same as the operational context (Impara & Plake, 1997). Practice items should follow the same distribution of content as the actual exam (Kane, 1998). This practice session allows SMEs to conceptualize the problem and gain a better understanding of the process and rating scale. The majority of standard setting training will include these steps (Raymond & Reid, 2001).

Three ways have been suggested to establish if training has been effective. (Berk, 1996; Mills, 1995; Reid, 1991). The first is that panelists' ratings are stable over occasions. If a panelist gives a rating for a specific performance level for a specific item, then the panelist should give a similar rating if the same pairing were given a second time. If panelists are inconsistent with themselves beyond a reasonable margin of error, then there are issues with the method. These issues may come from a lack of understanding of the standard setting procedure or poor training (Loomis, 2012). The second way of determining if training was effective is if there is consistency with assumptions of the method. For example, the Angoff method assumes that panelists can accurately make a probability judgment about minimally competent examinees in specific performance levels. Examinees with adequate training should be able to make accurate judgments. If examinees cannot perform this task, then perhaps the training was not effective. The third method of evaluating training is if the cut scores reflect realistic expectations. While defining realistic expectations is a subjective process, final cut scores should fall within a range of acceptable outcomes. Reid (1991) highlighted an extreme example. If a cut score produced a fail-rate of 100% in empirical data, this may be the result of poor training being manifest in an inaccurate cut score. However, it could also be because there were no competent examinees in the testing group.

Effective training is applicable to every standard setting method. While small differences in training may exist between methods, poor training in any circumstance will undermine the accuracy of a cut score. Panelists must understand the process in order to produce the most accurate cut scores, and understanding the process begins with effective training (Kane, 1998).

2.3.4 Reviewer Feedback

The final step of standard setting, as outlined by Hambleton et al. (2012), is to collect evaluations of the standard setting process as well as performance standards. This process is done by surveying the SMEs and other participants of the standard setting workshop. Cizek (2012) stressed that collecting this information is a key component to completing a standard setting workshop and can provide important validity information. In addition, the surveys can allow current SMEs to help inform future standard setting workshops in the content area.

Cizek also outlined the four different functions of the standards setting evaluations: 1) Formative, 2) Summative, 3) Policy Informing, and 4) Knowledge and Theory Advancement. The formative portion of the evaluation is to inform the current standard setting workshop. It is therefore important that panelists are given a chance to provide feedback during the standard setting process. The purpose of the summative evaluation is to gather appropriate forms of validity evidence from the panelists. This information includes the participant's view of the standard setting process, their opinions of the fairness of standard setting, and that the process was conducted appropriately. The third purpose, policy informing, relays information from panelists to the policy makers who decide to accept or change the suggested standards. Since a standard setting panel usually only recommends standards, information provided by the evaluation to the policy makers may help inform policy makers about accepting the proposed standards or making revisions. Finally, the fourth purpose of evaluations, knowledge and theory advancement, provides information about ways that the current methodology may be improved for future studies. The survey evaluation questions typically address these four different categories and

ultimately provide important validity evidence for current and future standard setting operations.

2.3.5 Validity of Standard Setting

The *Standards for Educational and Psychological Testing* states that "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed use of tests" (p. 9). Tests themselves are not validated; however, validity is a property associated with the interpretation of test scores. Just as tests are not validated, cut scores are not validated. Kane (2001) states "Just as we do not validate a test but rather the interpretation assigned to test scores, we do not validate a cut score or a performance standard in isolation. Rather, we evaluate the appropriateness of the performance standards, given the general purpose of the decision process. The aim of the validation effort is to provide convincing evidence that the cut score does represent the intended performance standard and that the performance standard is appropriate" (p. 57).

It is important to compile validity evidence to support the standard setting process and the proposed cut scores. Setting performance standards has a large impact on student scores, and even a small change in the location of a performance standard may have a large impact. As student raw scores are converted into an ordinal measure of performance, these performance categories are given meanings and have consequential outcomes, and then the consequential outcomes are interpreted. These consequential outcomes can be as varied as graduating from high school, receiving a medical license, or being approved to work as an accountant. Each outcome has high consequences for the examinee. For this reason, it is necessary to compile validity evidence to support the intended use and interpretation of performance standards and their corresponding cut scores.

Kane (2001) suggests three types of validity evidence that should be evaluated between performance standards and cut scores: procedural evidence, internal consistency evidence, and the agreement with external criteria.

2.3.5.1 Procedural Evidence

Procedural evidence refers to the appropriateness of the procedures used in the standard setting process and the completeness of the compiled information. Procedural evidence is especially important because of the limitations of adequately collecting validity evidence using empirical methods (Kane, 2001). In practice, procedural evidence is often considered adequate support for standard setting decisions. Poor procedural evidence makes a standard setting method difficult to defend and damages the confidence in cut scores.

The *Standards for Educational and Psychological Testing* are not specific on what standard setting procedures are applicable to use in the standard setting processes. However, the standards do give suggestions on properties of the method. The method should have "sound scientific basis" (p. 43). In addition, the 1985 standards state that the method should be "well documented, be based on an explicable rationale, be public, be replicable and be capable of producing a reliable result" (p. 15). Any method that satisfies these requirements is an appropriate method. However, the idea that different standard setting method yields reliable results is the subject of criticism. Jaeger (1989) concluded that standards set on the same test using different procedures often produce inconsistent results. This lack of consistency across methods is disturbing, as it shows that different standards may be set based entirely on whichever standard setting method is chosen. Additionally, numerous studies have shown the strengths and weaknesses of various standard setting methods (Clauser et al., 2009; Impara & Plake, 1998); however, there is no

general consensus as to which standard setting procedure produces the best results. Kane (2001) points out that this is because there is no perfect external criteria to use as a point of comparison for standard setting methods. While the "best" standard setting method remains a mystery, there is agreement that the cut scores should be set in a meaningful and systematic way. Kane (2001) described five different steps in the standard setting process that have an important impact on the compilation of procedural evidence:

1) Definition of Goals
2) Selection of Participants
3) Training
4) Definition of Performance Standard
5) Data Collection Procedures

Several of these areas of validity evidence require little explanation. Goals for the standard setting procedure should be well thought out and defined. Participants should be selected from a range of candidates who have a stake in the accuracy of the cut scores. The candidates should also be capable of performing the standard setting task. While the first steps are simple to explain, more literature exists emphasizing the importance of the final three steps.

A large body of literature exists that stresses the importance of training participants. Loomis (2012) pointed out that all participants should get thorough training in the standard setting process. This training should include details on how cut scores will be set, the importance of accurate ratings, an accurate description of the test, and even the opportunity to take the test themselves (Mills, Melican, & Ahluwalia, 1991). In addition to a thorough description of the task, participants should be allowed to practice setting standards to get a better feel for the task and receive feedback from the administrators (Reid, 1991). Other

researchers have focused on re-training participants at given intervals during the standard setting process if necessary (Plake, Melican & Mills, 1991).

Kane (2001) mentioned that defining the performance standards is usually not given the attention that the task deserves. Often policy makers believe that 'performing at a fourth grade level' is a construct that is understood by everyone. Often vague references or gaps between performance levels result in unsolved ambiguities that pollute the standard setting process. The defensibility of cut scores is likely to be improved when the definitions for the performance standards are clearly stated and participants agree on the definitions (Kane, 2001).

2.3.5.2 Internal Consistency

One important aspect of validity information that must be addressed in standard setting is the consistency of the standard setting results. While consistency of results is not the best source of validity evidence and justification for the interpretation of the cut score, it does help justify the use of the score. It is difficult to have confidence in a method that does not produce consistent results on the same test (Kane, 2001).

One way to evaluate the internal consistency of a method is to obtain an estimate of the standard error for the cut score. There are two approaches to obtain the estimate of the standard error with most standard settings methods. The first is to convene multiple panels and compare the results across different panels. Some difference is expected due to rater backgrounds (Plake et al., 1991) and different populations (Jaeger, 1991), but there should be a strong relationship between the two panels. The second way to estimate the standard error is to use generalizability theory to estimate the variance components associated for the different factors in the method. Generalizability theory allows the variance components

to be used as an estimate of the standard error of the cut score (Brennan & Lockwood, 1980).

Kane (2001) points out one more method that can be used to check for internal consistency for a test centered method like the Angoff. Panelists in the Angoff procedure are required to estimate the proportion of minimally competent examinees that will get each item correct. Once examinees have taken the test, the panelists' ratings for each item can be compared to the examinees' scores. When only candidates close to the cut score are used in the computation of p-values, the item difficulty for these minimally competent examinees should be similar to the SME ratings for each item. If the conditional p-values are similar to the SME ratings, then this is evidence that the panelists' item difficulty estimates were accurate.

Shepard (1993) suggested comparing cut scores between different types of items (multiple choice and constructed response) as well as comparing cut scores across different areas of content or benchmarks on the test. If content or item formats are judged differently by panelists, then these additional checks may help reveal potential problems in the methodology or training of SMEs (Cizek, 1993).

Kane (2001) emphasized the need for a method to produce reliable results as an essential component to a standard setting methodology. While Brennan and Lockwood (1980) suggested the use of generalizability to estimate the reliability of an entire method, Kane suggested evaluating intra-rater reliability as well. One way he suggested to obtain this measure was to have the same raters do the rating task twice. A correlation coefficient can be computed for both rounds of rating as an estimate of intra-rater reliability. If raters are independent of each other, then a measure of intra-rater reliability can provide valuable

information about the reliability of the standard setting method and the ability of SMEs to understand the required task.

2.3.5.1 External Criteria

The third body of evidence that should be compiled to evaluate the validity of cut scores is external evidence. External evidence can be obtained by comparing cut scores established during standard setting to an external measure. While many sources of data may be used in the comparison, there is never a perfect external criterion (Kane, 2001). For example, a potential external criteria for a certification test may be job performance reviews, but this criterion is subject to error in the manager's opinion and reporting avenues.

The first way to capture external evidence is to compare the standard setting results of one standard setting method to the results of another (Werner, 1978). This process is similar to the ideas behind convergent and divergent validity. This comparison has the most value when there is confidence in both of the standard setting methods (Webb & Fellers, 1992). If the two approaches agree, then there is convergent validity and also more confidence in the resulting cut scores. However, it is common for the methods not to agree, as different methods may ask different questions and provide different data to the examinees.

The second and most straightforward method is to compare the results for the test to some other assessment-based procedure (Kane, 2001). In this method, examinees who have recently finished an exam and were categorized into performance categories then take a second exam or participate in an activity related to the first. High performance in the activity should be related to the classification decision on the initial exam. However, this form of evidence is usually not satisfactory and is often difficult to obtain (Shimberg, 1981).

First, it is necessary to develop a second form of assessment as a point of comparison. Second, the alternative assessment must also have a cut score established using some standard setting method, which provides ambiguity in the relationship between the two measures. Third, the time commitment of taking two different assessments is usually too impractical for operational testing. Because of these weaknesses, this form of evidence is rarely, if ever, obtained (Kane, 2001).

The final method suggested by Kane (2001) involves comparing the cut scores to some other form of assessment. Classification data, such as grades in a course, SAT scores, job performance, or other assessments could be directly compared to the established cut scores and test performance. A positive relationship between cut score decisions and theoretically related constructs shows support for the accuracy of the cut scores.

While the standard setting field continues to grow and new methods are introduced, several of the core issues remain the same. There is a continuous struggle with how to set appropriate cut scores because no perfect method has been discovered. Despite the inconsistencies across standard setting methods, it is important to validate the interpretations and use of cut scores through the collection of validity evidence for whichever method is chosen for the standard setting workshop.

2.4 Standard Setting Methods

In practice, there are many different standard setting methods. Zieky (2001) made a list of six standard setting methods used in practice: estimated distribution, bookmark, Angoff, cluster analysis, generalized examinee-centered, and web based. However, these methods are just a few of the many different established standard setting methods. Berk (1986) identified over 37 different standard setting methods for criterion-reference tests, and this number has only grown (Raymond, 2001). The Angoff method has risen steadily in

popularity since its introduction in 1971 (Impara & Plake, 1998). The bookmark method was proposed by Lewis, Mitzel, and Green (1996), and has also become popular on many tests. Both of these methods will be discussed in greater detail because of their relevance to the current study.

NAEP has provided interstate trend data and has been supplemented by state assessment programs for within-state performance and trend analysis. The testing and accountability policies associated with No Child Left Behind (NCLB, 2001), required states to demonstrate that students were performing proficiently in key subjects by the 2013-2014 academic year. This also required regular assessment of students' performance through assessments in reading and mathematics in third through eighth grades and at least once in high school. This represented a major shift in most states' accountability policies and a significant investment of resources into assessment programs; not only was the actual movement of students from below to above proficiency a significant requirement of the law, the testing programs (and associated data systems) presented a major challenge for many states.

For low performing schools demonstrating adequate levels of proficiency and meeting annual growth objectives as required by NCLB was a significant challenge. Despite safe harbor policies, many schools struggled to show that enough of their students were participating in (and succeeding on) the required assessments. As schools began to implement the NCLB-required testing programs and accountability structure, it became clear that the testing and progress requirements differentially impacted both low performing and highly diverse schools (Kim & Sunderman, 2005). Though the full proficiency requirement has been adjusted to be more flexible, with many states applying for and being granted waivers, the notion of understanding and assessing students' current level of performance has remained integral to school accountability.

State accountability systems initially relied on status models, or snapshots of current performance, to judge whether students were making enough progress in a given year. Many states relied on comparing cohorts of students to one another (the fourth graders in 2002 compared to the fourth graders in 2004, for example) to judge whether students were improving across time. This requires a few potentially difficult assumptions, first that the cohorts are demographically similar. Assuming that comparing student cohorts can isolate student growth requires a belief that the cohorts are demographically comparable, have similar previous educational experiences, and have been exposed to similar [enough] educational programs. This is not always a feasible approach. It is particularly problematic when student populations are known to not be comparable based on a curricular or programmatic shift, like school re-structuring, or when there is a significant amount of student and/or teacher turnover within a school.

The proportion of students performing at or above proficiency may be very important, for example, when comparing schools within a district. Having a higher percent proficient could indicate that one school is outperforming another, even when their student populations, curricula, and basic methods are comparable. School accountability based on a status approach exacerbates several measurement issues, like comparing successive cohorts of students. The status approach also masks the performance of persistently low performing schools (Ho, 2008). By ignoring growth or progress below the proficiency cut point, schools that may be facilitating tremendous growth in their students *without* the students crossing the proficiency cut-score are not recognized for their success at increasing student achievement.

Critics of the status approach argue that test performance does not adequately represent academic progress and that the limitations of status measures fail to reflect the performance of students and schools. At the school level, Betebenner (2009) argues that dichotomous classifications of student performance (as proficient or not proficient) are inadequate for judging

a school's efficacy. Status models also introduce several measurement issues pertaining to how proficiency, or movement toward proficiency, is understood. Technically, student *progress* cannot be adequately assessed with a descriptive 'snap shot' approach given the dependence of proficiency measures on the location of the cut-scores, comparability issues across cohorts, and potentially problematic re-allocation of school resources to students performing just below proficiency (Booher-Jennings, 2005; Holland, 2002).

As this debate played out in testing and accountability policy, increasing attention was paid to the different factors influencing student performance. This led to comparative and exploratory study of teacher characteristics and qualifications as well as individual student factors that may lead to increased success in the classroom. The status approach was determined to be inadequate for assessing the effectiveness of a given school or teacher (see Linn, 2003; Linn, Baker, & Betebenner 2002), given the increasing political importance of both individual teachers and schools being held responsible for student success or failure. In response to the limitations of status modeling, particularly the masking of student progress below and above the proficiency cut, an alternative approach to demonstrating school efficacy was introduced through the 2005 Growth Model Pilot Program (GMPP, Spellings, 2005).

Growth modeling allowed schools to be accountable for the progress students were making toward proficiency instead of absolute proficiency (counts or percentages of the student body). This made demonstrating efficacy much simpler for historically low performing schools as well as those serving a diverse student body as their students were improving but were still operating below the proficiency cut point (Kim & Sunderman, 2005). The GMPP introduced four main types of models to contextualize student test score changes and estimate a student's growth. Through participation in the GMPP, several states used student test data to demonstrate accountability based on one of four approaches, a trajectory model, value table / transition matrix, value added modeling, or the student growth percentile. Each of these models operates

differently, but all take into account students' past and current test score(s) in estimating a student's growth based on his or her score trajectory.

2.4.1 Angoff Method

The most common and well known standard setting method carries the name of its inventor: The Angoff Method. Interestingly, the first mention that Angoff made of his procedure was in a chapter on scaling and equating which was written as a measurement reference (Thorndike, 1971). In the 100-page chapter, Angoff described the entirety of his method in a single 21 line paragraph. While the method carries Angoff's name, Angoff himself credited his colleague Ledyard Tucker, his colleague at the Educational Testing Service (Plake & Cizek, 2012).

> A systematic procedure for deciding on the minimal raw scores for passing and honors might be developed as follows: keeping the hypothetical "minimally acceptable person" in mind, one could go through the test item by item and decide whether such a person could answer correctly each item under consideration. If a score of one is given for each item answered correctly by the hypothetical person and a score of zero is given for each item answered incorrectly by that person, the sum of the item scores will equal the raw score earned by the minimally acceptable person. A similar could be followed for the hypothetical "lowest honors person". (1971 p. 514-515)

Plake and Cizek (2012) pointed out three critical components of Angoff's brief proposal. The first is that SMEs should cognitively conceptualize the "minimally acceptable person." This mental visualization of the minimally competent examinee remains a core component of the Angoff method today. The second important aspect is raters make judgments about each test item. Jaeger (1989) referred to methods which focus on rater

judgments about item parameters as a test-centered model. The third important aspect of Angoffs' original method is it can be applied and adapted to set more than one cut score. By simply performing the exact same exercise but conceptually imagining a different minimally competent group, a cut score for a different proficiency group could be established.

Angoff made one additional footnote in his initial introduction of the Angoff method. He stated:

> A slight variation of this procedure is to ask each judge to state the probability that the "minimally acceptable person" would answer each item correctly. In effect, judges would think of a number of minimally acceptable persons, instead of one such person, and would estimate the proportion of minimally acceptable persons would answer each item correctly. The sum of these probabilities would then represent the minimally acceptable scores (1971, p. 515).

This footnote introduced the first Angoff where raters would effectively attempt to provide probability judgments for borderline examinees.

The Angoff method procedure has changed little since its introduction and remains relatively simple. First, a panel of raters comprised of SMEs and other exam stakeholders is assembled. Each rater then conceptualized the probability is that each minimally competent examinee would get each item correct. The sum of the probabilities for each item equals the passing score for one rater. The average across all raters is the proposed cut score for the exam.

There are several modified Angoffs in practice today. One modification is including multiple rounds of ratings, where, between each round, panelists discuss their ratings as a group. Another modification is that impact data, or information about the test and

examinees, is given to the panelists between each round. However, in every modification, the core of the Angoff method remains constant.

The Angoff method is one of the most popular standard setting methods (Cizek, 2012). While popular, it has received much criticism. Impara and Plake (1998) expressed concerns about the capability of panelists to make accurate judgments about items and examinee performance. The authors asked teachers to rate the performance of their students on a classroom assessment that they had used many times over several years. The study findings indicated that individual panelists could not make accurate item difficulty estimates for their own students. Additionally, rater performance degraded when asked to estimate item difficulty for specific population subgroups such as the minimally competent examinee. The authors argued that it would be unlikely that a typical panel of raters could accurately estimate item difficulty by rater performance if teachers could not accurately perform the task for their own students, whom they had been working with for an entire academic year, on a familiar test they had used for many years. Raters become even less accurate in their estimates when additional factors are introduced, such as: setting multiple cut scores for different performance levels, presenting impact data on the test or examinees, facilitating discussion between raters, or accounting for the possible effect of guessing (Melican & Plake, 1985).

Shepard (1995) expressed similar concerns about the Angoff method, arguing that the cognitive task requires raters to 1) imagine the typical test taker, 2) condition the typical test taker on the *minimally competent* test taker, and 3) understand probability sufficiently to estimate the probability that the randomly selected, minimally competent examinee would get the item correct. This list of complexities creates a task that is too cognitively advanced for panelists and that exceeds their abilities as human beings. Thus,

ratings from an Angoff standard setting workshop would be inaccurate as panelist could not accurately complete the task.

While the Angoff method has been criticized in the literature, may prominent papers have been written defending the Angoff method. Kane (1995) defended the Angoff method and pointed out that it has been used on a multitude of certification and educational tests without major complaints from participants. Zeiky (2001) also pointed out that if the Angoff was indeed impossible for panelists to understand then there would be far more complaints from panelists.

2.4.2 Bookmark Method

A second standard setting method, the bookmark method, also deserves attention in this review because of its impact on standard setting and the reasons for its recent rise in popularity. The bookmark method (Lewis & Mitzel, 1995) is an item response theory (IRT) based standard setting method based on the concept of item mapping (Bourque, 2009). Bourque refers to *item mapping* as the attribution of the skills, knowledge, abilities, and other characteristics by test items to examinees with scores near the scaled difficulty of those items. For example, an item with an IRT difficulty of 1.5 may have skills associated required skills: graphical interpretation, problem solving, and table development. An examinee that gets the item correct and who has a total score near the scaled score of the item is attributed with the skills associated with that item.

The bookmark method, like most standard setting methods, is relatively straightforward. Lewis and Mitzel (1996) required each item to be calibrated and placed on the IRT theta scale with no guessing parameter. The items are ordered based on the probability of a student having a set probability of getting the item correct. The items are placed in an ordered item booklet (OIB) in this order. To determine the cut score, panelists

review each item in order and, keeping in mind the *minimally* qualified candidate, rate each item as to whether the candidate will have a greater, equal, or less than a given probability of getting the item correct. The cut score is then the average of all the item difficulty parameters for those items ranked equal to the given probability.

In practice the bookmark method can be much different than what was initially proposed by Lewis and Mitzel. Although the OIB is compiled in a similar way, panelists simply go through the book and literally place a bookmark between the item they believe the minimally competent candidate will answer correctly and the item the minimally competent candidate will answer incorrectly. An assumption with this method is that raters can conceptualize the item booklet as a step scale, where examinees will get all the items up to a certain difficulty correct and items thereafter, incorrect.

The bookmark method shares several characteristics in common with the Angoff method. The most notable similarity is that the panelist mentally conceptualizes the minimally competent examinee when rating items. However, a notable departure from the Angoff is that it does not require raters to make complex probability estimates for each item (Lewis, Mitzel, Mercado & Schultz, 2012).

Lewis et al. (2012) described several reasons for the rapid rise in popularity of the bookmark method. The first was the use of multiple performance levels following the 2002 NAEP (Bourque, 2009) and the requirement of at least three performance categories for the NCLB placed a heavy strain on the Angoff method, as it was primarily designed for a single dominant cut score (pass/fail). The difficulty of having panelists make a probability judgment for each item on the test, for each performance level, resulted in increased standard setting times for the Angoff method, which resulted in panelist fatigue and jeopardized the validity of the cut scores. In addition to increased time, the cost of

performing an Angoff workshop escalated. The authors suggested the BSSP was being adopted because it was better equipped to handle the writing of PLDs, as it is a natural outcome of the process. It also is better able to handle the use of constructed response items better than methods such as the Angoff, which are primarily tailored to single response items.

Lewis et al. attribute the bookmark method's rise in popularity to the dissatisfaction with the Angoff method. The Angoff method, they argue, requires panelists to make probability judgments, a task that is not well suited to panelists, such as teachers and educators. Finally Lewis et al. (1996) mentioned that the Angoff was widely criticized as being "fundamentally flawed" (Shepard, Glaser, Linn & Bohrnstedt, 1993, p. 132) and people were looking for alternative methods. The BSSP provided a sufficient solution.

For the purpose of the present study, the BSSP provides valuable information about future standard setting procedures. The BSSP attempts to integrate directly with the IRT scale values (Lewis & Mitzel, 1995) which provides a valuable statistical tool in the standard setting procedure, that of an equal interval scale.

2.5 Legal Issues in Standard Setting

An important consideration of any standard setting procedure is its defensibility in court (Kane, 1994). Carson (2001) outlined case law regarding the importance of standard setting. Carson noted the number of times that standards have been challenged, both in educational and certification testing. The necessity of setting standards is necessary has been upheld by the court, dating back to *Schware v. Board of Bar examiners of State of New Mexico* (1957), where the courts stated: "A state cannot exclude a person from the practice of law or any other occupation… A state can however require high standards for

qualification… but any qualification must have a rational connection with the applicant's fitness or capacity to practice a licensed occupation" (pp. 238-239).

It would initially appear that the courts would require some form of external validity evidence to support the standards. However, in practice, the most important form of evidence has been procedural validity (Plake, 1998). Given the difficulties of finding relevant external criteria for a point of comparison, the most valuable information is the evidence supporting the process used for defending standards (Kane, 1994). The standard setting process is "a psychometric due process" (Cizek, 1993) that is a rationally defined set of rules that govern the judgmental process. Because of the importance of the documentation of the standard setting process, it is necessary that any standard setting method contains a well-developed set of rules that oversee the process that can be well documented. Which procedures are used does not appear to be as important as the documentation and reasonableness of the procedure.

2.6 Conclusions Based on the Review of Literature

The literature review revealed that standard setting is a broad and versatile topic. Standard setting is frequently criticized for several reasons, one of which is the unreliability across methods. Each individual method comes with specific problems and criticisms that range from the complexity of the cognitive task to insufficient statistical justification. The importance of standard setting begins with the selection of panelists and ends with the collection of appropriate validity evidence to support the use of intended cut scores. Kane (2001) highlighted three important facets of validity information that should be collected for every standard setting method: procedural validity, internal validity and external validity. Each of these sources of validity provides evidence that cut scores are as defensible and accurate as possible.

IIT has been shown to be applicable in a wide array of situations. At the core of IIT is the idea that the mental process of making judgments can be inferred through the use of a factorial design and the detection of a cognitive algebra model. While IIT has never been applied to standard setting, the processes seems well situated to the standard setting field. The most common form of IIT analysis is the visual detection of a cognitive algebra model through the use of a factorial graph. If this inspection reveals a linear fan or parallelism, then the underlying cognitive scale utilized by the raters has desirable properties and IIT may help inform a standard setting method.

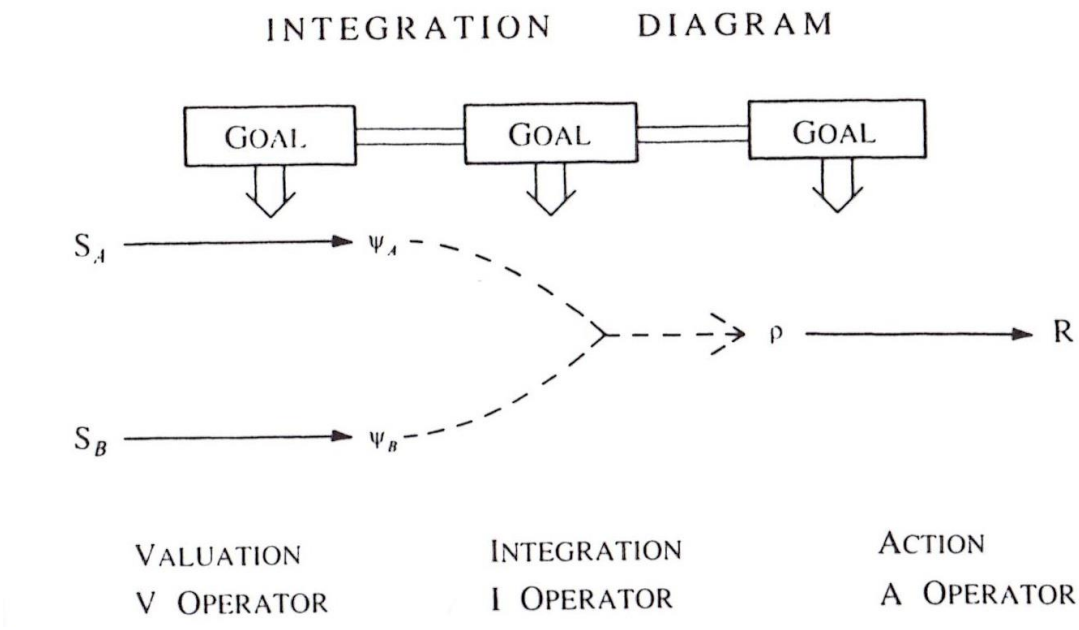


Figure 1 IIT design

	Factor A			
Factor B	$\psi_{A1}+\psi_{B1}$	$\psi_{A2}+\psi_{B1}$	...	$\psi_{An}+\psi_{B1}$
	$\psi_{A1}+\psi_{B2}$	$\psi_{A2}+\psi_{B2}$	...	$\psi_{An}+\psi_{B2}$
	$\psi_{A1}+\psi_{B3}$	$\psi_{A2}+\psi_{B3}$	...	$\psi_{An}+\psi_{B3}$
	$\psi_{A1}+\psi_{B4}$	$\psi_{A2}+\psi_{B4}$	...	$\psi_{An}+\psi_{B4}$

Figure 2 Example Factorial Design Using Additive Cognitive Algebra Model

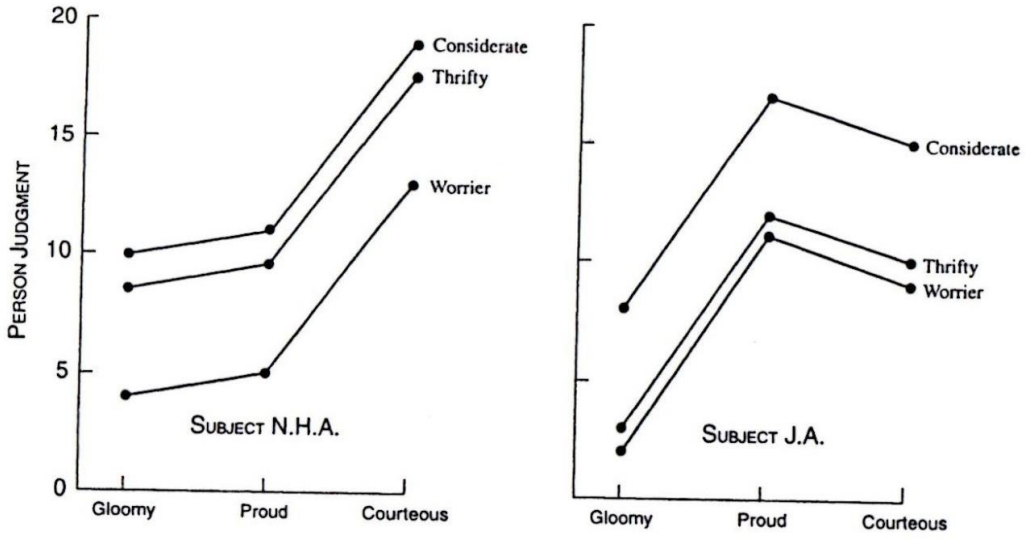


Figure 3 Observed Parallelism Example.

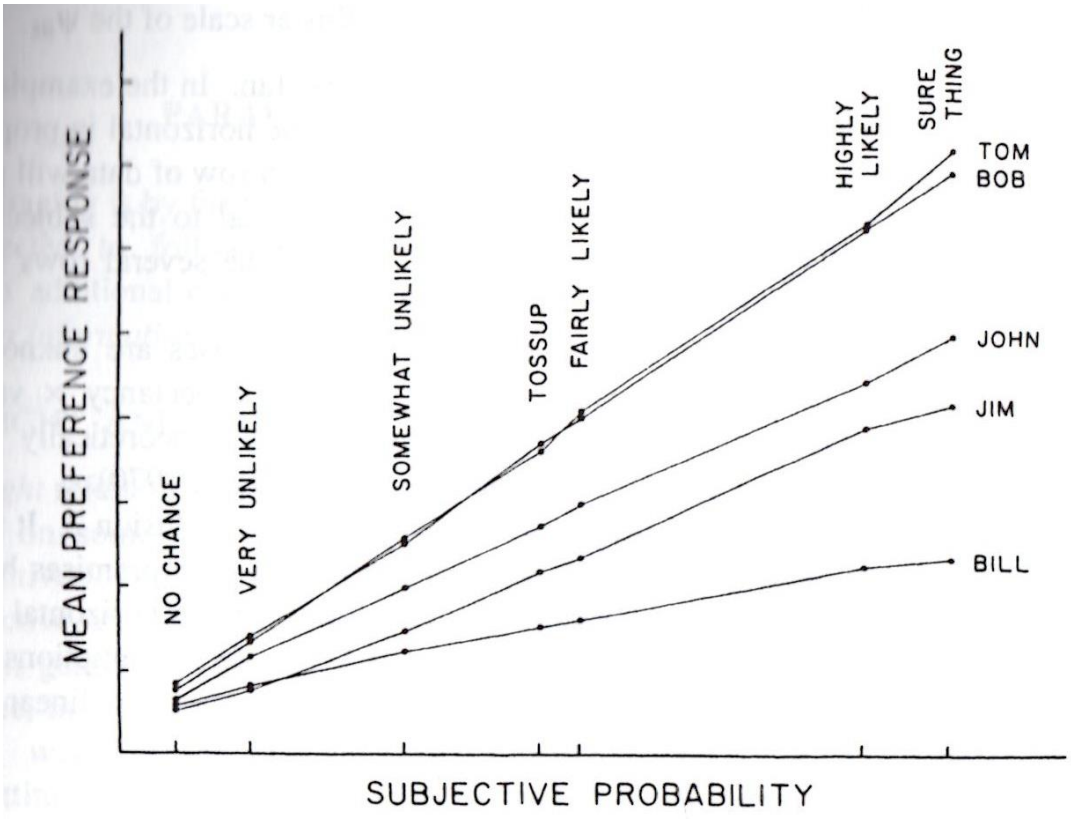


Figure 4 Linear Fan Example

CHAPTER 3

METHODOLOGY

3.1 Overview

The main purpose of this study is to evaluate if information integration theory (IIT) can be effectively applied to standard setting. Additionally, this study will offer a brief comparison between the IIT standard setting method and the Angoff method. More specifically, the following three research questions will be addressed:

(1) Can IIT be useful in conducting a standard setting meeting?

(2) Do expert judgments follow a known cognitive algebra model?

(3) How does an IIT based standard setting method compare to the commonly-used Angoff standard setting method?

The first question addresses the overarching issue of the appropriateness of IIT to standard setting. The appropriateness of IIT will be evaluated using Kane's (2001) validity framework for evaluating the standard setting process. The second question investigates specific questions common in an IIT study, mainly the positive identification of a cognitive algebra model. This question will be answered through an analysis of the factorial graphs . If a cognitive algebra model can be identified, then the third question will compare the appropriateness of cut scores set by the IIT and Angoff methods by following Kane's (2001) framework for evaluating the validity of a cut score through the collection of procedural, internal and external validity evidence. The general procedural outline of the study follows:

1. Develop a method and program which allows for SMEs to participate in a standard setting method governed by IIT.

2. Perform standard setting operations on three exams from widely varying areas using both the Angoff method and IIT method.

3. Identify and analyze sources of internal validity evidence for both methods.

4. Identify and analyze sources of external validity evidence for both methods.

3.2 IIT Standard Setting Procedure

As mentioned, the principle point of analysis for IIT is the factorial graph, which requires a factorial experimental design. The factorial design in turn requires a minimum of two factors, or variables, to be used. Two factors commonly used in test-centered standard setting methods are perceived item difficulty and performance levels. An example of this factorial design is given in Figure 2. Similar to the Angoff method, SMEs participating in the IIT standard setting method will be asked to rate the difficulty of an item for a PLD. Each rater will be presented with an item and a PLD and asked to rate the difficulty of the item for a typical candidate for the particular PLD. This process will continue until each SME has completed every combination of PLD and item in the factorial design.

After each rater has completed the task, both the individual factorial graphs for raters and the aggregated factorial graph for all raters will be evaluated to determine the specific cognitive algebra pattern. The factorial graphs will be investigated for either observed parallelism or a linear fan, as evidence of an additive or multiplicative model, respectively. A model will only be identified through the inspection of factorial graphs and accompanying ANOVA tests. If an adding or multiplicative cognitive algebra model can be confirmed, then the use of IIT for standard setting has valuable evidence. An additive model will be confirmed by first identifying observed parallelism in the factorial graph followed by the absence of a significant interaction in the repeated measures ANOVA. If there is a

significant interaction, Eta-squared will be calculated as a measure of effect size. If the effect size is small ($\eta^2 < .058$; Cohen, 1988) then this will also be evidence of an adding-type model. A multiplicative model will be identified by evidence in the factorial graph of a linear fan and a significant interaction with a large effect size ($\eta^2 \geq .058$) in the repeated measures ANOVA. If either model is identified, the benefits described by Anderson (1981, 1982), such as the ability to use an equal interval scale, will then be applied to the rating scale and help inform the placement of cut scores.

3.2.1 Estimating the Cut Score

After evaluating if a cognitive algebra model is appropriate, the next step will be to determine the best way to set a cut score using the raters' judgments and the benefits of IIT. As with any standard setting method, IIT will be used to divide continuous examinee identifiable buckets (Pass/Fail, Qualified/Not Qualified).

The Angoff method provides valuable theoretical information about where to place a cut score. The task behind the Angoff method is for raters to conceptualize the "competent" examinee and then condition that conceptualization on the *minimally* competent. The average across this rating measure eventually becomes the suggested cut score. *Minimally competent* is used in the Angoff method because the cut score should be placed on the continuous scale just as the point of transition between the most proficient examinee in one category and the least minimally competent examinee in the next. Figure 5 shows two performance categories (basic and competent) separated by a single cut score. If the location of two performance categories is known, the cut point should be placed somewhere on the scaled score between the two performance categories. The location of the cut score on the scaled score should be right as the most proficient examinee in the lower category becomes the least proficient examinee in the higher category.

The IIT method of standard setting does not delineate within performance levels using the concept of *minimally competent*. Instead IIT sets cut points by obtaining the midpoint of several different performance levels simultaneously using a matrix based on the factorial design. Since cognitive algebra provides information about an equal interval scale, each point between performance midpoints is equal distance. Therefore, the point directly between two performance level midpoints is the location where one performance level transitions to the next. The significance of this is that the midpoint between two performance levels is where the new cut score should theoretically be located. To derive a numerical cut point, the marginal means of the rows for each performance level in the factorial matrix will be calculated and the midpoint between two performance levels will be the cut score. The cut however will be placed initially on the rating scale, but since the scale is equal interval it can be transformed into either a percent scale, the raw score scale of a test or even an IRT theta scale. This process is illustrated briefly in Figure 6, which shows how a linear transformation would convert the cut scores from a 0-20 scale into a raw score on a 65 item test.

3.3 Program Development

Since IIT has never been applied in standard setting, there is no software program that can be adequately used by SMEs. Therefore, it will be necessary to develop a program that allows the application of IIT to standard setting and adheres to the specific methodological characteristics described by Anderson (1981, 2004, & 2008). The program will facilitate the following tasks:

1) Present SMEs with each item by PLD combination.
2) Randomize the presentation order of each combination.

3) Present the SMEs with practice ratings. The user interface for this process is shown in Figure 7. Each rater will be asked to rate the difficulty of a random item for a random proficiency level on a fixed scale.
4) Create a factorial graph for each SME.
5) Create a factorial graph for the aggregated data across all SMEs.
6) Run a repeated measures ANOVA, including F-tests for both main effects and the interaction.
7) Compute the suggested cut scores based on the aggregated SME data.

One important consideration in program development is the presentation of the stimuli and the user interface. In general, the interface will be constructed to make it as user-friendly as possible with few possibilities to make errors. Users will not be permitted to return to previous ratings and must continue to the next stimuli must present a rating for the current one. Currently, the rating scale can toggle between 1-1000, 1-100, and 1-20. The scale itself is arbitrary and Anderson (2008) suggested using a scale unfamiliar to the rater. Since a functioning IIT study hinges on the importance of a linear (equal interval) scale, the numerical scale values themselves are relatively unimportant and Anderson has even suggested using a slider scale to remove the confusion associated with a numerical scale. Anderson specifically cautions against the use of a 1-100 scale because it adds increased difficulty to the cognitive task by adding typically unused points as users generally treat a 100 point scale as a 20 point scale, only using multiples of five. Additionally, the 1-100 scale may interact with scales familiar to the raters such as a percent scale (Anderson, 1981). Finally, Anderson points out raters usually utilize a 1-100 rating scale similar to a 1-20 rating scale, frequently just selecting multiples of 5 even when given the freedom of other numbers. The goal of the program development process was to incorporate Anderson's

suggestions on conducting an IIT study into a user-friendly program that can automate much of the standard setting process.

3.3.1 Reducing Threats to Validity

While many of the tasks required of the program are standard practice for a within-subjects factorial design. However, steps 2 and 3 are suggestions given by Anderson (1981) to help reduce threats to the validity of an IIT study. He suggests that three of the largest threats to the validity of an IIT experiment are position effects, carryover effects, and memory effects. *Position effects* occur when the rating of a particular stimulus depends on its serial position. The earliest stimuli may be more inaccurate than later stimuli because of learning effects and the need to internalize the response scale through practice. Later stimuli may suffer as well since SMEs may become fatigued. Stimuli order are randomized by the program to control for fatigue, and ten practice items are given to control the initial learning process.

Carryover effects occur when one response is dependent on a previous response. For example, if each item by performance level stimuli were given in order, a SME would see the same item three times in a row and would know that the item should be easier for more advanced groups. The proximity of each of these stimuli would result in carryover effects. To help reduce this problem, stimuli are presented in a random order to SMEs. *Memory effects* are related to carryover effects and create dependencies among stimuli when the rater remembers and utilizes previously viewed information. While difficult to control, randomizing the presentation order of stimuli helps create a more balanced design that can help control for memory effects.

3.4 Design

The first task after data collection will be to estimate cut scores on exams using both the Angoff and IIT methods. Cuts cores will be set on three different exams in three different content domains. These exams are: HP's Designing HP storage solutions exam, Excelsior College cultural diversity exam and the Trends for International Math and Science (TIMSS) exam. Each test will have cut scores set by both the Angoff and IIT methods. Both methods will be as faithful as possible in adhering to the nine standard setting steps proposed by Hambleton, Pitoniak & Copella (2012). Descriptions of each test, including information about panelists, standard setting operation and examinee descriptions are given below.

3.4.1.1 HPs Designing HP Enterprise Storage Solutions Exam

The HP storage solutions exam is comprised of 120 items. The item formats for these items range from multiple choice, multiple correct multiple choice, matching, pull down and hotspot items. Most items are scenario based and include images. The test is a high stakes exam that offers certification in the use of HP database software.

3.4.1.1.1 Panelists

The HP designing HP storage solutions exam will use ten SMEs for both the Angoff Method and the IIT method of standard setting. HP initially will provide twenty SMEs and they will be randomly assigned to either the Angoff method condition or the IIT condition. There will be no interaction between the two sets of panelists. The composition of each panel will include 50% content specialists and 50% educators in storage solutions. Panelists received compensation equally for their participation in both groups and consistent with what HP normally provides SMEs for a standard setting workshop.

3.4.1.1.2 Standard Setting Operation

The HP exam will set standards on the exam using both the modified Angoff method and the IIT method described above. Standard setting workshops will take place on consecutive days with the modified Angoff workshop first and the IIT workshop on the second day. The same facilitator will be used for the training and operation standard setting operation for both methods.

3.4.1.1.3 Examinees

The examinees for the HP exam are typically professional workers in the HP company structure wishing to get certified in the next level of HP software development. Examinee level data will be collected and examined after approximately 1500 examinees complete the storage solutions exam.

3.4.1.2 Excelsior College Nursing Exam

The nursing exam measures the skills and knowledge obtained in a standard broad spectrum nursing course. The test is 100 multiple choice items with a range of graphics and scenarios.

3.4.1.1.1 Panelists

Sixteen panelists will be chosen that all have at least two recent years of teaching experience as college professors in the field of cultural diversity or a related field. Panelists will be compensated for their time according to standard Excelsior college compensation requirements. The SMEs will be randomly assigned to the IIT standard setting process or the Angoff standard setting process.

3.4.1.1.2 Standard Setting Operation

The standard setting operation will take place over the course of three days. The first day will include training panelists in the Angoff method and the first round of Angoff ratings. The second day will include discussion of the Angoff ratings and subsequent rounds of evaluations. On the afternoon of the second day training will begin on the second group of panelists for the IIT method. On the third day, the SMEs will complete the IIT standard setting workshop.

3.4.1.1.3 Examinees

The examinees for the cultural diversity test are college students in the cultural diversity class taught by Excelsior College. Examinees typically range from 18 – 50 years old and represent a typical, if slightly older college classroom. After 200 examinees have taken the exam, examinee level data will be investigated and compared to the estimated difficulties from the standard setting workshops.

3.4.1.2 Trends for International Math and Science

The Trends for International Math and Science Study (TIMSS) is an international assessment designed to measure math and science achievement in the United States and throughout the world at the 4th and 8th grade levels. The TIMSS was administered in 1995, 1999, 2003, 2007, and 2011. For the purpose of this study, only the 2011 data for 8th grade math will be used. As an international assessment, the TIMSS was administered in more than 60 countries; however, more than 20,000 students in 1000 schools across the United States participated in the assessment. The current study focuses only on students from the United States, as the recruitment of panelists for the standard setting procedures will also be limited to the United States.

The TIMSS uses a matrix sampling design to administer questions to students. While many forms of the test are available, they are roughly equivalent, and each will include 30 items (with 15 shared items on another form). The current study will focus on only a single form of the 8th grade TIMSS math assessment for the standard setting process.

3.4.1.1.1 Panelists

The final set of panelists was selected for the TIMSS. However, no specific company was in charge of setting standards using IIT for the TIMSS exam, so thirty panelists will be recruited and offered compensation for their time. The composition of these panelists will be roughly 75% teachers and 25% school administrators or math curriculum specialists. As a requirement, teachers will be required to be currently employed as 8th grade math teachers or curriculum specialists. Panelists will be compensated a fixed hourly rate for their participation. Each participant will be offered $50 an hour for their services. Panelists will be randomly assigned to one of three standard setting groups. The first group will set standards on the thirty item test using the IIT method. The second group will set standards using the modified Angoff method with items and ability levels presented in a random order. The third group will be perform a traditional Angoff rating procedure with items presented in a fixed order within each performance level.

3.4.1.1.2 Standard Setting Operation

The standard setting workshop for the TIMSS exam will be done online for both the IIT method and Angoff method. Each of the panelists will be required to participate in a 1-2 hour training session. After the training session is complete they will be able to log onto the standard setting website and make IIT or Angoff judgments depending on their assignment.

Each participant will have a total of one week to complete the required ratings for the three performance levels.

3.4.1.1.3 Examinees

The examinees for the TIMSS portion of the exam are 20,000 8th grade math students from over 1000 schools across the United States. An additional 15,000 8th grade students will be randomly selected from Asian, European and African countries.

3.4.2 Training of Panelists

Training is an essential part of the standard setting procedure. The quality of training directly contributes to procedural validity evidence. Therefore, one important focus of the study will be to give panelists adequate training in each method. Training will be done by following the procedures outlined by Loomis (2012), as well as suggestions by Hambleton, Pitoniak, and Copella (2012). Each company will provide facilitators to train the panelists for both the Angoff and IIT methods. Care will be taken to ensure that the training for both methods is as equivalent as possible given the differences in methodology.

3.4.3 Perform Standard Setting Operational Tasks

After training panelists, both the HP certification exam and the Excelsior college cultural diversity test will have cut scores set using both the Angoff Method and the IIT method. The Angoff method will follow each step proposed by Hambleton, Pitoniak, and Copella (2012). For the Angoff method, each panelist will begin by individually reviewing each item and providing the probability that a random minimally competent examinee will get the item correct. Next, the panel will convene, and individual differences in item ratings will be discussed within the panel for each item. Panelists will then rate each item

individually once again. After this second rating process, the ratings will be compiled and cut scores will be derived according to modified Angoff rules as described in section 2.4.

After training for the IIT method, each panelist will log into the IIT standard setting program via the internet. Each rater will see all the items for the three competency levels in a complete factorial design (3 x n, where n is the number of items in the exam. After all the panelists have completed their ratings, the program will compute the IIT cut scores according to the methodology described above in section 3.2. In addition, each rater will rate 10 items twice to calculate an intra-rater reliability coefficient. This intra-rater reliability coefficient will then be adjusted by the Spearman-Brown prophecy formula.

3.4.4 Collection of Additional Evidence

A large amount of validity evidence can be obtained strictly by recording the proceedings of the standard setting workshops. The main type of validity information obtained this way is procedural. Statistical information can be obtained by analyzing the rater responses. However, statistical evidence is not the only important information to support the use of a new standard setting operation. Testing programs may be interested practical information, such as the length of time it takes to complete a standard setting workshop in order to calculate potential costs. For the Angoff Method, the standard setting operation will be timed, including training and the time it took for the administrator to prepare materials. For the IIT method, time will be recorded for the preparation of materials and the time it took each rater to finish the rating procedure. In addition, the time it takes to analyze standard setting results will be computed for each method.

3.5 Identify Sources of Validity Evidence

Kane (2001) proposed three sources of validity information that should be compiled to help validate the interpretation of a given cut score. These sources were: procedural validity, internal validity, and external validity. This section focuses on the collection of validity evidence to support the setting of cut stores established for both the Angoff and IIT methods. Procedural evidence will support that proper and accepted steps were followed in the standard setting workshop by recording the proceedings of both standard setting workshops. Two main statistical indices of internal validity will be calculated and reported when applicable, for each method: inter-rater consistency using intra-class correlations and intra-rater consistency.. TIMSS data will be used to determine external evidence by comparing cut scores obtained from both Angoff and IIT methods to external criteria based on parent, teacher and student surveys.

3.5.1.1 Procedural Validity Evidence

The first form of validity that will be collected is procedural validity. Information will be recorded about the proceedings of the standard setting workshop. Information such as the selection of panelists, panelist training, panelist discussion, facilitator involvement in discussion and other information suggested by Kane (2001) will be recorded. The purpose is to collect information that the established standard setting rules for each method were properly followed. In addition, raters will be asked to complete a survey on the perceived effectiveness of the standard setting workshop and their confidence in the recommended cut scores. The survey will be similar to the survey found in Hambleton, Pitoniak & Copella (2012), with modifications made when appropriate for each standard setting workshop. The general survey is provided in Appendix A.

3.5.1.2 Internal Validity Evidence

One obtainable foundation of validity evidence for most standard setting procedures is internal validity information. The first source of internal validity is ensuring that panelists are reliable among themselves. While a portion of within-rater reliability can be inferred from the factorial graph and observed parallelism or non-overlapping performance levels, the strongest support for this form of evidence is obtained by having raters perform the standard setting operation twice. In many cases, this variation of test-retest reliability is unfeasible due to financial and timing constraints. However, in the current study a small group of items from each test will be rated multiple times by each panelist. This subtest will be selected based on item specifications and test objectives that match the total content of the test. While the entire exam will not be rated twice by panelists, the small subset of items should provide data to evaluate for intra-rater consistency. Since only a small portion of items will be used to compute intra-rater reliability, the Spearman-Brown prophecy formula will be used to predict the intra-rater reliability for the entire test.

The second method for obtaining internal validity evidence for each standard setting method is inter-rater reliability. Intra-class correlation (ICC) coefficients using a one-way random effects model will be calculated for each standard setting workshop.

Other descriptive information about the cut score will be obtained, including the standard deviation of the cut score in order to evaluate the error of cut scores set by both methods. Additionally, the standard deviation of the mean will be calculated for each standard setting workshop. While most internal validity evidence will be collected for both methods, an additional form of validity is only applicable to the IIT method. This validity is the detection of identifiable cognitive algebra models. Detection of models will be done through the inspection of the factorial graph provided by panelists' ratings. Both the

individual graphs and the graph of the aggregated rater data will be examined. If no basic cognitive algebra model is discernible, more effort will be placed into identifying more complex cognitive algebra models. However, if a cognitive algebra model can be identified from the factorial graphs, then this is strong internal validity evidence that IIT may be appropriate to standard setting.

If a cognitive model is visually identified, then a repeated measures ANOVA will be conducted on the factorial design to establish further support of the algebraic model. Both main effect F-tests will be analyzed in addition to the interaction. The main effect for performance level will show if cognitively the raters believe there are significant differences between the performance levels. However, the most compelling significance test is for the interaction effect. If there is observed parallelism, there should not be a significant interaction. If there is a linear fan, there should be a significant interaction. However, the effect size will also be computed for each of the main effects and the interaction. If there is a significant interaction, but it has a small effect size, then this is also support for a parallel pattern.

3.5.1.3 External Validity Evidence

The final source of validity information that will external validity. *External validity* is the comparison of the cut scores proposed by the standard setting panel to external criteria. Kane (2001) mentioned that this type of validity is difficult to obtain for standard setting because it is difficult to determine the quality of the external criteria. However, in the current study, we will attempt to compare cut score decisions to external evidence of student performance by correlating the cut score classification with student, teacher and parent evaluations as well as other variables associated with high performance. In addition

to these external criteria, cut score classifications of examinee data will be compared across the Angoff and IIT methods.

3.5.1.3.1 TIMSS External Validity Evidence

The TIMSS assessment is administered with surveys for the student, teacher, and parent, as well as demographic information on each student. The demographic and survey data will be used for two different analyses of external validity information.

The first analysis will correlate several variables theoretically related to higher performing students with cut scores set by the Angoff and IIT methods. These variables will be: number of hours in math class, teacher's perception of student's achievement level, parent's perceptions of student achievement level, the student's perception of their own achievement level, SES status, and mother's level of education. A correlation between these variables individually will help provide evidence of external validity.

The second analysis will use the same demographic and survey variables as the first, but with a more complex analysis. In the second analysis, these variables will be used as independent variables in a logistic regression function to predict student performance levels without using test scores. The TIMSS data set includes students from a broad spectrum of student performance. Ten thousand students will be randomly selected from each of the top, middle and bottom 10 percent of performers on the exam and used to compute an ordinal logistic regression equation. Examinee performance (top 10%, middle 10%, bottom 10%) will be used as an approximation of student performance levels and will be the outcome variable in the logistic regression. Next, SES status, mother's level of education, number of hours in math class, teacher's predicted performance of the examinee, parents' predicted performance of the examinee, and the student's beliefs about themselves will be used as predictor variables in the logistic regression.

The logistic regression equation will then be applied to a second random sample of 10,000 examinees from the TIMSS data. The logistic regression equation will assign each examinee to a predicted performance category (high, medium, and low). The predicted performance category will then be correlated with the placement categorization assigned by cut scores obtained from both the Angoff and IIT standard setting workshops.

3.5.1.3.2 Comparison of Examinee Data Across Methods

The final evidence of external validity will be the comparison between the Angoff and IIT methods for each of the three tests. The first comparison will compare the reliability and precision of the cut scores using internal validity evidence. This comparison will show which method provides more precise estimates of the cut score.

The second comparison will investigate the percentages of examinees in each performance level category for each method. Kane (2001) suggested that comparing the percentages of examinees in each category in different methods provides evidence of convergent and divergent validity. In general, it is not ideal for both methods to produce the exact same cut score unless one method is arriving at the cut score in a more efficient manner.

Finally the third evaluation of external validity will investigate the accuracy of rater judgments of item difficulty. The data for the examinees that barely passed the exam will be collected and used to compute conditional p-values. Since the Angoff method requires panelists to compute the p-value for the minimally competent examinee, then the rater derived p-values for the Angoff method should be similar to the empirical conditional p-values based on the candidates who barely passed the exam. A comparison of these values should yield roughly similar results if the raters performed the task accurately.

3.7 Conclusion of Methods Section

The methods section summarizes the design for the research project. The current plan is designed to follow Kane's (2001) framework for collecting validity evidence for standard setting methods. The collection of validity evidence will either help validate IIT as a potential standard setting method or show the theory's inadequacies in standard setting situations. The specific procedural, internal, and external validity evidence collected for both the Angoff method and the IIT method will help establish a comparison between the two methods. While the comparison between methods provides valuable information, the most important aspect will be the direct application of IIT to standard setting and the discovery of a cognitive algebra model. The discovery of such a model will help validate IIT as a potential standard setting method in the future.

CHAPTER 4

RESULTS

4.1 Overview

This study consisted of a total of seven different standard setting workshops for three different exams. Each exam had a minimum of two standard setting workshops, one using the IIT method and another using the Angoff method. The TIMSS exam had a third standard setting which was the randomized modified Angoff, or in other words, the Angoff question and scale with randomized performance levels and items. Results for each of these exams will be discussed in turn. Each study had a minimum of seven and a maximum of ten raters with each rater being randomly assigned to either the Angoff workshop or the IIT workshop. Where possible, the two different standard setting workshops were run in the same manner. Results for the standard setting workshops are divided into six sections: (1) detection of a cognitive algebra model, (2) estimating the cut score, (3) procedural validity evidence, (4) internal validity evidence (5) any additional analysis pertinent only to the current exam, and the evaluation of the external consistency for the TIMSS exam.

Results for the HP storage solutions exam are presented first, followed by the Excelsior college nursing exam. Findings based on the TIMSS standard setting workshop are reported last.

4.2 HP Standard Setting

4.2.1 Detection of Cognitive Algebra Models

The detection of cognitive algebra models was done through the inspection of the factorial graph found in Figure 9, which is an average across all raters. In addition to an inspection of the averaged factorial graph, each individual rater graph was inspected and

can be found in Appendix B. The second analysis performed to confirm a cognitive algebra model was a repeated measures ANOVA. The visual inspection of the factorial graph revealed nearly parallel lines for the performance levels, which is indicative of an adding or averaging model. The repeated measures ANOVA produced significant main effects for level ($F(2,12) = 93.51$, $p < .01$) and items ($F(97,582) = 6.35$, $p < .01$) and a significant interaction term ($F(194,1164) = 2.05$, $p < .01$). However, the interaction term was associated with an epsilon of .02, a very small effect size. Since the main effects were large, and the effect size for the interaction was small, these results support an additive model. The results of the ANOVA can be found in Table 1.

Table 1 ANOVA table for HP Storage Solutions Exam.

	Sum of Squares	df	Mean Square	F	*p*
Level	40602.7	2	20301.3	46.89	<.001
Item	3444.45	97	35.5098	6.35	<.001
Level x Item	1091.30	194	5.73	2.05	<.001

4.2.2 Estimating the Cut Score

Since an adding model was positively identified, estimates of the cut scores using the IIT data were calculated based on previously discussed methodology. The three different proposed methods of setting cut scores using IIT data produced different results. The first method, which took the difference between the marginal means of adjacent performance categories, produced a suggested cut score of 52.42% between unqualified and qualified and 73.32% between qualified and highly qualified. The second method set the cut score two standard deviations below the marginal mean and produced suggested cut scores of 53.73% and 68.64% for qualified and highly qualified. The third method, which

estimated the raters weighting factor from the valuation stage of IIT produced cut scores of 50.34% and 63.32% for the different performance categories. Only the cut score for the qualified examinee was calculated for the modified Angoff method. The estimated cut score for this method was 68.75%. The estimated cut scores are reported in Table 2.

Table 2 Value Estimated cut scores for HP Storage Solutions Exam

	Level 2 (Competent)	Level 3 (Highly Competent)
Angoff	68.75%	-
IIT Cut Score 1	52.42%	73.32%
IIT Cut Score 2	53.73%	68.64%
IIT Cut Score 3	50.34%	63.32%

4.2.3 Procedural Validity Evidence

Procedural validity insured the steps involved in the standard setting workshop were adequately followed by documenting the proceedings of both workshops. Overall, both standard setting workshops proceeded with few issues. However, one rater did not wish to participate in the study and opted out of data collection due to time constraints for the IIT method. Due to a programmatic error, a second rater's data were corrupted, leaving a pool of 7 raters on the IIT side and 10 raters for the Angoff.

Both sessions were timed, including training, discussion, practice and actual rating sessions. The training for each method took just over an hour as all participants had participated in previous standard setting workshops. After training, each participant performed 20 practice ratings and then continued with actual ratings.

The ten participants in the Angoff method took just under one hour and fifty-seven minutes on average to complete the ratings for one performance level. The participants then took two hours to discuss the Angoff ratings and did not have enough time to complete

the ratings for the other performance levels. Therefore, there are no data for highly competent or below competent for the Angoff method. The participants in the Angoff rating took approximately 6 hours to complete the entire standard setting workshop, not including breaks.

The 7 participants in the IIT study took three hours five minutes on average to complete all 324 ratings, or approximately 40 seconds per rating. Since there was no discussion among panelists after ratings, the entire IIT standard setting process took an average of four hours and thirty-five minutes for the participants.

4.2.4 Internal Validity Evidence

Internal validity evidence was collected by evaluating both inter-rater reliability and intra-rater reliability. Inter-rater reliability was assessed using two-way mixed ICC's. The ICC for the Angoff method was computed after the second round of Angoff ratings, after one round of discussion. The ICC for the Angoff method was .793 after the second round of ratings. The ICC for the IIT method was .782. Since each judge rated ten items twice for the IIT method, it was possible to compute an intra-rater reliability by correlating the first round of ratings with the second round for each items. The spearman brown prophecy formula was then used to predict the intra-rater reliability for the complete form of 100 items. The intra-rater reliability and the predicted intra-rater reliability for each of seven raters is shown in Table 3. The intra-rater reliability for each of the seven judges were all above .55 for 10 items with a predicted reliability of over .8 for a 100 item test. However, many of the predicted intra-rater reliabilities were above .99 for a 100 item test, indicative that raters were extremely reliable with themselves.

Table 3 Intra-rater reliability for 7 raters on HP Storage Solutions Exam

	10 item intra-rater reliability			100 item predicted reliability		
Rater	Not Competent	Ideal Competent	Highly Competent	Not Competent	Ideal Competent	Highly Competent
1	.673	.828	.82	.954	.98	.98
2	.365	.183	.852	.92	.817	.993
3	.906	.945	.82	.99	.994	.978
4	.698	.609	.643	.958	.94	.948
5	.822	.843	.77	.979	.981	.971
6	.408	.555	.971	.932	.961	1
7	1	.866	1	1	.987	1

4.3 Excelsior College Nursing Exam

4.3.1 Detection of Cognitive Algebra Models

The detection of cognitive algebra models was done through the inspection of the factorial graph found in Figure 10, each individual rater graph found in Appendix B, and statistically through a repeated measures ANOVA. The visual inspection of the factorial graph revealed nearly parallel lines for the performance levels, which is indicative of an adding or averaging model. The repeated measures ANOVA produced significant main effects for both levels ($F(3,33) = 771.15$, $p < .01$) and items ($F(99,1089) = 6.12$, $p < .01$) as well as a non-significant interaction term($F(297,3267) = 1.02$, $p = .41$). These results support an addition model. The results of the ANOVA can be found in Table 5. An effect size was not reported because the interaction was not significant.

Table 4 ANOVA table for Excelsior College Nursing Exam.

	Sum of Squares	df	Mean Square	F	P
Level	142666	3	47555.4	771.15	<.001
Item	1850.99	99	18.69	6.12	<.001
Level x Item	336.43	297	1.13	1.02	.41

4.3.2 Estimating the Cut Score

Since an adding model was positively identified in the nursing model, estimates of the cut scores using the IIT data were calculated. Similar to the HP standard setting data, the three different proposed methods of calculating cut scores using IIT data produced different results. The first method, which took the difference between the marginal means of two performance categories, produced a suggested cut scores of 44.54% between weak and marginally competent, 63.7% between marginally competent and competent, and 83.38% between competent and highly competent. However, Excelsior desired four different cuts so this methodology is not ideal as it can only produce cuts equal to the number of categories minus 1. The second method set the cut score two standard deviations below the marginal mean produced suggested cut scores of 29.6%, 48.24%, 68.29% and for weak, marginally competent, competent and highly competent, respectively. The third method, which estimated the raters weighting factor from the valuation stage of IIT produced cut scores of 25%, 42.76%, 62.14% and 81.88% for the same ability levels. The estimated cut score for the Angoff method was also calculated using traditional Angoff calculations and resulted in a suggested cut score of 33%, 59% 75% and 87% for weak, marginally competent, competent and highly competent, respectively. The estimated cut scores are reported in Table 5.

Table 5 Estimated cut scores for Excelsior College Nursing Exam

	Level 1 (Weak)	Level 2 (Marginally Competent)	Level 3 (Competent)	Level 4 (Highly Competent)
Angoff	33%	59%	75%	87%
IIT Cut Score 1	-	44.54%	63.7%	83.38%
IIT Cut Score 2	29.6%	48.24%	68.29%	88.34%
IIT Cut Score 3	25%	42.76%	62.14%	81.88%

4.3.3 Procedural Validity Evidence

Procedural evidence was collected through observation and rater surveys. The standard setting workshop took place over two days. The first day was devoted to training. In the morning on the first day, all 12 raters were assembled in a single room to receive an introduction to the test. The training began with each rater taking the exam so raters could get a feel for the difficulty of the test. After each rater finished the test, they were provided with results that summarized their performance. At this point, the raters were encouraged to discuss strategies items they got incorrect or they believed were incorrectly keyed. After discussing the test, all 12 raters received information on the basics of standard setting and the population of interest. The 12 raters then spent one hour developing PLD's for each of four performance categories.

After finishing the PLDs for each level, raters were randomly assigned to one of two groups. Each group then received information about the standard setting method they would use. Group 1 began with the modified Angoff method while group 2 began with the IIT method. After both groups finished with their respective method, group 1 then received training on the IIT method and group 2 received training on the modified Angoff method. Each group then proceeded to develop cut scores using the second method.

After the cut scores were developed using both methods, raters were asked to complete a survey detailing their experience during the standard setting workshop. Every rater reported they felt the training for both methods was adequate and they felt they adequately performed their job as a SME. All raters felt positive about both standard setting methods. Overall, 7 of 11 raters said they found the IIT method to be easier and more intuitive and 7 of 13 raters stated that if they were to return to do another standard setting workshop, they would prefer to use the Angoff method over the IIT method.

4.3.4 Internal Validity Evidence

Internal validity evidence was collected by evaluating inter-rater reliability. Inter-rater reliability was assessed using two-way mixed ICC's. The ICC for the Angoff method was computed for the second round of Angoff ratings, after discussion. The ICC for the Angoff method for the cut score were .813, .804, .832 and .848 for weak, marginally competent, competent and highly competent, respectively, while the ICCs for the IIT method were .735, .643, .711and .790 for the same performance levels.

4.3.5 Additional Analysis

The excelsior college nursing exam standard setting workshop was comprised of two independent panels of 7 raters. The first panel set cut scores on the 100 item test first using the modified Angoff method followed by the IIT method. The second panel began with the IIT method and finished with the modified Angoff. Due to the crossed design, it is possible to look at differences across panels to see if each panel produced similar results. The inter-rater reliability for the Angoff for panel 1 was .762, .731, .767 and .766 for the four levels. The ICC's for panel two for the same levels were .572, .616, .597 and .755. Panel one produced suggested cut scores of 37%, 66%, 78% and 89% for weak, marginally

competent, competent and highly competent, respectively. At the same time, panel two suggested cut scores of 30%, 53% 71% and 85% for the same levels.

The ICC for panel one on the IIT was .695, .582, .661 and .834 for highly competent, competent, marginally competent and weak. Panel two produced slightly lower results of .565, .535, .531 and .557 for the same performance levels. Two different suggested cut scores were computed for each panel using IIT. A summary of the differences in panels is reported in Table 6.

Table 6 Differences in cut scores between Panel 1 and Panel 2 on the Excelsior College Nursing Exam

	Level 1 (Weak)	Level 2 (Marginally Competent)	Level 3 (Competent)	Level 4 (Highly Competent)
Angoff	37%	66%	78%	89%
IIT Cut Score 1	-	42.02%	61.83%	83.23%
IIT Cut Score 2	28.07%	45%	66.77%	88.02%
		Group B		
Angoff	30%	53%	72%	86%
IIT Cut Score 1	-	47.06%	65.55%	83.51%
IIT Cut Score 2	29.67%	48.8%	67.9%	87.26%

Both panels were administered a rater satisfaction survey after they completed their ratings using both the Angoff and IIT methods. Overall every rater felt comfortable and confident in the ratings they provided using both methods. Raters were asked which method they preferred, where just over half responded they preferred the IIT method and found it more intuitive. However, there did seem to be a panel effect, where the panels preferred whichever method they used most recently. Seven of 11 raters preferred the Angoff method, but 6 of the 7 were all on the same panel. Similarly, 7 of 13 raters found the IIT method more intuitive, but these were the same 7 that preferred the IIT method and 6 of

the 7 were from the same panel. Overall, it seemed like preference displayed a proximity effect, where the preferred method was the most recent method used.

4.4 TIMSS Standard Setting

The TIMSS standard setting study consisted of 30 total SMEs randomly assigned to three different standard setting panels. The first panel performed the standard setting using the modified Angoff and answered the question: What is the probability a minimally competent examinees will get this item correct? The second method answered the same Angoff question but items and ability levels were presented randomly. The third panel performed the IIT standard setting method. Each panelist rated 25 items for three performance levels, resulting in 75 ratings for each panelist. Unfortunately two panelists failed to arrive for the modified Angoff method and one failed to show for the random Angoff method, resulting in panels of 8 and 9 individuals, respectively.

4.4.1 Detection of Cognitive Algebra Models

The detection of cognitive algebra models was done through the inspection of the factorial graph found in Figure 8, each individual rater graph found in Appendix B, and statistically through a repeated measures ANOVA. The visual inspection of the factorial graph revealed nearly parallel lines for the performance levels, which is indicative of an adding or averaging model. The repeated measures ANOVA produced significant main effects for levels ($F(24,216) = 8.01$, $p < .01$) and items ($F(2,18) = 291.33$, $p < .01$) as well as a significant interaction term ($F(48,432) = 2.37$, $p < .01$). However, the partial eta-squared was .01 for the interaction, representing a very small effect size. Since the main effects were significant and the interaction had a very small effect size, these results support an additive or averaging model. The results of the ANOVA can be found in Table 7.

Table 7 ANOVA Table for TIMSS Exam

	Sum of Squares	df	Mean Square	F	*p*
Level	18087.88	2	9043.94	291.33	<.001
Item	2143.91	24	89.33	8.01	<.001
Level x Item	317.18	48	6.61	2.37	<.001

4.4.2 Estimating the Cut Score

Since an adding model was positively identified across the TIMSS raters, estimates of the cut scores using the IIT data were calculated based on previously discussed methodology. The three different proposed methods of setting cut scores using IIT data produced different results. The first method, which took the difference between the marginal means of two performance categories, produced a suggested cut score of 53.64% between needs improvement and proficient, and a cut score of 79.12% between proficient and advanced. The second method set the cut score two standard deviations below the marginal mean and produced a cut of 25% for needs improvement, 48.17% for proficient and 81.36% for advanced. The third method, which estimated the raters weighting factor from the valuation stage of IIT produced a cut score of 30.60% for needs improvement, 50.34% for proficient and for 63.33% advanced. The estimated cut score for the Angoff method was also calculated using traditional Angoff calculations and resulted in a suggested cut score of 57.87% for needs improvement, 75.10% for proficient and 87.51% for advanced. The estimated cut scores are reported in Table 8.

Table 8 Estimated cut scores for TIMSS exam.

	Level 1 (Below Competent)	Level 2 (Competent)	Level 3 (Highly Competent)
Angoff	57.87%	75.10%	87.51%
IIT Cut Score 1	-	53.64%	79.12%
IIT Cut Score 2	25%	48.17%	81.36%
IIT Cut Score 3	30.60%	50.34%	63.33%

4.4.3 Procedural Validity Evidence

Thirty different raters participated in the TIMSS standard setting process. Each of the thirty raters were recruited with requirements that they had a masters in math education and were either currently teach math at the 8th grade level or a math curriculum specialist for 8th grade. Each group received a one hour introduction to the test and the task for their specific method from the same facilitator. After one hour, each panelist completed practice ratings for 7 items (for a total of 21 different ratings). After completing the practice ratings, if the panelists felt uncomfortable with the task they were encouraged to practice on seven additional items. Once panelists felt comfortable with the rating task they performed ratings for the 25 items from the TIMSS form.

After completing the rating, panelists in each of the three groups were encouraged to fill out a survey documenting their experiences. Each panelist reported they felt comfortable with the rating task and that the PLDs supplied were adequate for each group. Overall, the feeling for each group about the standard setting workshop was positive. The only complaints centered around deficiencies in the program where raters entered ratings. Since these comments were more about program functioning and not the method, these comments will not be discussed here.

4.4.4 Internal Validity Evidence

Internal validity evidence was collected by evaluating inter-rater reliability. Inter-rater reliability was assessed using two-way mixed ICC's. The ICC for the Angoff method was computed for the first round of Angoff ratings. The ICC for the Angoff method for the cut score was .812, .845, and .88 for the needs improvement, proficient, and advanced. The ICCs for the IIT method was .837, .829 and .832 for the same categories. The ICCs for the randomized Angoff were .056, .399, and .493. The modified Angoff method and the IIT method had relatively similar ICCs. However, when the modified Angoff method was randomized in the exact same way as the IIT method, the ICC's dropped significantly. This decrease may indicate a problem with conceptualizing the Angoff question.

4.4.5 External Validity Evidence

The TIMSS was the only exam with data to examine external validity. Each student was assigned to categories based on the cut scores suggested by each method. External validity evidence was then investigated in two steps. In the first step, correlations were used to assess the relationship between performance category assignments and variables which should correlate with performance levels. The second step assigned examinees to theoretical performance categories based on demographic and performance variables using a logistic regression function. After assigning each examinee to a performance level, correlations were used to assess the relationship between the theoretical performance level and assigned performance level from each method. Since the IIT method suggested three cut score for each level, each different method of deriving the cut score was analyzed.

Correlations between assigned cut scores (three from IIT and one from modified Angoff) and seven different variables were computed and are reported in Table 10. These variables were: how the student values math, the students belief in math being important,

the students expectations of their math performance, how prepared the teacher was to teach math, the teacher expectations the student, the mothers level of education and the number of books in the home. In general, the correlations between the IIT performance category assignments and the variables correlated higher than the Angoff method. These results are reported in Table 9.

Table 9 Correlations between cut score classifications and other variables

Method	Angoff	IIT1	IIT2	IIT3
N	402	402	402	402
Math Value	0.06	0.118	0.112	0.112
Math Effect	0.107	0.149	0.11	0.11
Math Expectations	0.278	0.389	0.339	0.339
Teacher Preparation	0.231	0.257	0.229	0.229
School Expectation	0.231	0.229	0.242	0.242
Mother Education	0.213	0.241	0.252	0.252
Books	0.224	0.337	0.362	0.362

*all correlations are significant at the $p < .001$ level.

The logistic regression function was computed on 2240 students, sampled from high and low performing students. The variables used in the logistic regression were: how the student values math, the students belief in math being important, the students expectations of their math performance, how prepared the teacher was to teach math, the teacher expectations the student, the mothers level of education and the number of books in the home. The regression equation was then used to assign 402 students to performance categories. The regression coefficients are reported in Table 10 Overall, 300 students were assigned to the proficient category and 102 to the basic category using this equation. Afterwards, correlations with the assigned group membership using the Angoff method was .241. The IIT method produced slightly higher correlations, with .394, .404 and .404 with the first, second and third methods, respectively. This information is presented in Table 11.

Table 10 Regression Coefficients for the TIMSS logistic regression predicitons

Regression Beta Coefficients			
Parameter	Estimate	Standard Error	Probability
Intercept	2.87	.63	<.001
Confidence in Math	-1.17	.14	<.001
Value Math	.15	.17	.37
Belief math	.29	.09	.003
Homework time	-.99	.13	<.001
Expectations	-1.28	.14	<.001
School expectations	.23	.05	<.001
Father education	.20	.04	<.001
Mother education	.10	.04	.02
Books in home	.76	.07	<.001
Teacher prep	-1.04	.1	<.001
Instruction time	-.003	.001	.04

Table 11 Correlations between logistic regression group membership prediction and different cut scores.

Method	Correlation of Logistic Pass Prediction
IIT method 1	.394
IIT method 2	.404
IIT method 3	.404
Modified Angoff	.241

4.5 Summary of Data Analysis

This section presented the results from seven different standard setting workshops from three different tests. There were three different modified Angoff workshops and three different IIT workshops. The final standard setting workshop was the modified Angoff task randomly presented, similar to the IIT method.

Across all three IIT standard setting workshops, an addition cognitive algebra model was positively identified through a visual inspection of the factorial graph and the lack of a

significant interaction in the repeated measures ANOVA. The positive identification of a cognitive algebra model indicated the accurate use of the 1-20 scale by the raters, as well as the conceptualization of an interval level scale. Since raters were utilizing an equal-interval scale, it was possible to linearly transform the 1-20 scale to a percent scale, which could then be used for performance category assignation of examinees.

Across all three exams, the cut-scores set by the IIT method were consistently less than the cut-scores set by the Angoff method. A complete overview of the cut scores is displayed in Table 12. This result is not a benefit or a detriment to either method, but just indicates that there is a systematic difference in the results of both methods.

Table 12 Overview of cut scores for each test and method

	Cut 1		Cut 2		Cut 3		Cut 4	
	Angoff	IIT	Angoff	IIT	Angoff	IIT	Angoff	IIT
HP	-	-	68%	50 - 53%	-	63 - 73%	-	-
Excelsior	33%	25 - 29%	59%	42 - 44%	75%	62 - 63%	87%	81-83%
TIMSS	58%	25 - 31%	75%	48 - 53%	87%	63- 81%	-	-

Finally, information about how long it takes to complete the rating task for both methods was collected for each of the conditions. For the HP Storage Solutions exam, it took raters less time to complete the IIT method, even though the raters were asked to do three times the ratings. This result primarily occurred because raters during the Angoff standards setting workshop had to discuss ratings and complete a second round of ratings, which was not the case for the IIT method. However, the total time required to complete the ratings for

the other two exams was similar, with the IIT method taking slightly less time than the Angoff method. These results are reported in Table 13.

Table 13 Average time for raters to complete the standard setting task

	HP (98 ITEMS)	Excelsior (100 items)	TIMSS (25 items)
Angoff	360 minutes (1 cut)	125 minutes	30 minutes
IIT	127 minutes (3 cuts)	118 minutes	27 minutes

Overall, the IIT method performed well compared to the Angoff method. Worries that raters would not be able to remain consistent because of the randomization of the method were unfounded. A score card comparing the Angoff method with the IIT method for all the aspects of the results section is found in Table 14. The table summarizes the procedural, internal and external validity data obtained during the course of the study. The two methods were evaluated on each criterion, and if one performed significantly better than the other then it is noted on the score card.

Table 14 Score Card Comparing Angoff and IIT methods

	Angoff	IIT
Procedural Validity		
Time Required	Equal	Equal
Preferred by Raters	X	
Perceived as more Valid		X
Internal Validity		
Inter-rater reliability	Equal	Equal
Intra-rater reliability		X
External Validity		
Corr math value		X
Corr math effect	Equal	Equal
Corr math expectation		X
Corr teacher prep	Equal	Equal
Corr school expectation	Equal	Equal

Corr with books in home		X
Corr with mother education	Equal	Equal
Corr with logistic regression prediction		X
Location of Cut Score		
Location of Cut (Percent Scale)	Higher	Lower

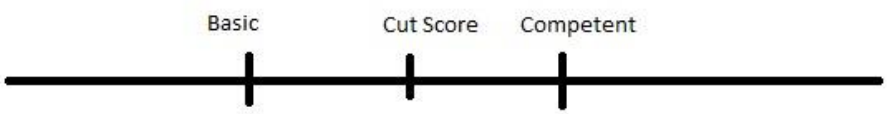


Figure 5 Theoretical Depiction of Cut Score

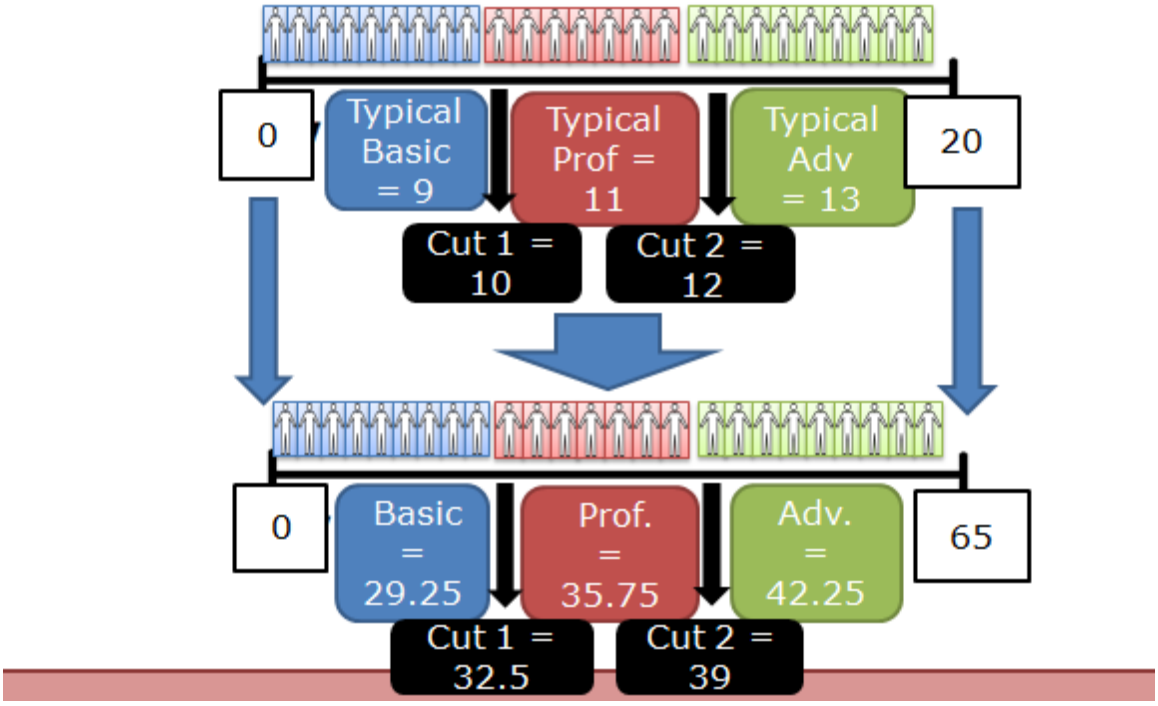


Figure 6 Example of Linear Transformation for IIT Scale

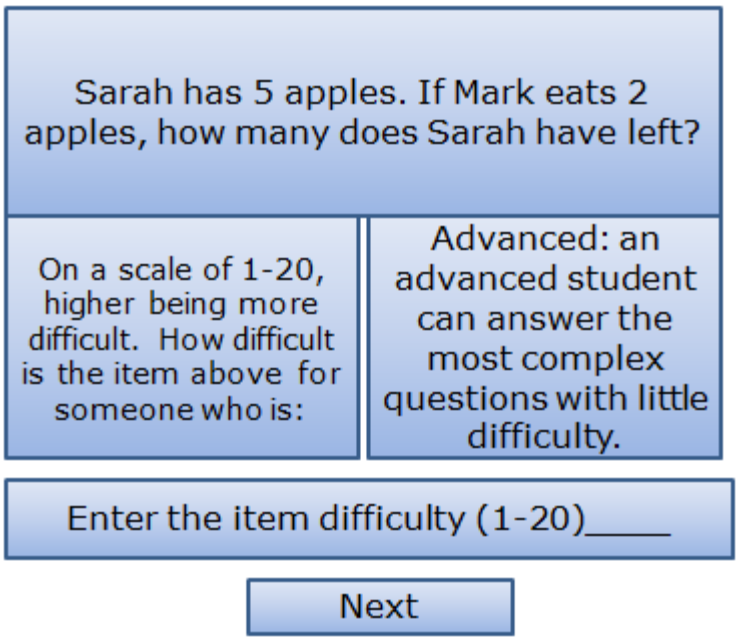


Figure 7 Computer Interface for IIT Method

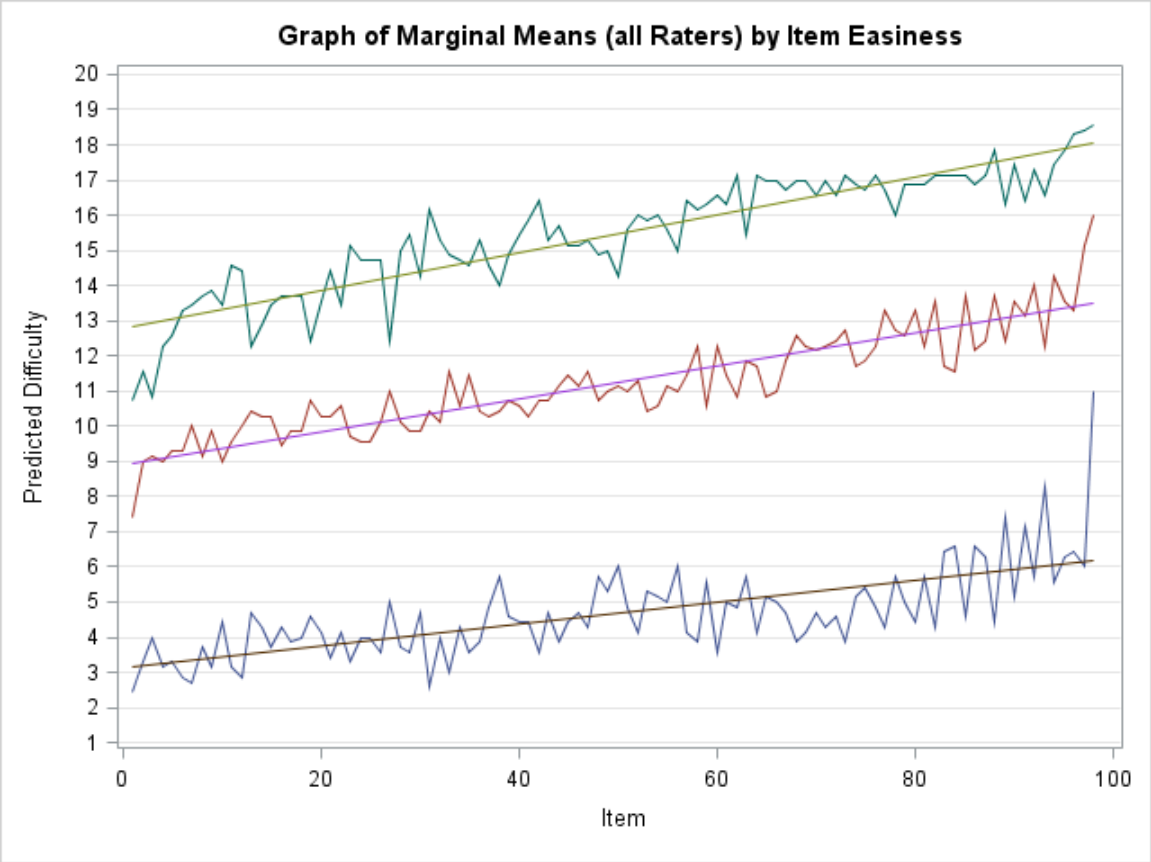


Figure 8 Average IIT graph for HP Storage Solutions

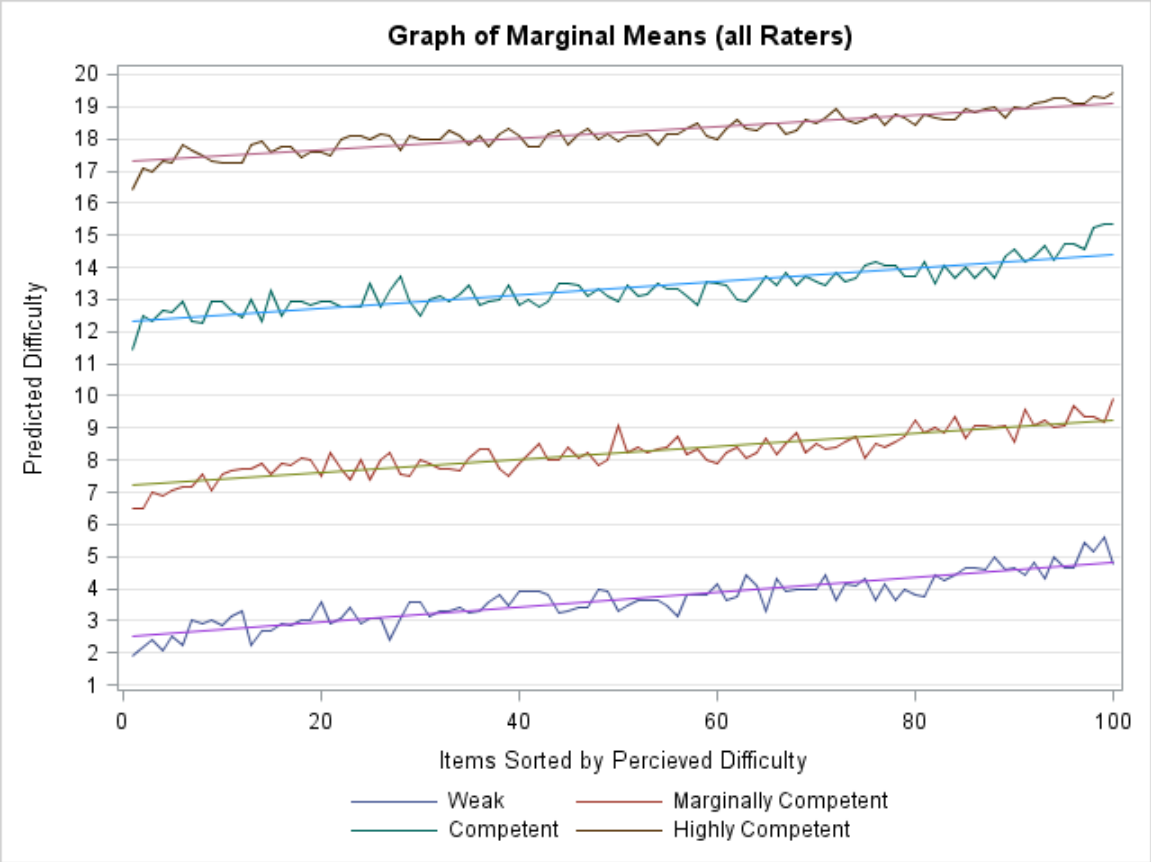


Figure 9 Average IIT graph for Excelsior College Nursing Exam

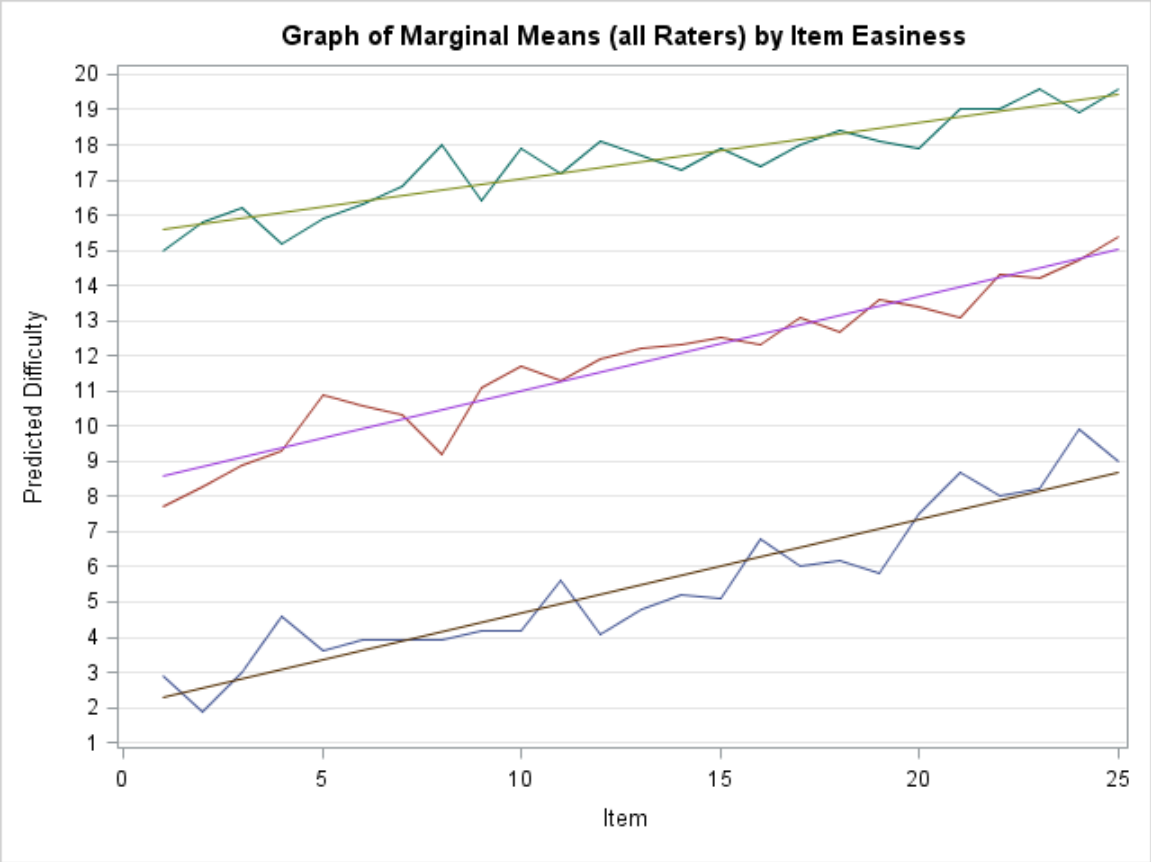


Figure 10 Average IIT graph for TIMSS Exam

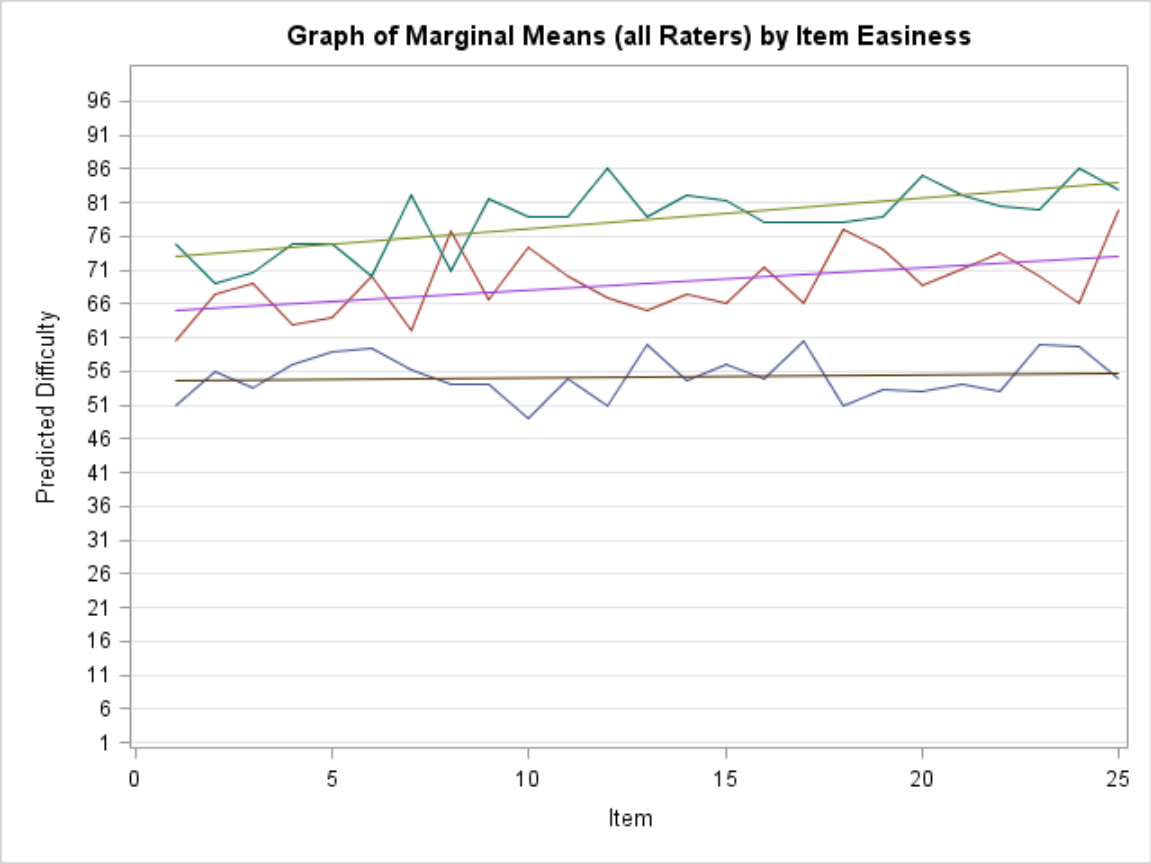


Figure 11 Average Randomized Angoff Graph for TIMSS Exam

CHAPTER 5

DISCUSSION

5.1 Introduction

The previous chapter detailed the results of this study. Chapter 5 serves to discuss prominent results, acknowledge limitations of the study, suggest future research directions, and present concluding remarks and recommendations for operational assessment practices.

5.2 Discussion of Findings

The primary goals for this study were to explore IIT as a potential method for setting cut scores. To accomplish this goal, cut scores were set on three exams using both the traditional modified Angoff method and the IIT method. To aid the interpretation of the IIT method, I utilized Kane's (2001) framework for evaluating the validity of cut scores by evaluating procedural, internal and external sources of validity when available. A discussion of the major findings follows.

5.2.1 Identifying Cognitive Algebra Models

The keystone of the current study was to determine if IIT was applicable to standard setting. The initial step involved the identification of cognitive algebra models. We hypothesized in the Methods section that we would identify either an additive or a multiplicative model through investigating the factorial graphs in ANOVA. Across all three exams, the factorial graphs for individual raters, as well as the average across all raters, displayed evidence of parallelism. It should be noted that individual raters occasionally

made mistakes and logical inconsistencies, where they rated an item harder for a more proficient examinee group than a less proficient group. However, these logical inconsistencies never occurred when rater responses were averaged across all raters. Therefore, based on the visual inspection of the factorial graphs, we concluded that the raters were combining factors using an additive model.

However, simply interpreting factorial graphs was not the only analysis conducted to determine which cognitive algebra model was being utilized. In addition to the visual analysis, a repeated measure ANOVA was utilized to see if a significant interaction existed between items and ability levels. The results of the ANOVA for each exam produced large effect sizes for both the performance level and item effects. For the TIMSS and the HP storage solutions exam there was also a significant interaction. However, the interaction in both cases had a very small effect size. These results provided further support of the additive cognitive algebra model. Since a cognitive algebra model was identified in the rater data, we concluded that IIT may provide valuable information when applied to standard setting.

5.2.2 Procedural Validity Evidence

Since a cognitive algebra model was identified, the next step included the collection of validity evidence to support the use of IIT in standard setting situations. The first type of validity evidence collected was procedural validity via three sources: facilitator observations, timings of standard setting workshops, and rater satisfaction surveys. Each standard setting workshop proceeded with no problems worth discussing. However, it should be noted that the facilitators were the same for the standard setting workshops in order to remove any facilitator effects on the examinees.

There were two important points that should be noted about procedural validity. The first is the difference in time required to complete the rating task. I hypothesized the IIT task would take more time, especially in the case of the HP storage solutions exam. However, this was not the case as the IIT method typically took about the same amount of time to complete ratings. The second important piece of procedural validity came from the Excelsior College nursing exam where raters rated items using both the modified Angoff method and the IIT method. Over 50% of the raters said that they preferred making ratings using the Angoff method as it allowed them to view previous ratings in order to make decisions. At the same time, most of the raters also stated they felt the IIT method produced more valid cut scores. Their reasoning was that since the ratings came randomly, it forced them to refer to the performance level descriptors more frequently and refresh their memory about the specific performance categories. The raters also stated that they preferred the 1-20 scale over the percent scale used in the modified Angoff method. Overall, important validity evidence was collected for both methods on each test. In general, the surveys provided by raters from both panels were very similar and all raters expressed satisfaction and confidence in their ratings.

5.2.3 Internal Validity Evidence

The second form of validity evidence analyzed was internal validity evidence. ICCs were used to evaluate the inter-rater reliability. Overall, the inter-rater reliability for both the Angoff and the IIT methods were comparable and were between .75 and .85 for all tests. However, the IIT method produced lower ICC's than the Angoff method for the Excelsior College nursing exam. This difference may be due to the fact that the Excelsior College ICC was calculated after the first round of panelist discussion. Since panelists had discussed and changed their ratings, dependencies were created between panelists that may have inflated

the inter-rater reliability. Despite this limitation, it appears that the ICCs for both methods were similar.

The IIT method provided two additional sources of internal reliability evidence beyond what was available for collection during the Angoff workshop. The study design hypothesized the use of an additive model if raters could adequately understand the task required. If raters are performing consistently, then the factorial graphs will rarely display logical inconsistencies (e.g., items that are rated easier for less proficient groups). When raters perform inconsistently, logical inconsistencies would be more common. For all three tests, logical inconsistencies were uncommon for the majority of raters. This result indicates raters understood the rating task similarly as well as provided logically consistent results despite item randomization. The only situation where raters did not perform consistently was when item by ability combinations were randomized using the modified Angoff question.

The second form of validity evidence that is difficult to collect for the modified Angoff method but simple for the IIT method is a measure of intra-rater reliability. Intra-rater reliability is the degree to which a rater is consistent with themselves. The most common way to measure intra-rater reliability is through test-retest procedures where raters perform a task and return weeks later to perform the same task after they had forgotten their previous ratings. Therefore, test-retest reliability is difficult to obtain with the Angoff method as raters may simply review their work and discover a repeated item. The IIT method however presents stimuli randomly without the ability to return to previous ratings. Because raters reported in the rater satisfaction surveys that they could never remember what they had put for a previous item by ability combination, it was possible to have raters rate 10 items twice. The results were impressive, as each of the 7

raters who completed this task had intra-rater reliabilities above .75 based on 10 items. The predicted reliability for each of the raters for a 100 item test was over .95 for each rater, with some raters having perfect reliability for certain cut scores. This information indicates that raters were remarkably consistent with their own previous ratings even though item presentation was randomized.

While most of standard settings based on the IIT method produced observed parallelism in the factorial graph, this was not the case when the question and scale were taken from the Angoff method. In the random Angoff method, raters were asked about the proportion of minimally competent examinees in each ability level who would get the item correct. Similar to the IIT method, the items and ability levels were randomized. The only difference between the IIT method and the randomized Angoff method was the question and the scale. However, raters were unable to remain consistent, despite these being the only changes. There was no observed parallelism, and there were no discernible patterns in the factorial graphs. Inter-rater reliability was also very low. These results suggest that raters remain consistent when performing the Angoff rating task because they are allowed to review previous ratings. It may also suggest a fundamental flaw with the Angoff rating question that deserves more attention in future research.

5.2.4 External Validity Evidence

External validity evidence was only available for the TIMSS exam since the other tests had not been used operationally and there was no examinee level data. The current study examined correlations between examinee classifications on the TIMSS using the Angoff and IIT methods with variables empirically shown to predict student achievement. In each case the IIT classifications (regardless of the method of deriving a cut score), correlated higher with these external criteria than the Angoff classifications. In addition to

simple correlations, a logistic regression function was developed to predict if a student would likely be proficient based on external criteria. The predicted classification membership was then correlated with actual group membership. Similar to the correlations, IIT classifications correlated better with the logistic regression prediction of classification membership than the Angoff method. While all external validity should be interpreted with caution, what we were trying to achieve by comparing student scores to external criterion was to demonstrate that the IIT method can produce quality cut scores that are related to external variables.

5.2.5 Evaluating Rater Graphs

Perhaps one of the greatest contributions of IIT to standard setting is it provides a framework through which it is possible to evaluate rater performance. To date, all operational standard settings which used this method have demonstrated an additive cognitive algebra model. However, not all raters responded to the IIT ratings used this model. Take for example Rater 3 and Rater 5 as shown in Figures 12 and 13. These two raters were responding to the same items and ability levels, but rater 3 completed the ratings in a much more consistent manner than rater 5. The three different ability levels in rater 5's graphs are almost indistinguishable. The rater had numerous logical inconsistencies, where he rated an item easier for less proficient groups than higher proficient groups resulting in crossing lines. The rater also did not utilize the full 20 point scale. This pattern does not fit the hypothesized cognitive algebra model and indicates the second rater was not performing the same cognitive task as the first. This problem could occur for several reasons. For example, the rater may have misunderstood the task, provided random responses or is simply not good at identifying the difficulty of items. Whatever the case may be, this is an example where the rater may need to be retrained and

asked to repeat the task or removed from the final analysis when determining cut scores. Since the IIT method provides a hypothesis for how raters should interact with the rating process, it is possible to evaluate raters who do not fit the hypothesis. If raters are not performing the cognitive task, or IIT is not applicable for a given rater, then it may be possible to eliminate or weigh underperforming raters less in the calculation of the final cut score. This rater by rater analysis may provide a way to improve the validity of cut scores by identifying raters for removal or retraining.

5.3 Limitations of the Current Study

A few limitations exist for this study and many will be discussed within this section. First, the current study represents the first application of IIT to the measurement literature and inevitably could not cover everything necessary to completely explore a new standard setting method. One important limitation is the lack of understanding with how raters were conceptualizing and integrating the 1-20 scale. It is difficult in any study to understand the cognitive processes of the individuals involved. While mathematics and IIT dictate that because a cognitive algebra model was identified the 1-20 scale is a simple linear transformation to any other scale. However, this may not be true. While mathematically it is possible to map the 1-20 scale onto a percent scale, a proportion scale or even the theta scale, the two scales my not be conceptualized cognitively in the same way, introducing error into the transformation. The raters themselves may not conceptualize the 1-20 scale and the percent scale in the same way, making the transformation cognitively incorrect.

The second limitation is the lack of quality in the external validity information. This problem is not limited to the variables investigated in the current study, but is a general problem inherent in all standard setting external validity studies. Due to measurement

issues and multiple sources of error in the external data, it is difficult to draw strong conclusions from the external validity evidence gathered in the current study.

The current study covered a very broad range of topics with respect to IIT and standard setting. The study covered three exams, seven different workshops and provided analysis for both modified Angoff results and IIT methods using experimental conditions. However, with such a broad scope, many specific topics of the method were left uninvestigated. These topics should still be researched through critical evaluation and experimentation.

5.4 Directions for Future Research

There are several possibilities of future research that could provide valuable information about the quality of the IIT method. These future research studies should focus on the areas of research not covered by the current study and provide empirical research to fill gaps in the research surrounding the application of IIT to standard setting.

One important area for future research was previously discussed in the limitations section. There still needs to be research focused on understanding how the rater cognitively approaches the IIT standard setting process and how they cognitively utilize and interact with the scale. While such research is not limited to the IIT method in standard setting, the novelty of the IIT method provides interesting opportunities to investigate rater cognition. Such research is especially important for the IIT standard setting method, as IIT is based in cognitive psychology and provides a framework for the evaluation of the cognitive processes of raters.

A second avenue of future research involves investigating the accuracy of rater judgments with relation to empirical item estimates of difficulty. Many studies have focused on the accuracy of rater perceived difficult with respect to the Angoff method

(Clauser, 2011). While many of these studies have shown human ratings poorly reflect empirical item difficulty, the IIT method should still be subject to the same rigorous research. Similar to the Angoff method, the IIT method asks raters to evaluate the conditional difficulties for items. Given this similarity, one would expect a positive relationship between perceived conditional difficulties and empirical conditional difficulties.

A third branch of research could focus on the salient benefits of IIT, such as the ability to analyze rater performance to identify poorly performing raters. Studies could focus on how best to evaluate the factorial graphs or utilize intra-rater reliability in order to weigh, or eliminate completely, ratings from poorly performing judges in the final suggested cut scores. IIT provides interesting and unique ways to evaluate rater performance, which may prove to be one of the greatest contributions of integrating IIT into standard setting workshops.

One final area of research should involve a closer inspection of the mathematical factors at work within IIT. Specifically, there were three areas that would need better mathematical justification through empirical research: estimating the weighting of the factors, how best to scale rater responses to an appropriate test scale and developing methods to derive suggested cut scores from IIT ratings. Estimating weighting factors could be improved through the application of an iterative estimation procedure. The current study used a simple linear transformation to scale the 1-20 scale onto a percent scale; however, there may be applicable research in equating that could more accurately scale suggested cut scores. Finally, three different methods of deriving cut scores using IIT were investigated in the current analyses; however, there are undoubtedly other methods of producing cut scores using IIT data that may provide more accurate results.

5.5 Benefits of the IIT Method

This paper has demonstrated the similarities and differences between the Angoff standard setting method and the IIT method. In general both methods produce similar levels of inter-rater reliability. The IIT method demonstrated better correlations with external criteria than the Angoff method. The IIT method offers several unique benefits that deserve additional attention. The current section emphasizes some of the potential contributions offered by applying IIT to standard setting.

5.5.1 Theory Driven

Messick (1989) stated that validity refers to "the degree to which evidence and theory support the adequacy and appropriateness of interpretations." While Messick was primarily concerned with test score interpretations, his argument can be appropriately applied to the validity of standard setting.

One important point highlighted throughout Messick's article is the importance of both empirical evidence and theory to validity arguments. Typical evaluation procedures of standard setting workshops focus on empirical evidence through the collections of ratings and reliability coefficients. The theory behind each standard setting workshop is infrequently and insufficiently addressed.

Perhaps the greatest benefit of the application of IIT to standard setting is that cognitive psychology theory is applied and evaluated in each workshop. Theory is used in the development and evaluation of the standard setting workshop. Each rater is subjected to a hypothesis that they are combining elements of the standard setting procedure using cognitive algebra. Inferential tests and hypotheses can be evaluated and discarded based on theory. The IIT method is perhaps the only application of psychological theory to standard

setting and provides additional support to standard setting validity claims above typically utilized empirical evidence due to its theoretical nature.

5.5.2 Evaluation of Raters

A second meaningful contribution of IIT to standard setting is that it provides a framework for evaluating raters. Because IIT is based on theory, it is possible to derive expectations and a hypothesis for the performance of raters. In this first exploratory phase of IIT in standard setting, we concluded that raters typically use an additive model when combining different stimuli to make an item difficulty judgment. We may therefore approach future studies with the theory that raters will continue to utilize an additive model, and raters who are not performing the task adequately may be declared unqualified raters or simply do not understand the task. IIT provides an empirical framework and theory for evaluating the performance of raters during the standard setting workshop.

5.5.3 Additional Sources of Reliability

Important validity information about the standard setting process can come from theory or empirical evidence. The majority of empirical evidence collected for standard setting involves the calculation of inter-rater reliability. These reliability estimates give the general cohesion of all the raters who participated in the standard setting workshop. However, these reliability estimates do not give sufficient evidence that a rater performing the task a second time would produce similar results.

The IIT method provides two additional ways of evaluating reliability that are not currently calculated in standard setting practice: a calculation of intra-rater reliability and factorial graphs. Both of these additional sources are easy to gather within an IIT framework. Factorial graphs are a natural product of IIT and can be evaluated in different

ways to evaluate the reliability of a rater. Intra-rater reliability can be calculated by having raters rate the same item multiple times during the rating phase of the standard setting workshop.

While inter-rater reliability estimates were roughly equivalent between the Angoff and IIT methods, the IIT method provides more sources of reliability information that cannot be gathered in typical operational standard setting.

5.6 Conclusions and Recommendations

The goal of the current study was to show that IIT may be applicable in standard setting situations. The general conclusion is that cognitive algebra models were utilized during the rating process utilized by SMEs when making item rating judgments. However, despite the quantity of data collected in the current study, there still remain a large number of research projects that need to be undertaken. The method is still in development, so it is important to conduct additional research.

The current study demonstrates several areas where IIT may offer improvements to current standard setting methods. IIT can provide important information about the cognitive processes involved in the rating process. Applying this additional information may provide ways to evaluate rater performance and evaluate if raters understand the rating process. Setting up the standard setting workshop using an IIT design provides additional sources of internal validity evidence.

Based on the research conducted in this study, IIT is applicable and useful to the standard setting process. However, much more research needs to be conducted before the standard setting method is ready to be utilized in high-stakes standard setting workshops. However, the research does highlight the potential benefits of IIT in standard setting. With

additional effort and research, the IIT method will provide a practical and valuable tool to improve the quality of standard setting.

5.6 Figures

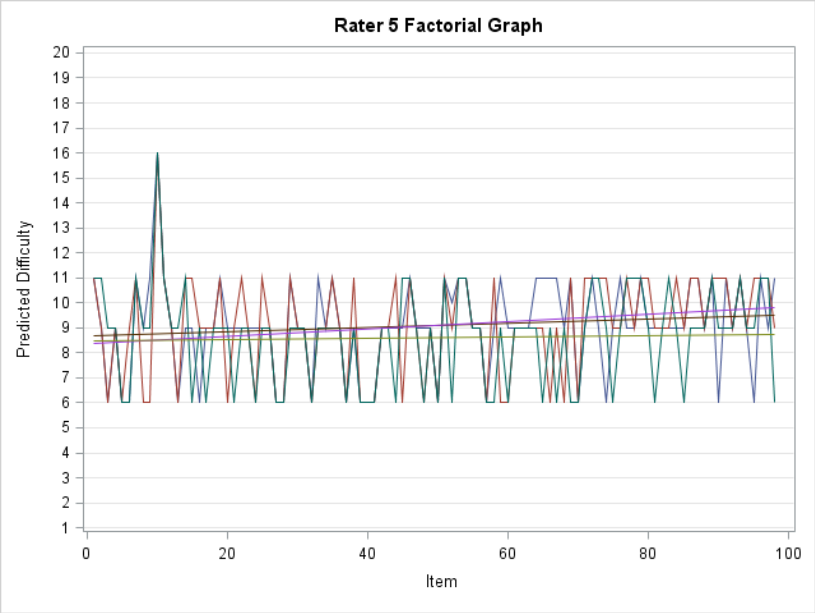


Figure 12 Rater 5 from HP Storage Solutions Exam

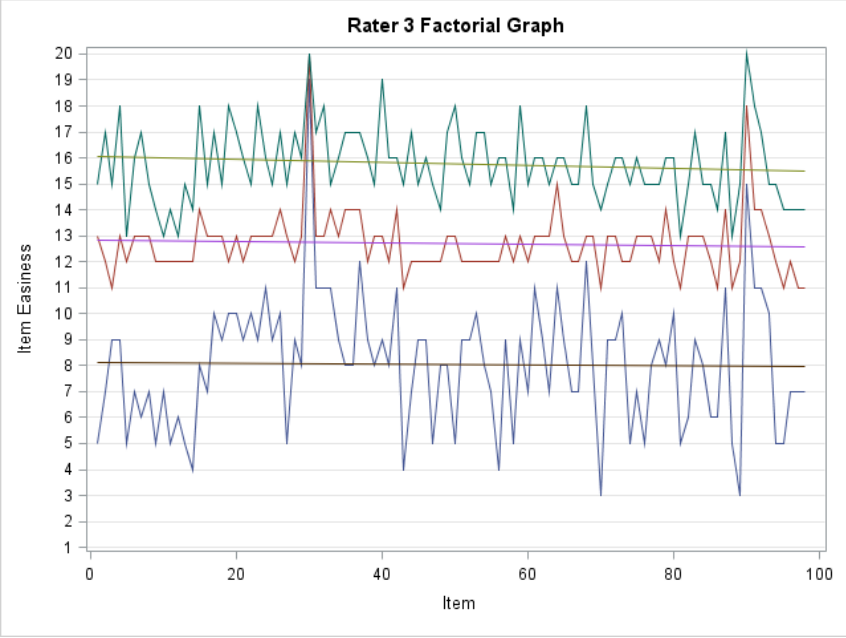


Figure 13 Rater 3 from HP Storage Solutions Exam

APPENDIX A

RATER EVALUATION FORM

Evaluation Form

The purpose of this evaluation form is to obtain your opinions about the standard-setting study. Your opinions will provide a basis for evaluating (1) the training you received, (2) the standard-setting method you applied for the last week, and most importantly, (3) the performance standards that you and other panelists would be recommending for the given exam.

1. We would like your opinions concerning the level of success of various components of the standard-setting study. Mark in the column that reflects your opinion about the level of success of these various components of the standard setting study.

	Not Successful	Partially Successful	Successful	Very Successful
Advance information about meeting				
Introduction to Exam				
Review of ability levels				
Training activities				
Practice Exercise				

2. In applying the standard-setting method, it was necessary to use three ability levels: Unqualified, Qualified, Highly Qualified. Please rate the definitions provided for these performance levels in terms of <u>adequacy</u> for standard setting.

	1	2	3	4	5
Unqualified					
Qualified					
Highly Qualified					

3. How comfortable are you with your understanding of the purpose of this exam?
    a. Very Comfortable
    b. Comfortable
    c. Somewhat Comfortable
    d. Not Comfortable

4. How comfortable are you with your understanding of the uses of the scores from this achievement test?
    a. Very Comfortable
    b. Comfortable
    c. Somewhat Comfortable
    d. Not Comfortable
5. How would you judge the amount of time spent on training in preparing yourself to make item difficulty judgments?
    a. About right
    b. Too little time
    c. Too much time
6. How adequate was the training provided on the standard setting method used?
    a. Totally Adequate
    b. Adequate
    c. Somewhat Adequate
    d. Totally Inadequate
7. How would you judge the amount of time spent on training?
    a. About right
    b. Too little time
    c. Too much time
8. How would you rate the amount of time allotted to perform the judgment task?
    a. About right
    b. Too little time
    c. Too much time
9. Indicate the importance of the following factors in your judgments.

	Not Important	Somewhat Important	Important	Very Important
The descriptions of unqualified, ideal qualified and highly qualified				
Your perceptions of the difficulty of the items				

Your own experience				
Your knowledge of content				
Previous judgments made on the item for other ability levels				

10. What confidence do you have in ratings of examinees at the UNQUALIFIED level?
    a. Very High
    b. High
    c. Medium
    d. Low
11. What confidence do you have in ratings of examinees at the IDEAL QUALIFIED level?
    a. Very High
    b. High
    c. Medium
    d. Low
12. What confidence do you have in ratings of examinees at the HIGHLY QUALIFIED level?
    a. Very High
    b. High
    c. Medium
    d. Low

13. What strategies did you use to assign difficulty ratings to items?

14. Please provide ways to improve the METHOD.

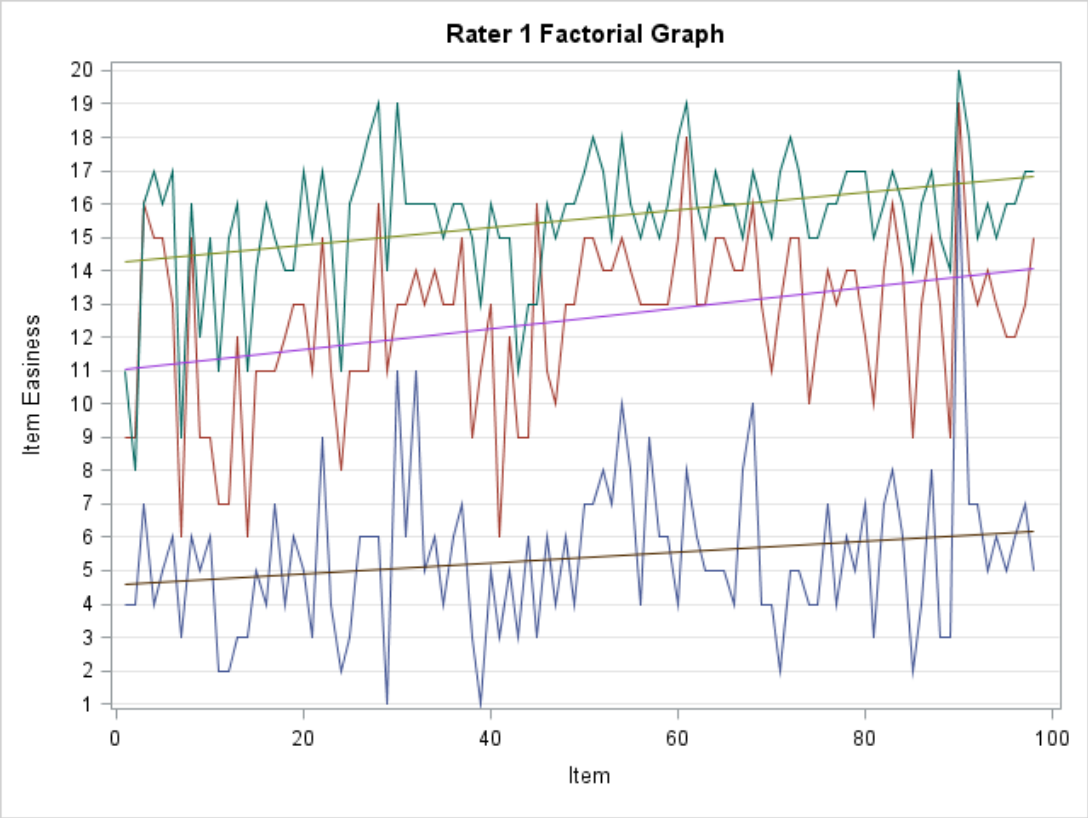
15. Please provide ways to improve the program (Which implements the method).

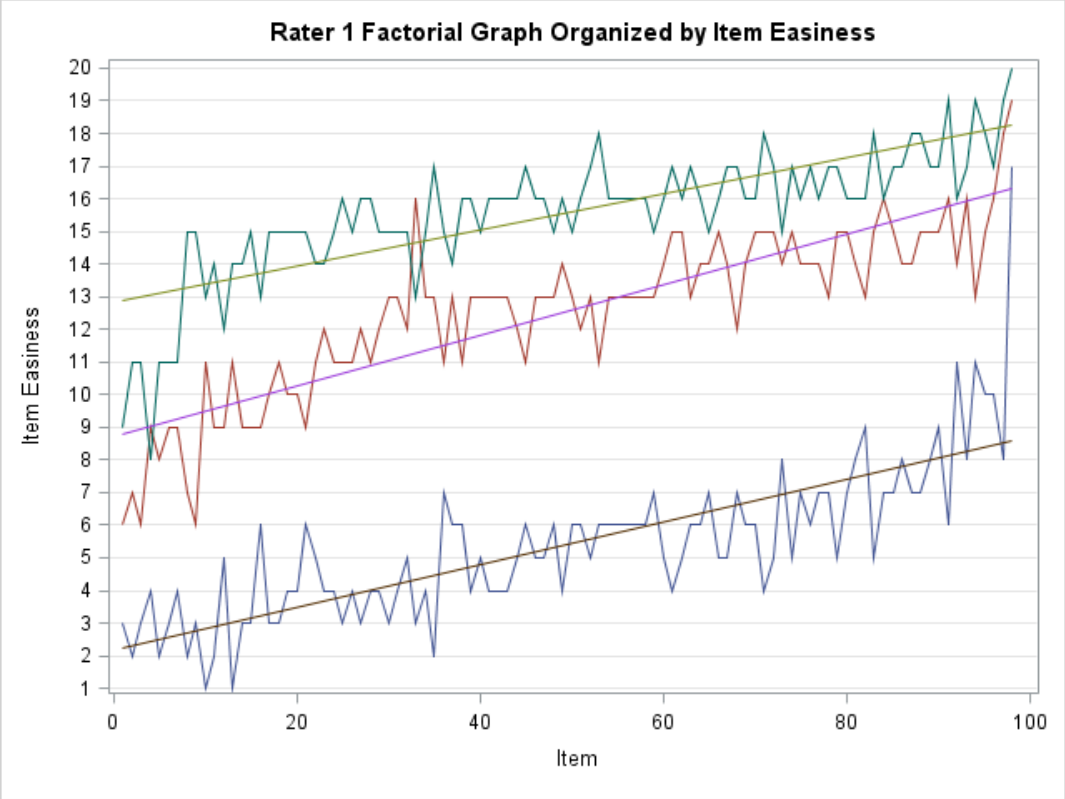
APPENDIX B

FACTORIAL GRAPHS

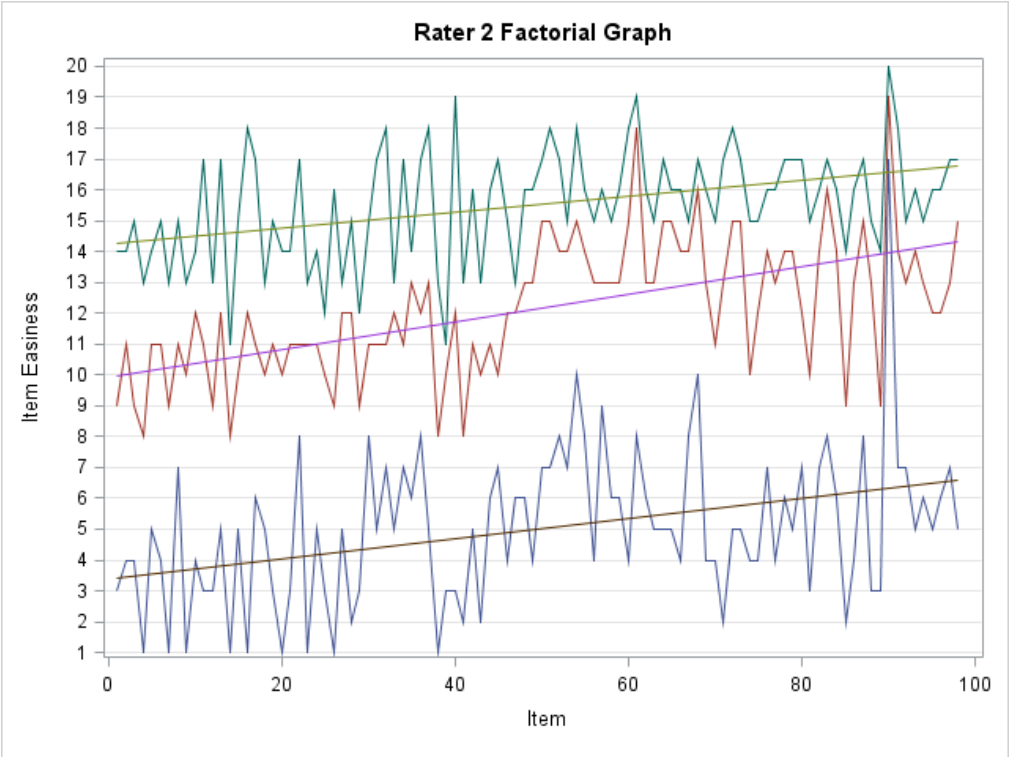
B.1 IIT Factorial Graphs For HP Storage Solutions Exam

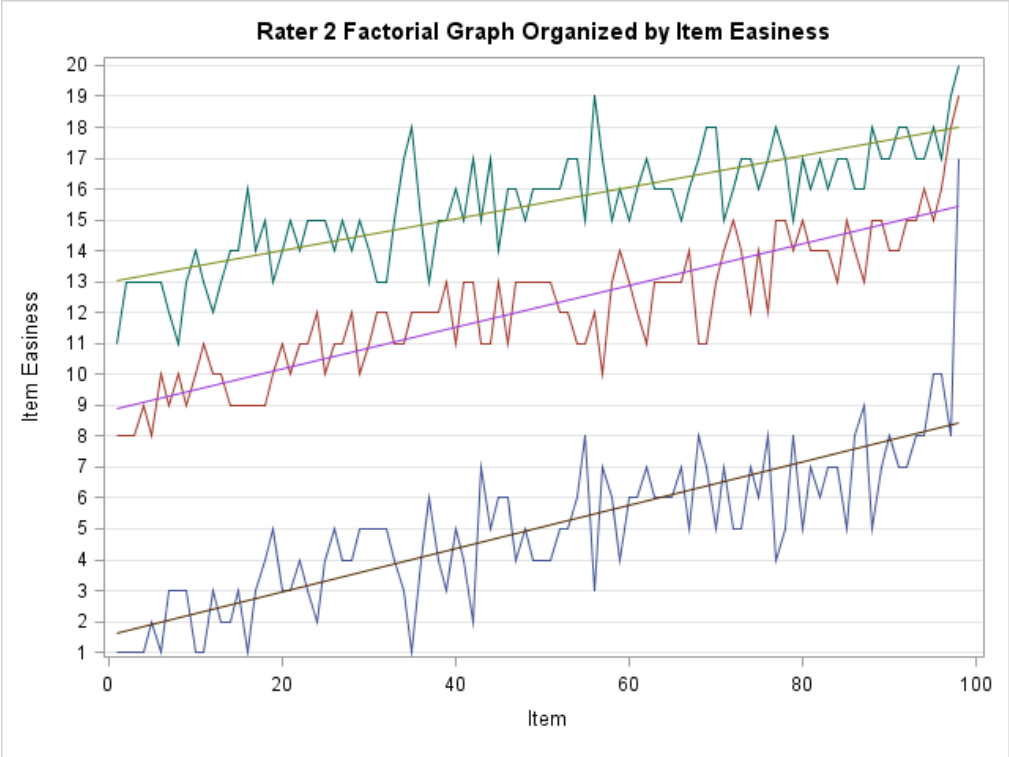
B.1.1 IIT Factorial Graph for Rater 1.



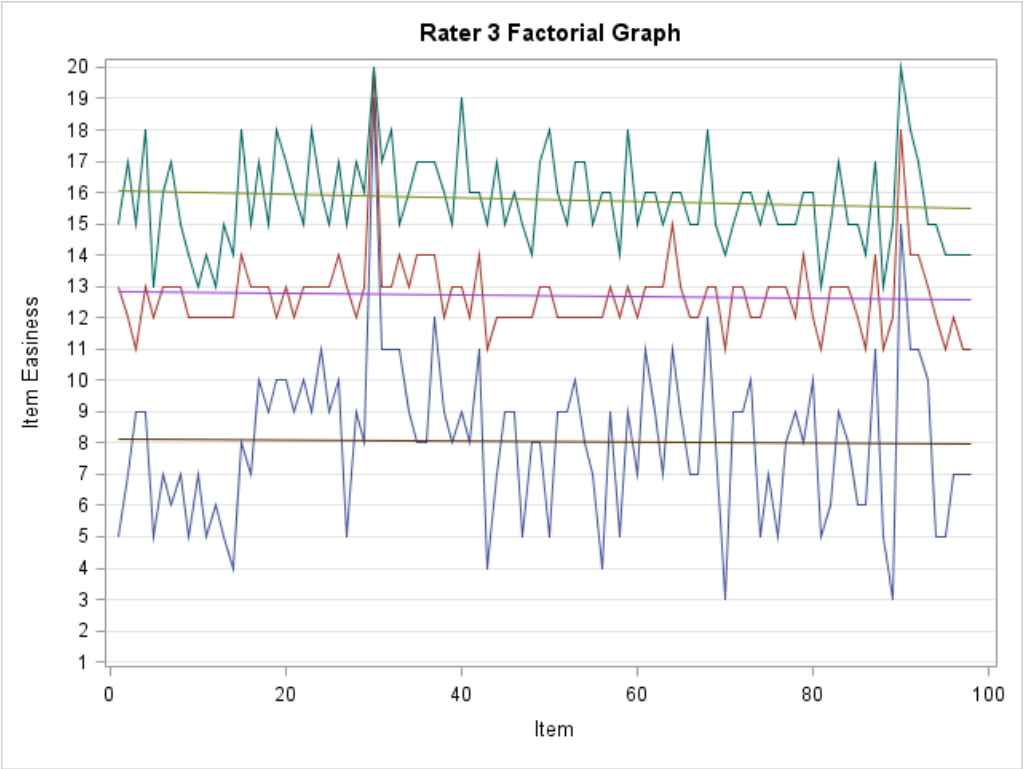

Rater 1 Factorial Graph Organized by Item Easiness
Item Easiness
Item
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
20
40
60
80
100

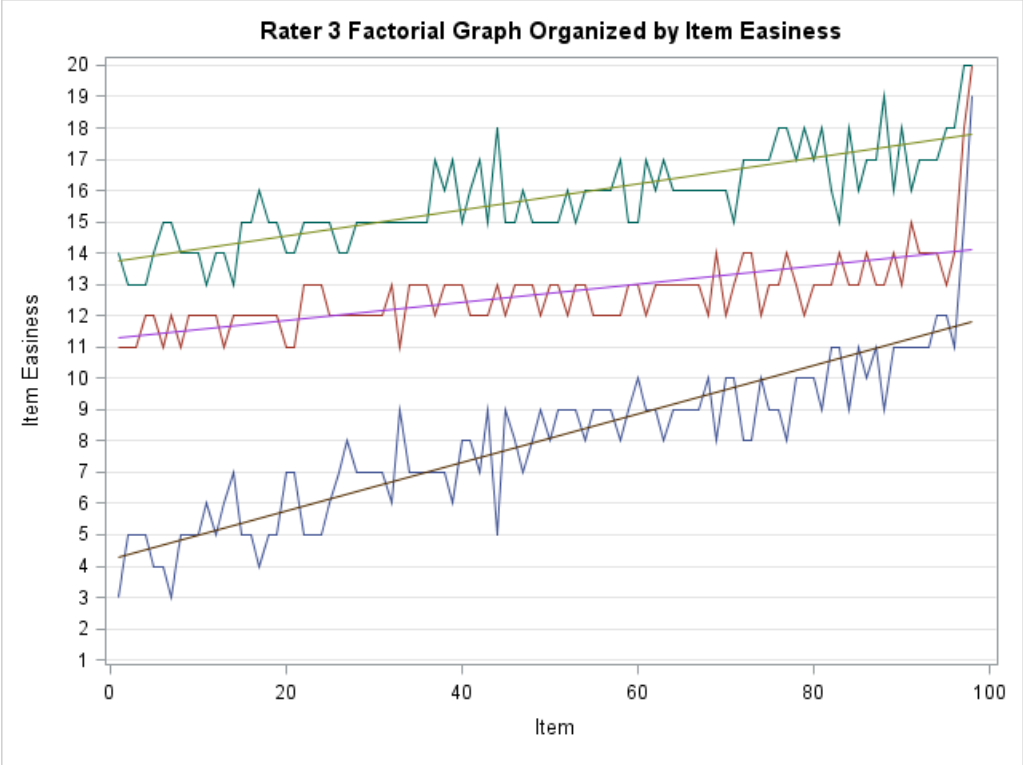
B.1.2 IIT Factorial Graph for Rater 2.



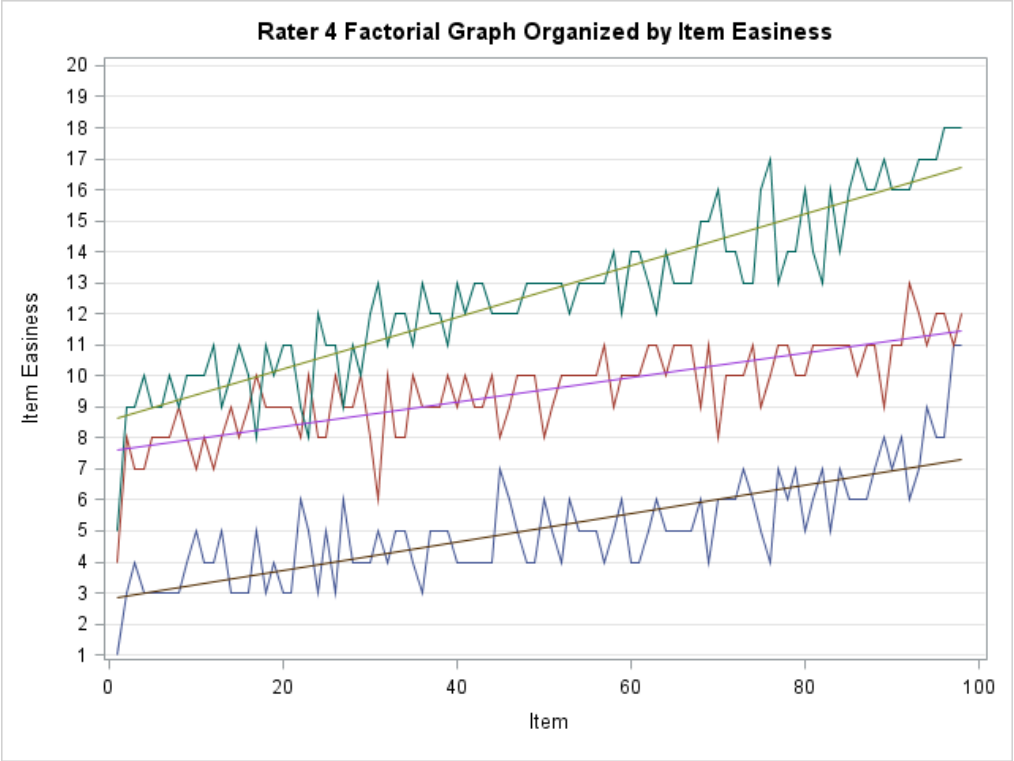


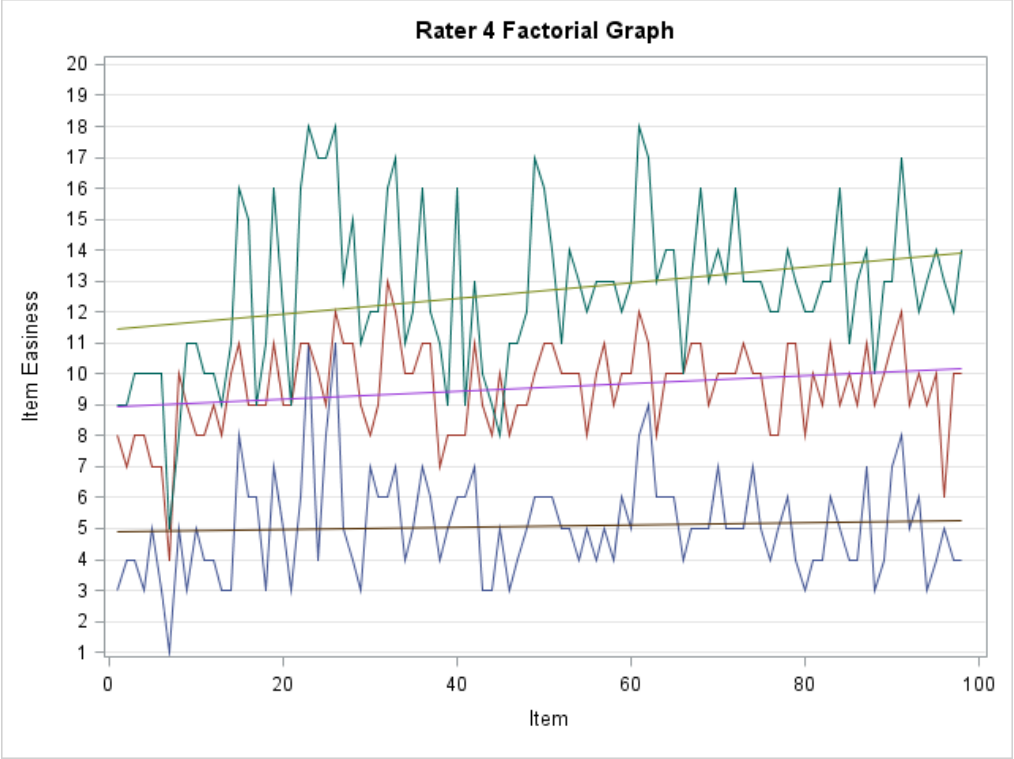
B.1.3 IIT Factorial Graph for Rater 3.



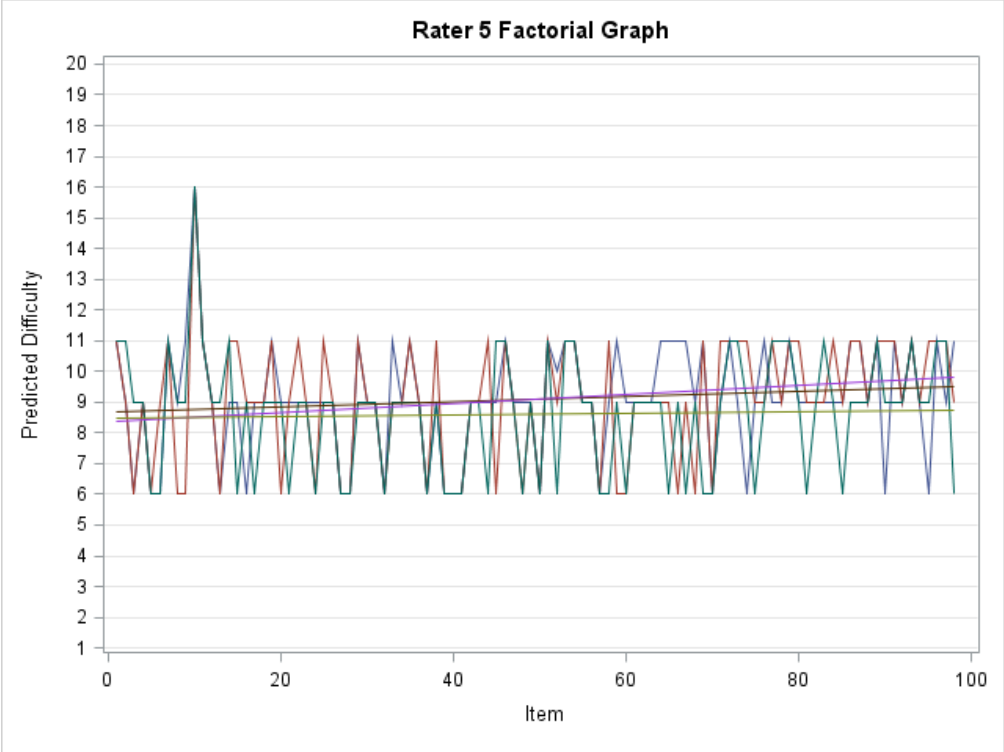


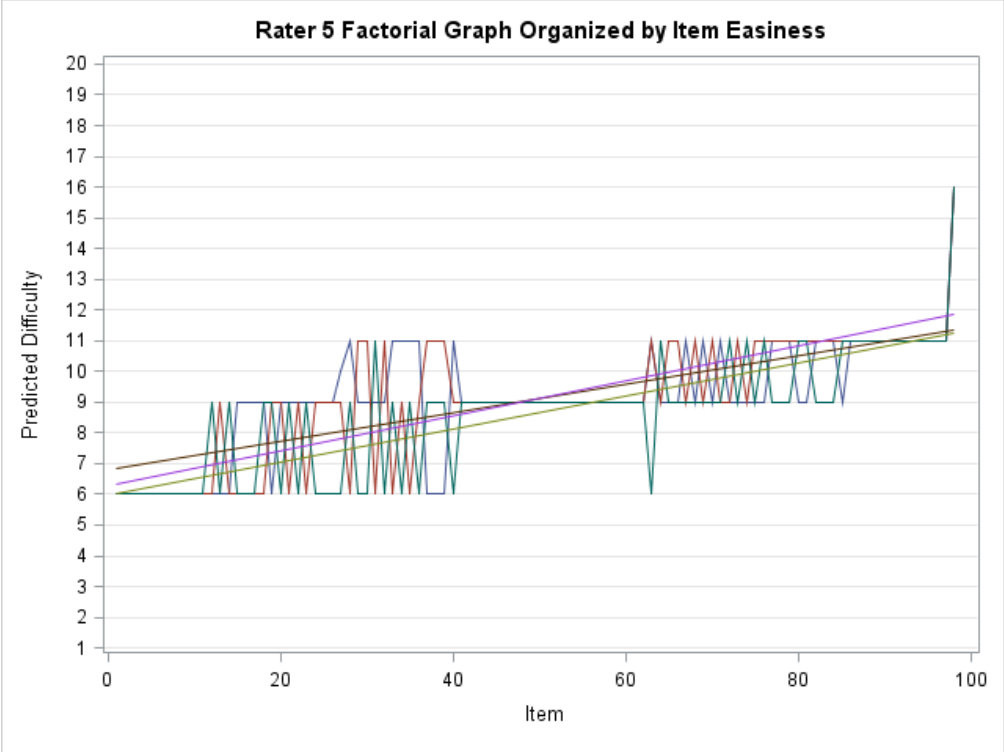
B.1.4 IIT Factorial Graph for Rater 4.



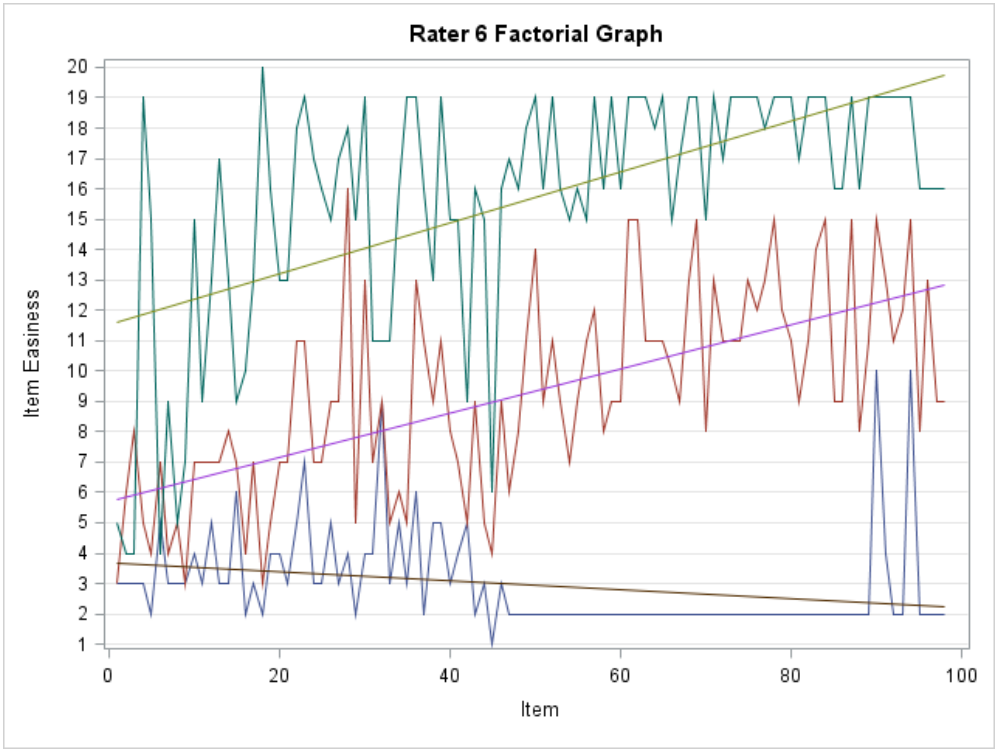


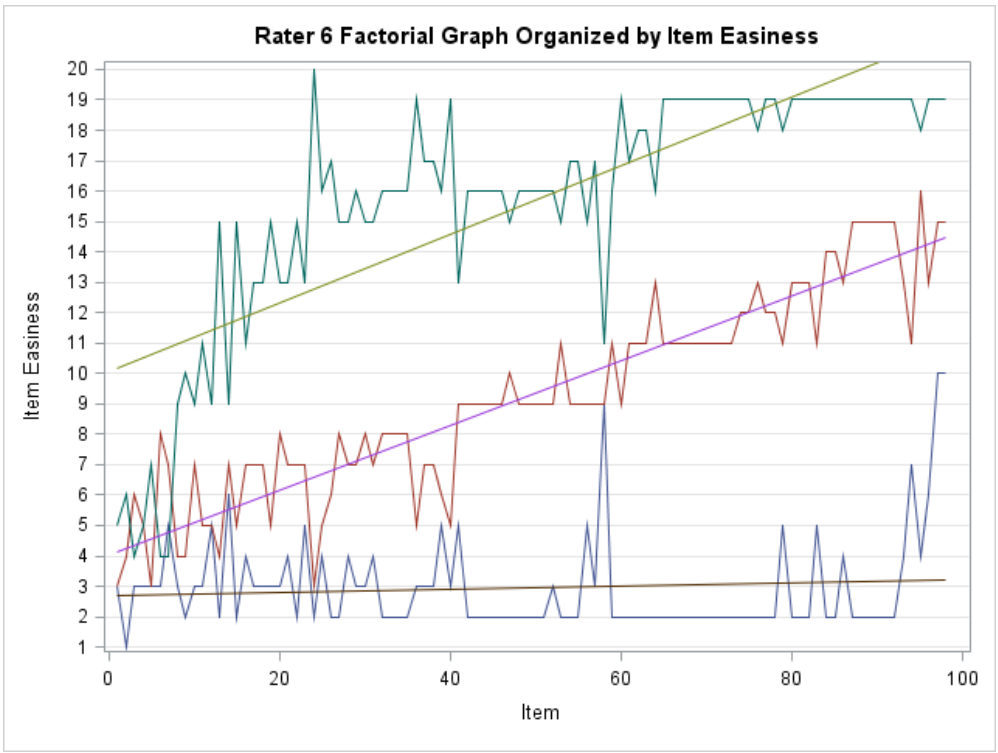
B.1.5 IIT Factorial Graph for Rater 5.



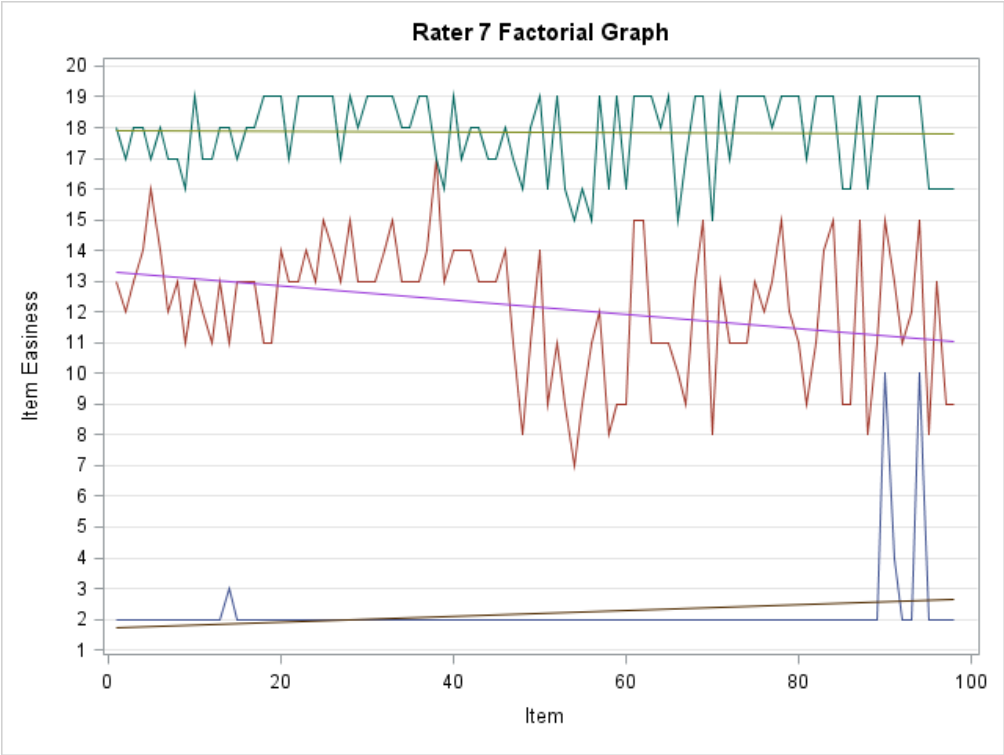


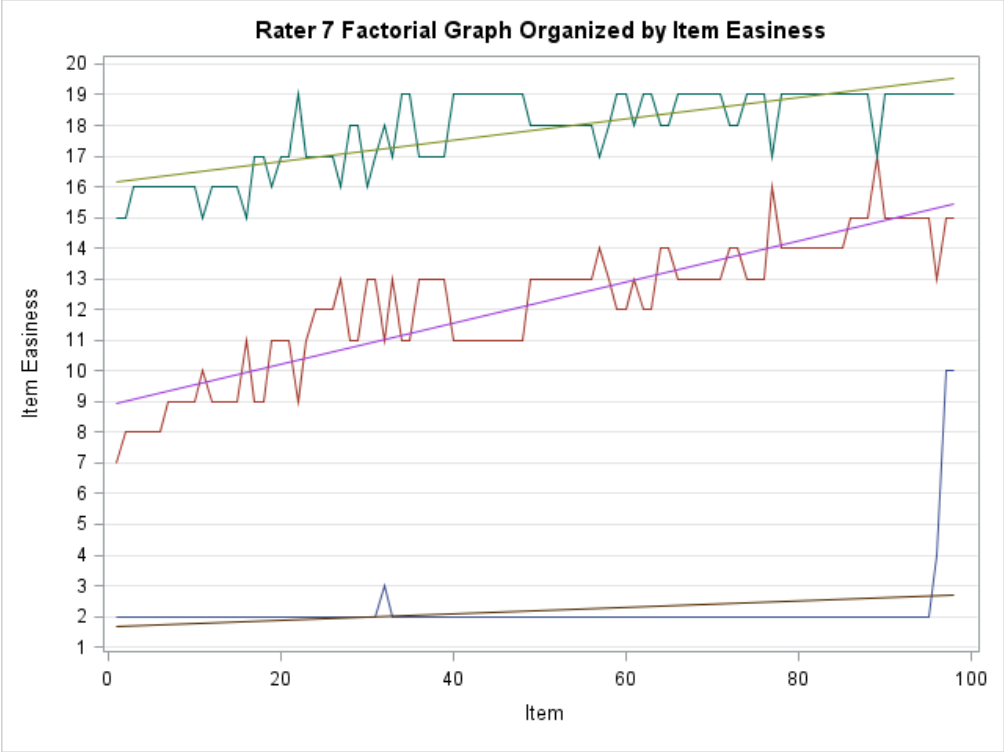
B.1.6 IIT Factorial Graph for Rater 6.





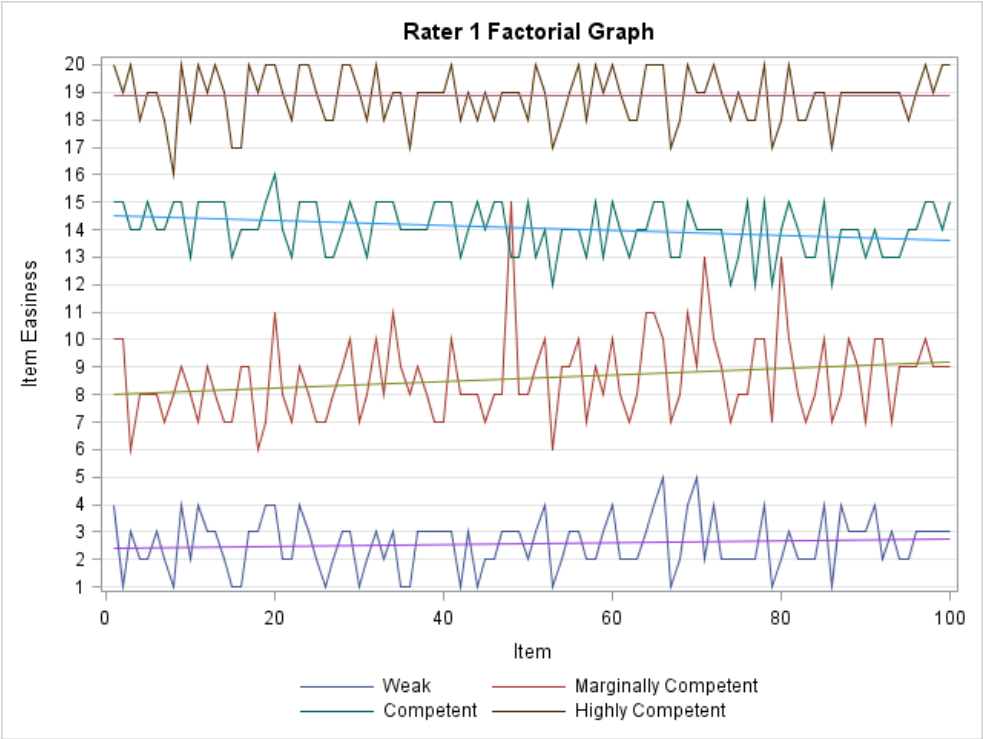
B.1.7 IIT Factorial Graph for Rater 7.

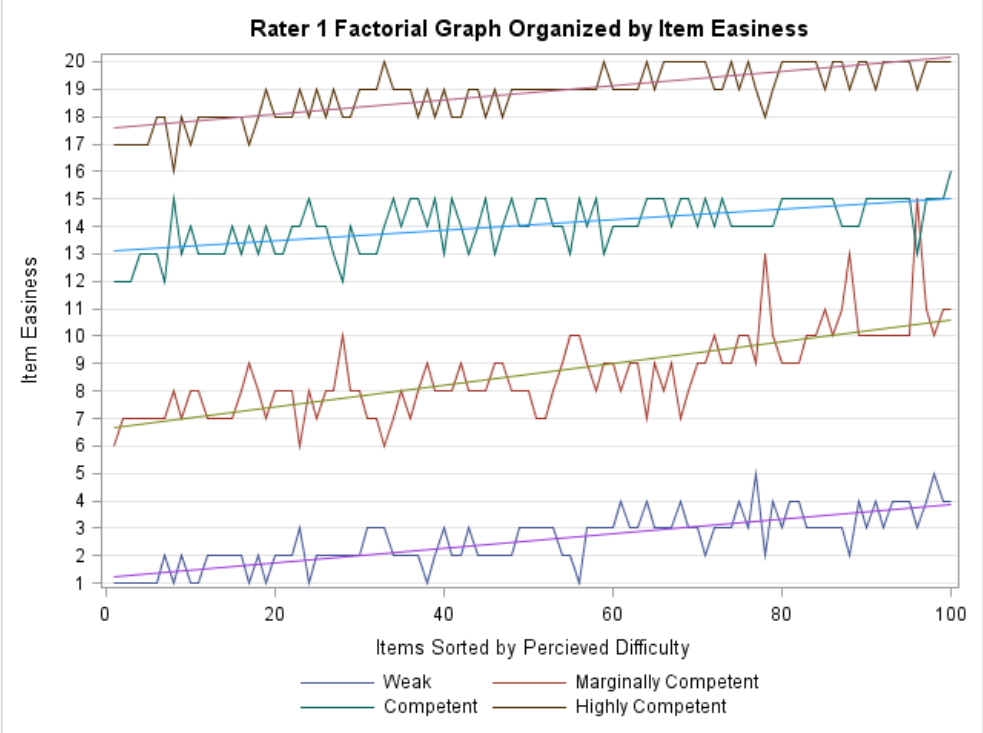




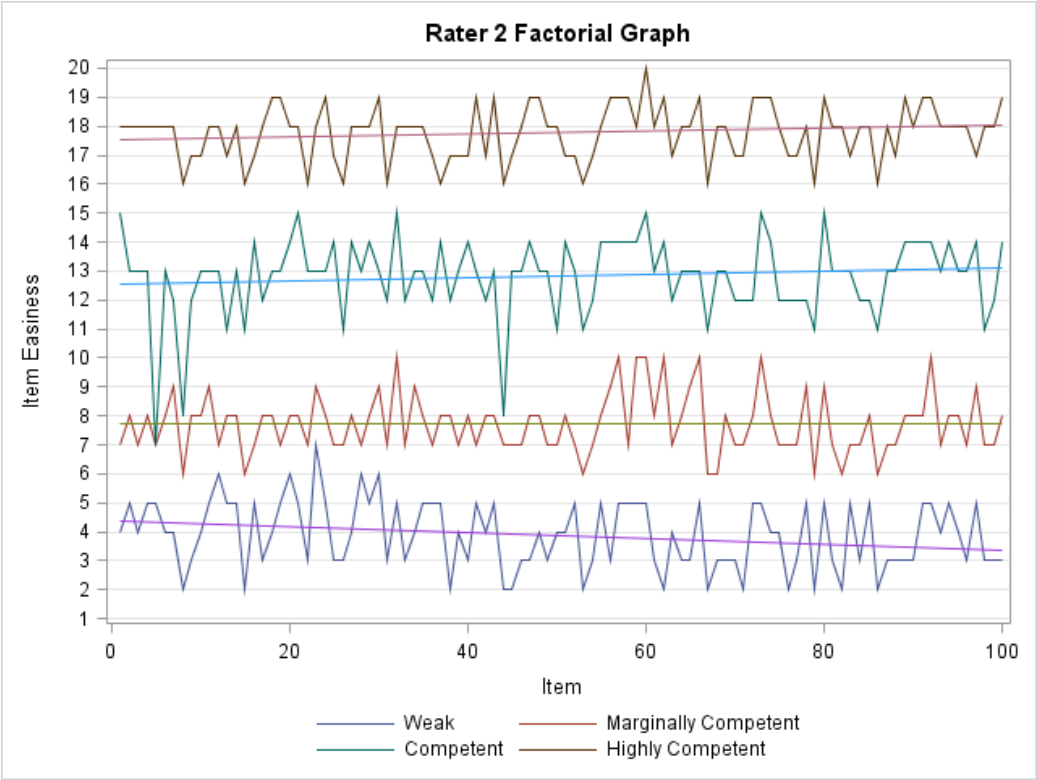
B.2 IIT Factorial Graphs For Excelsior College Nursing Exam.

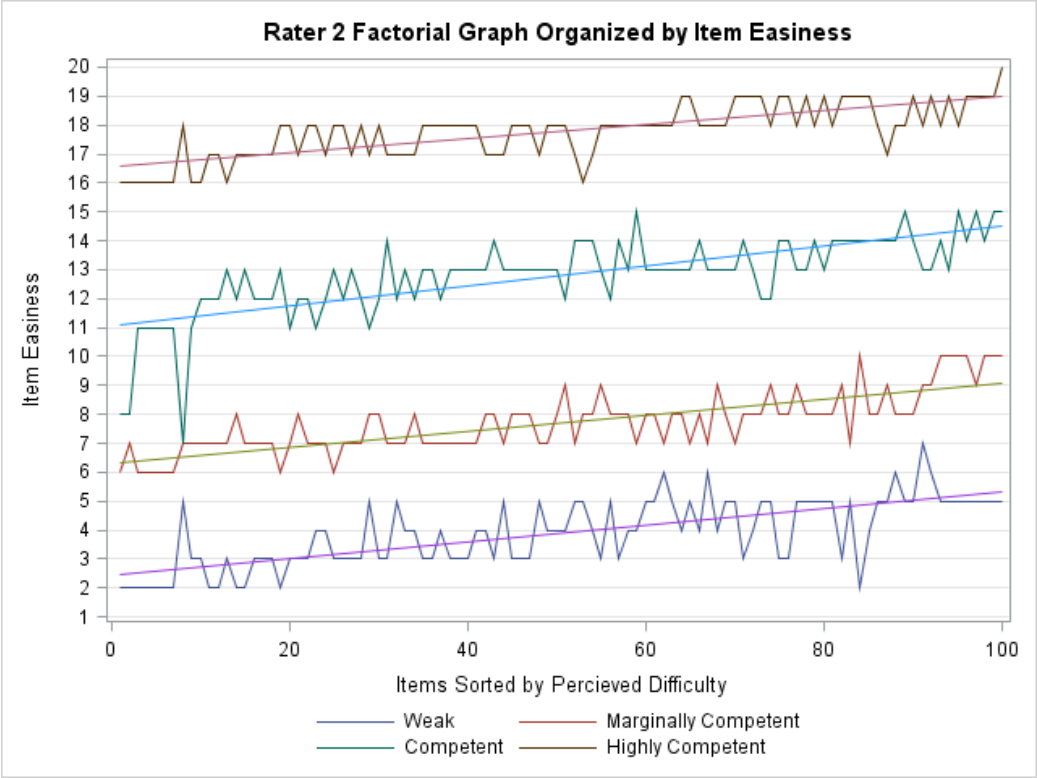
B.1.1 IIT Factorial Graph for Rater 1.



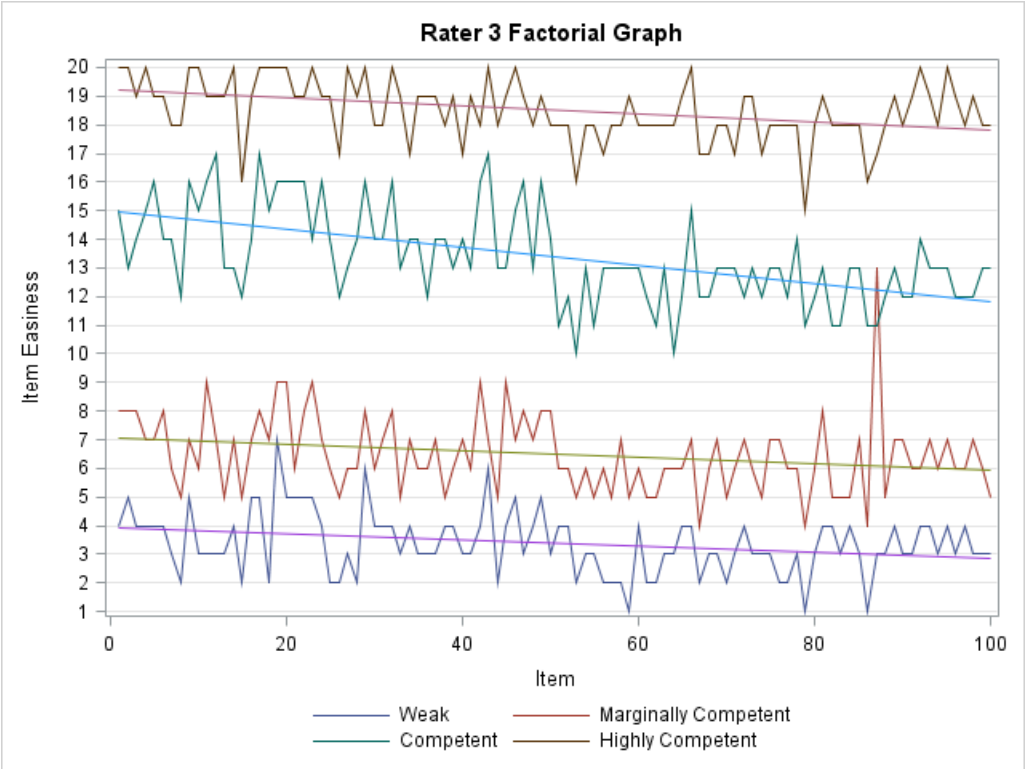


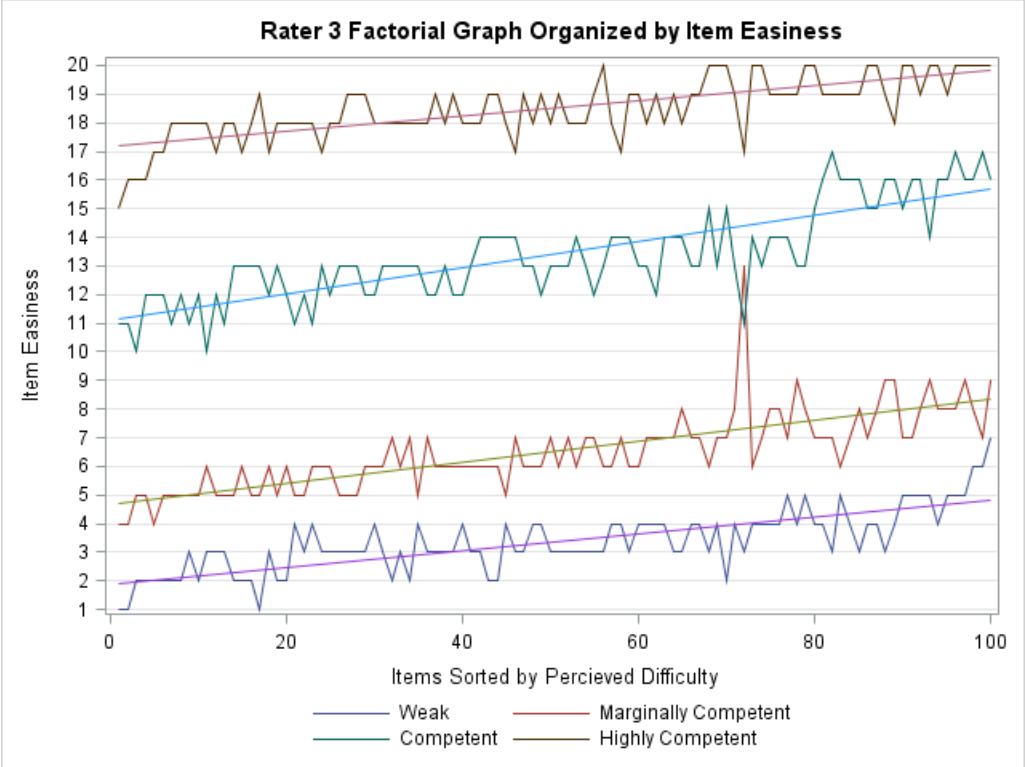
B.1.2 IIT Factorial Graph for Rater 2.



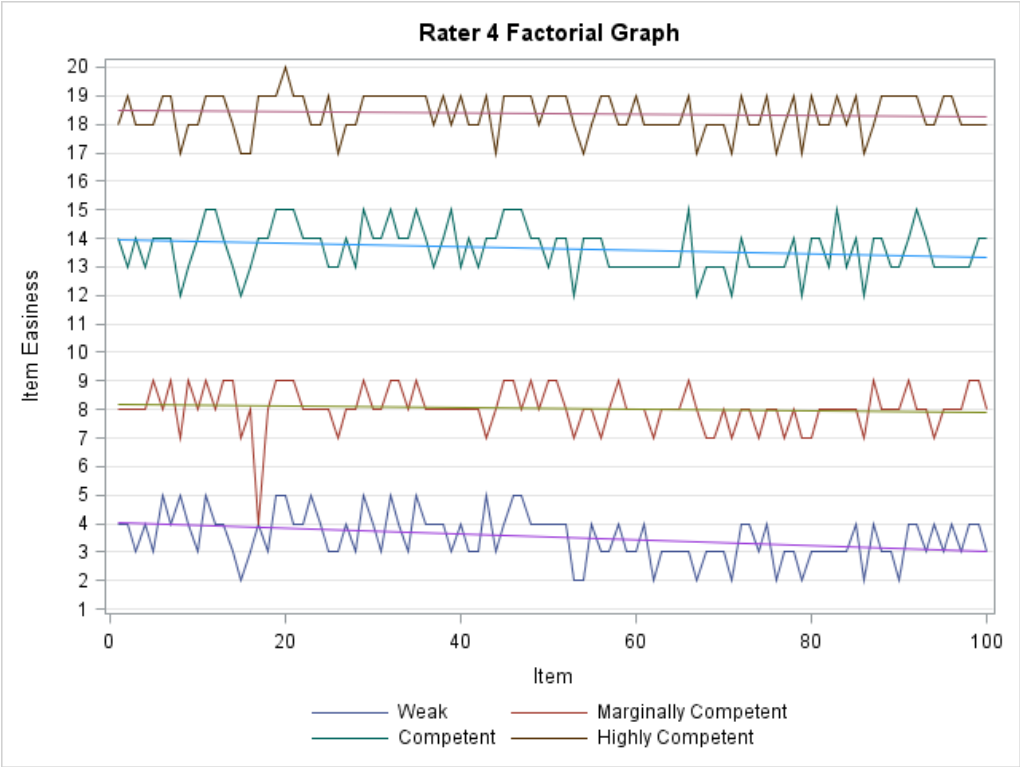


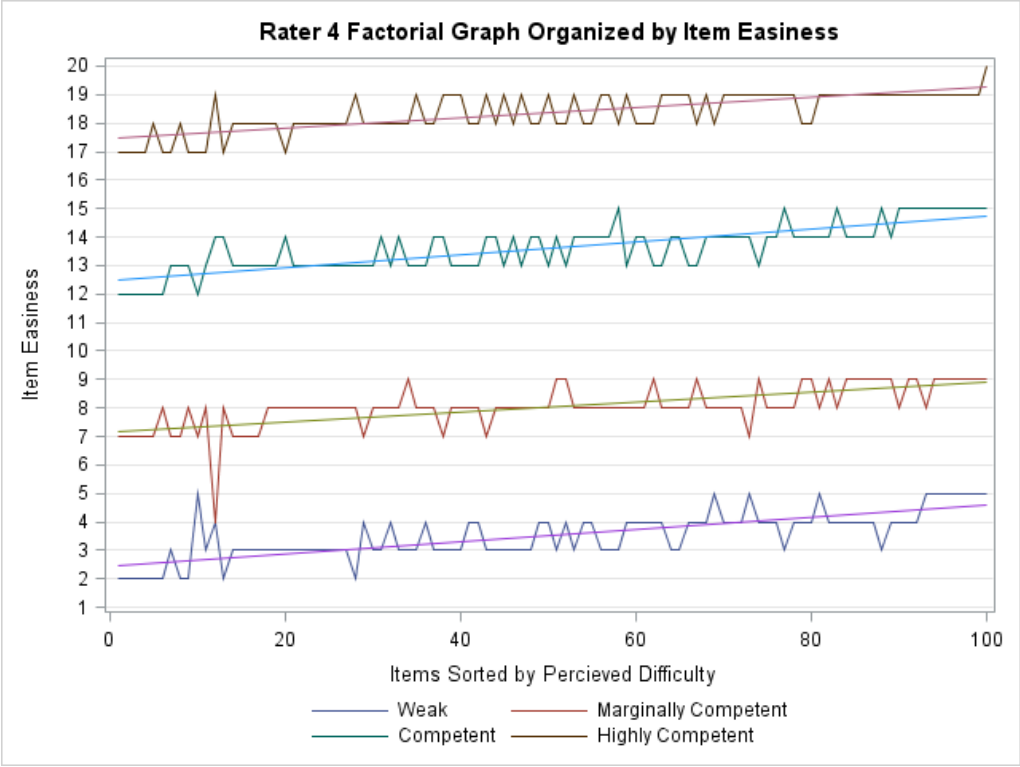
B.1.3 IIT Factorial Graph for Rater 3.



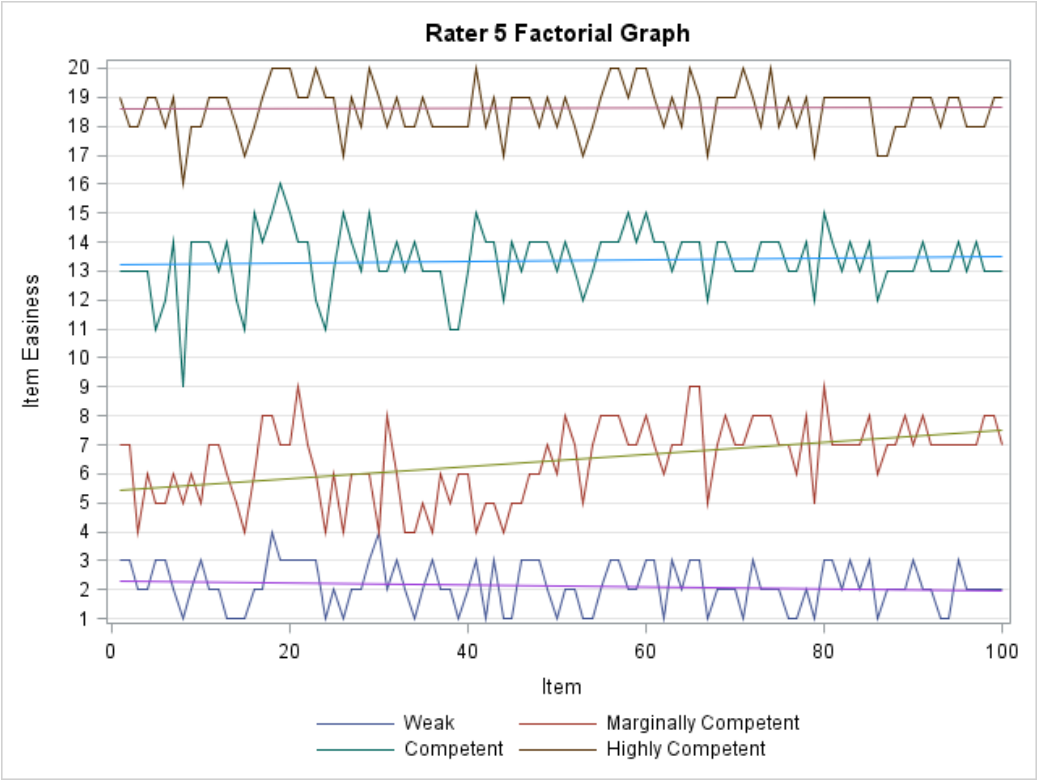


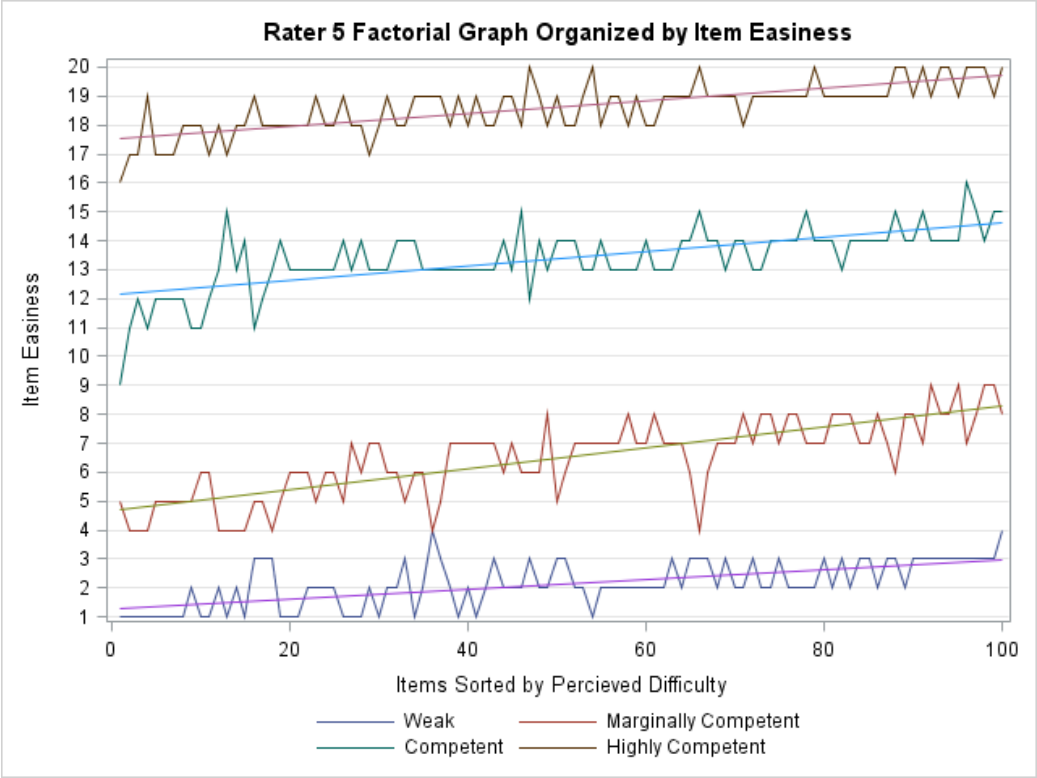
B.1.4 IIT Factorial Graph for Rater 4.



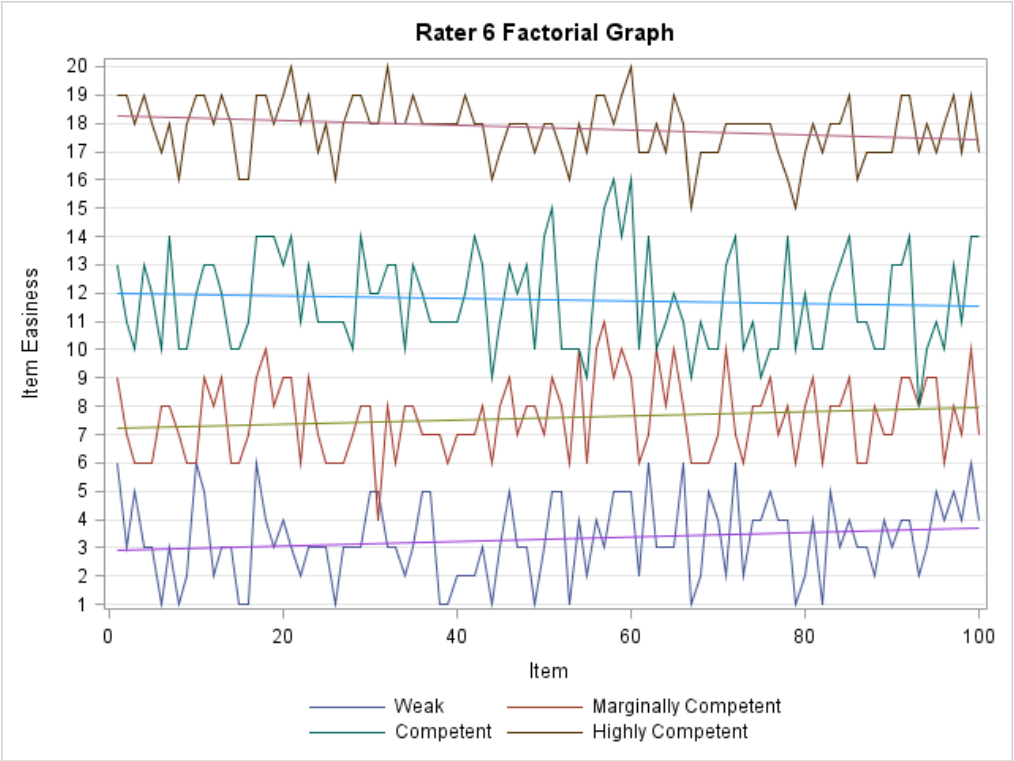


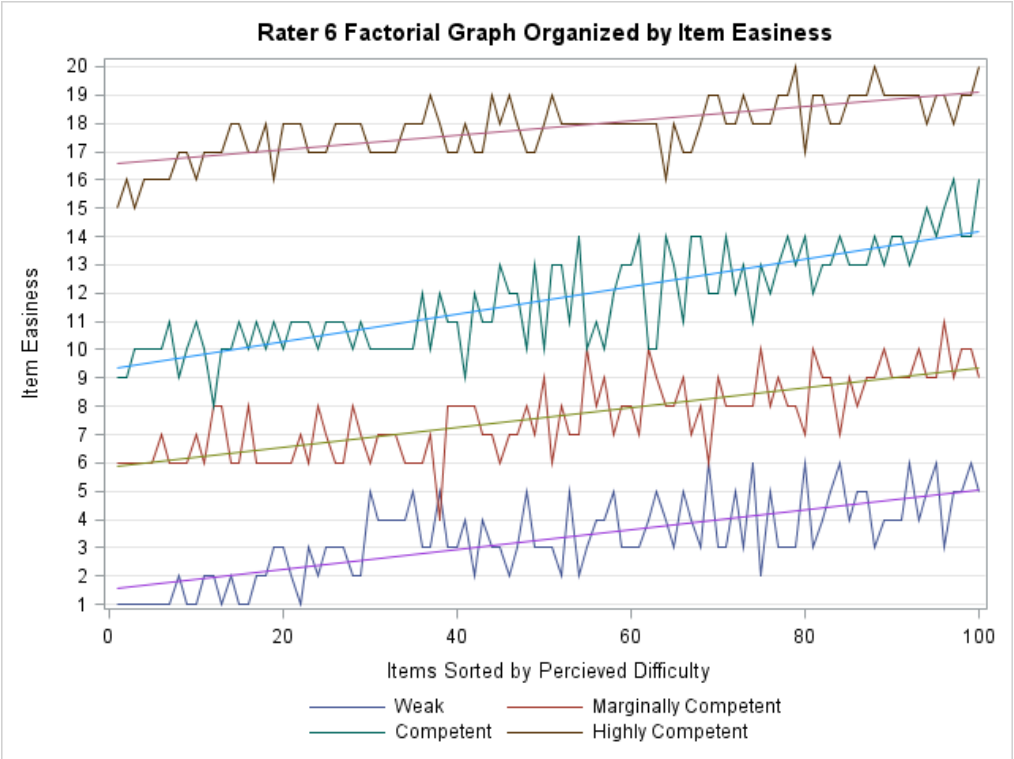
B.1.5 IIT Factorial Graph for Rater 5.



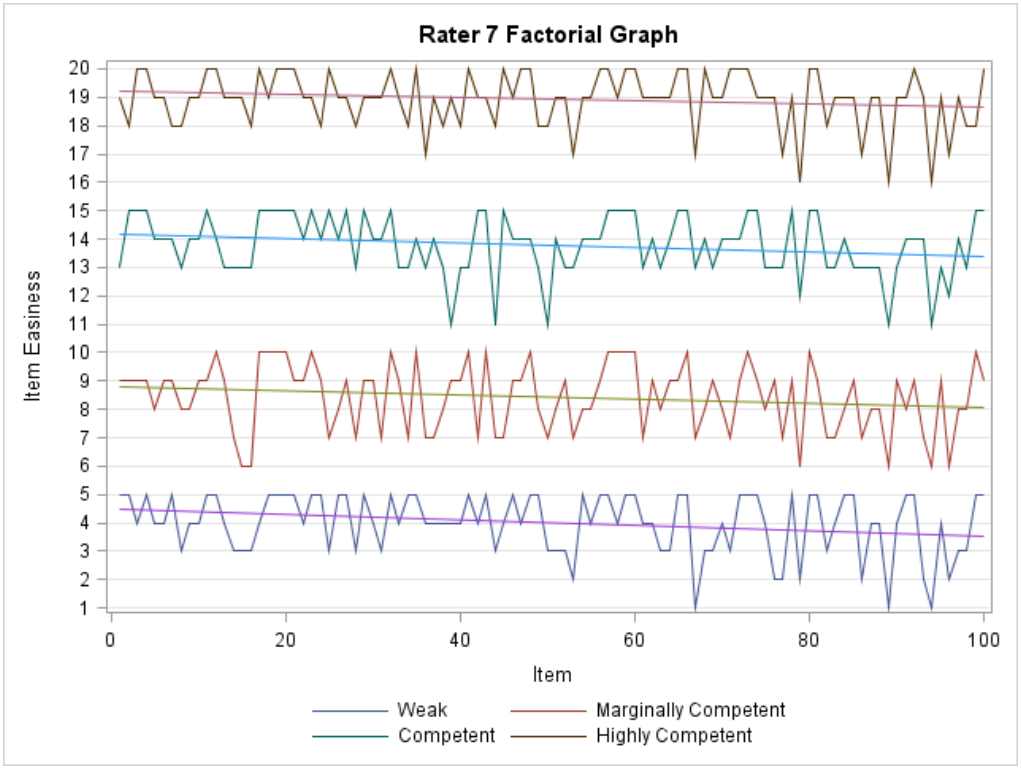


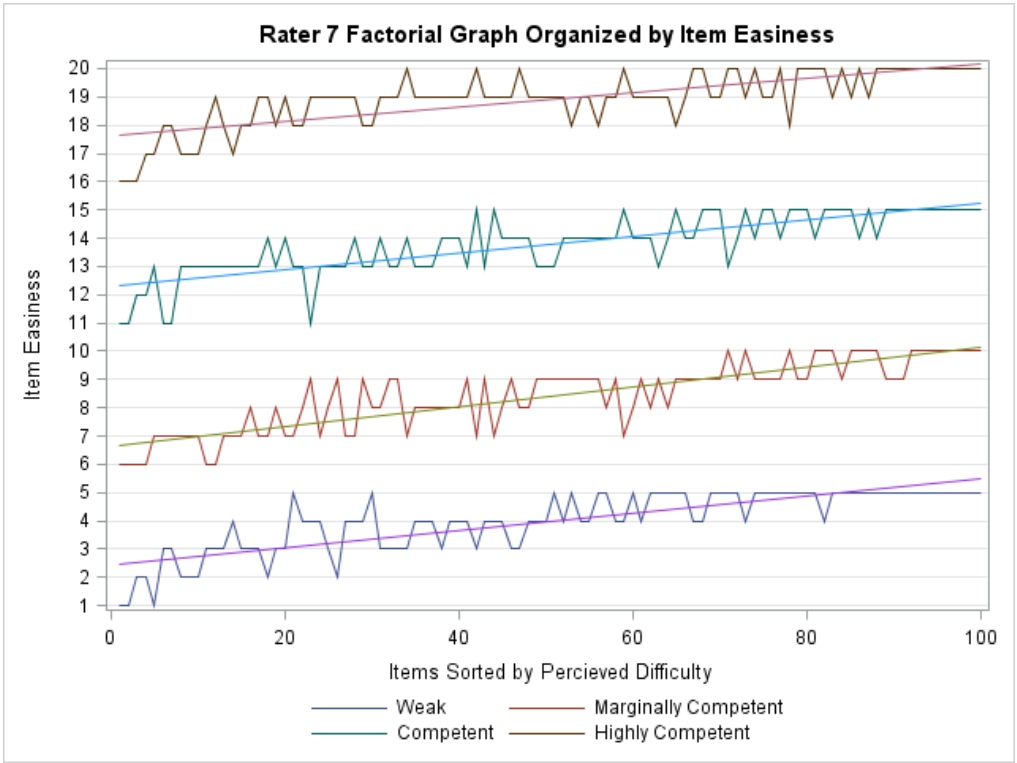
B.1.6 IIT Factorial Graph for Rater 6.



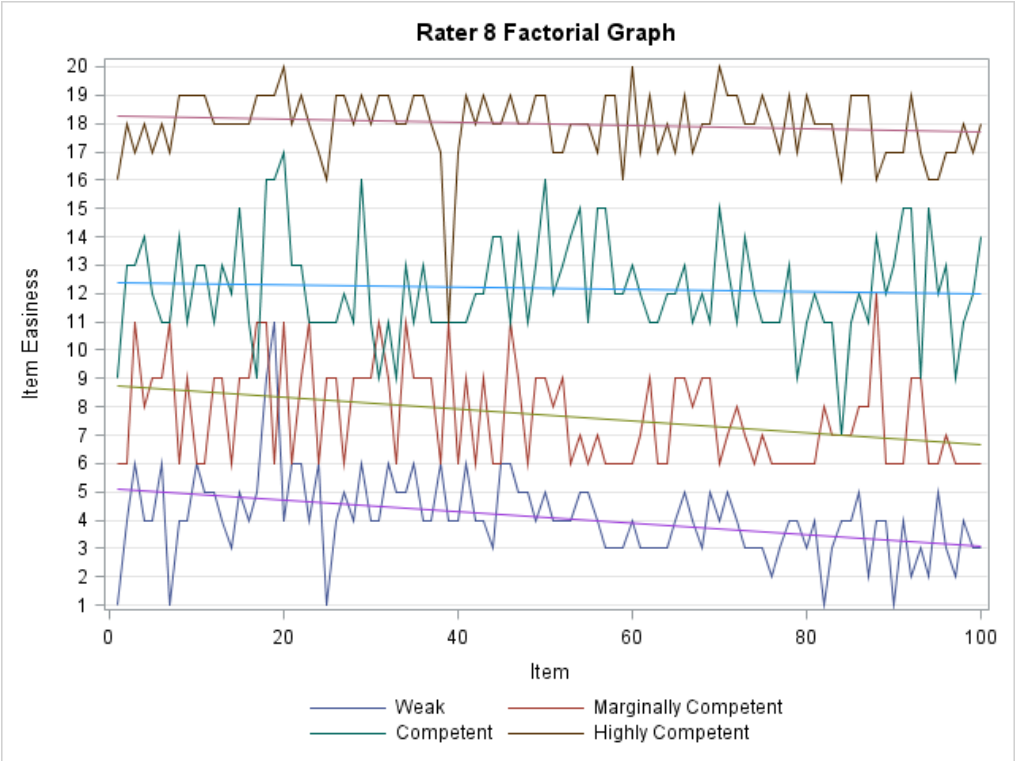


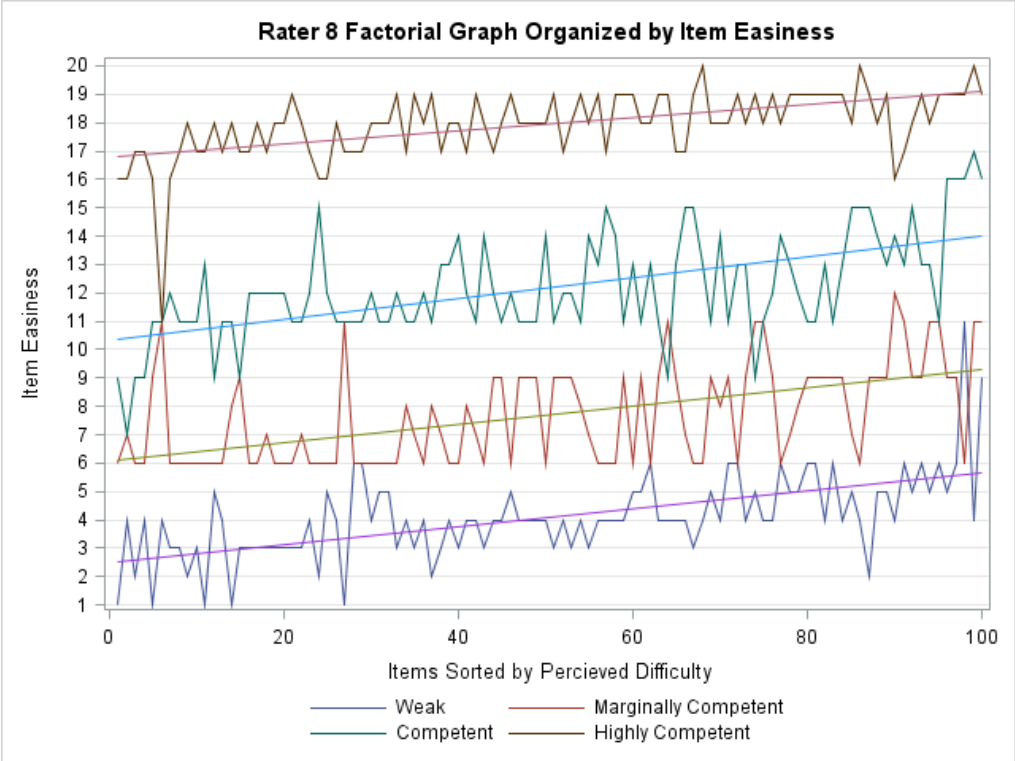
B.1.7 IIT Factorial Graph for Rater 7.



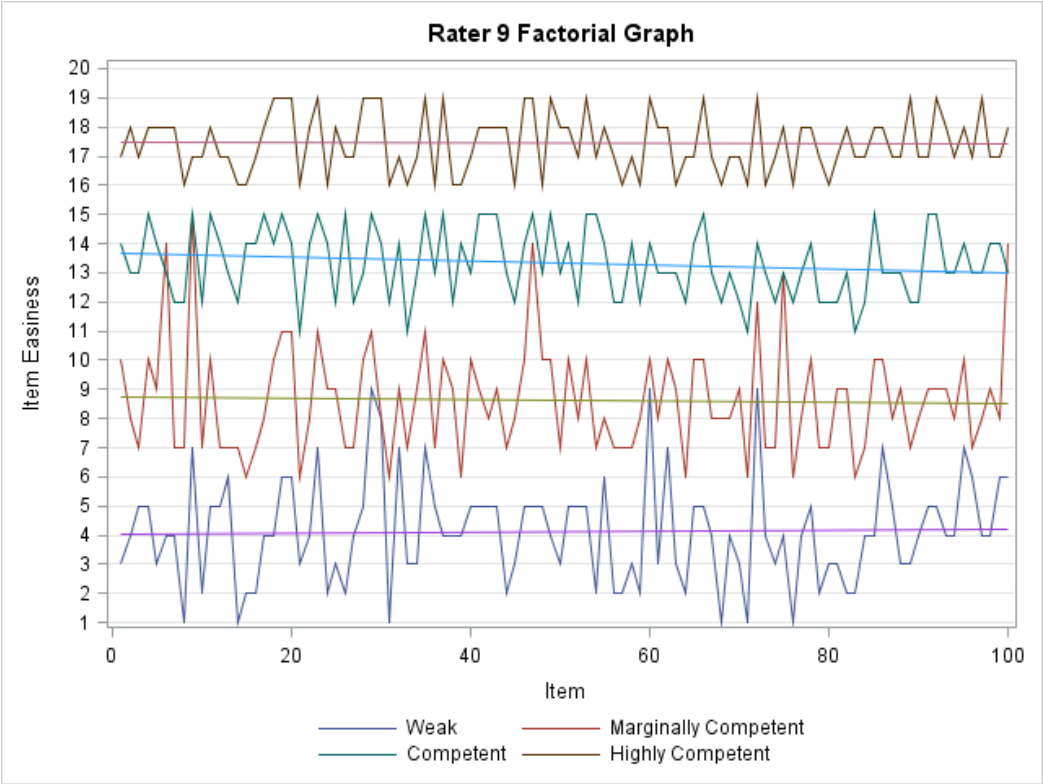


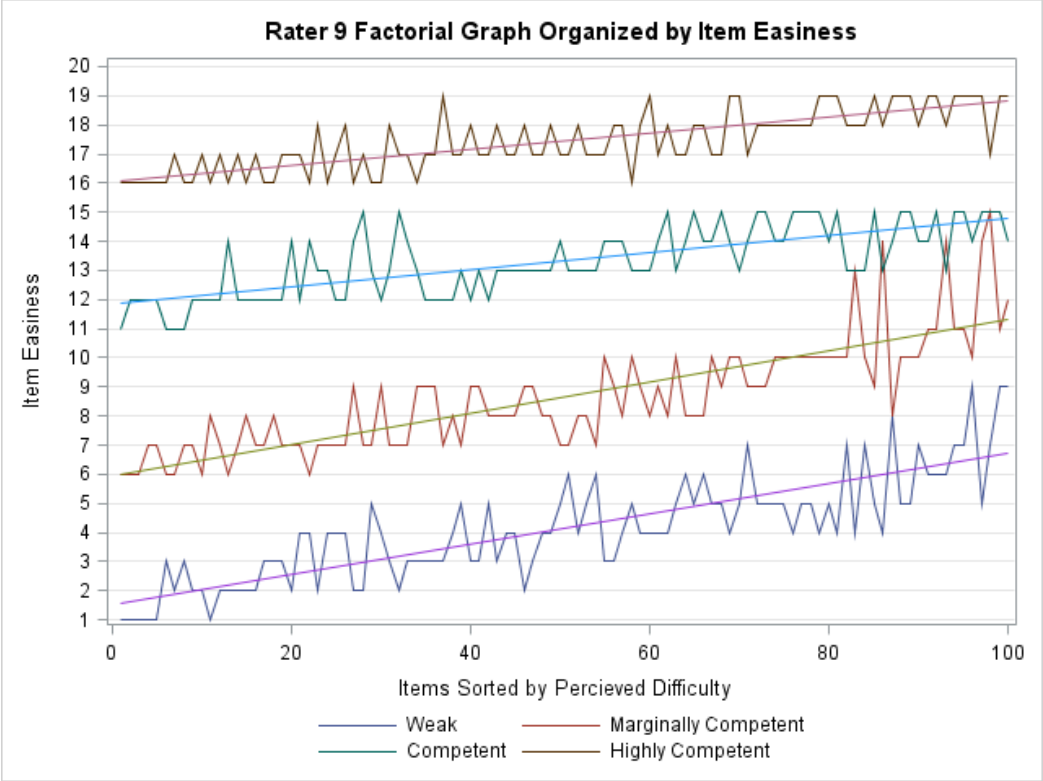
B.1.8 IIT Factorial Graph for Rater 8.



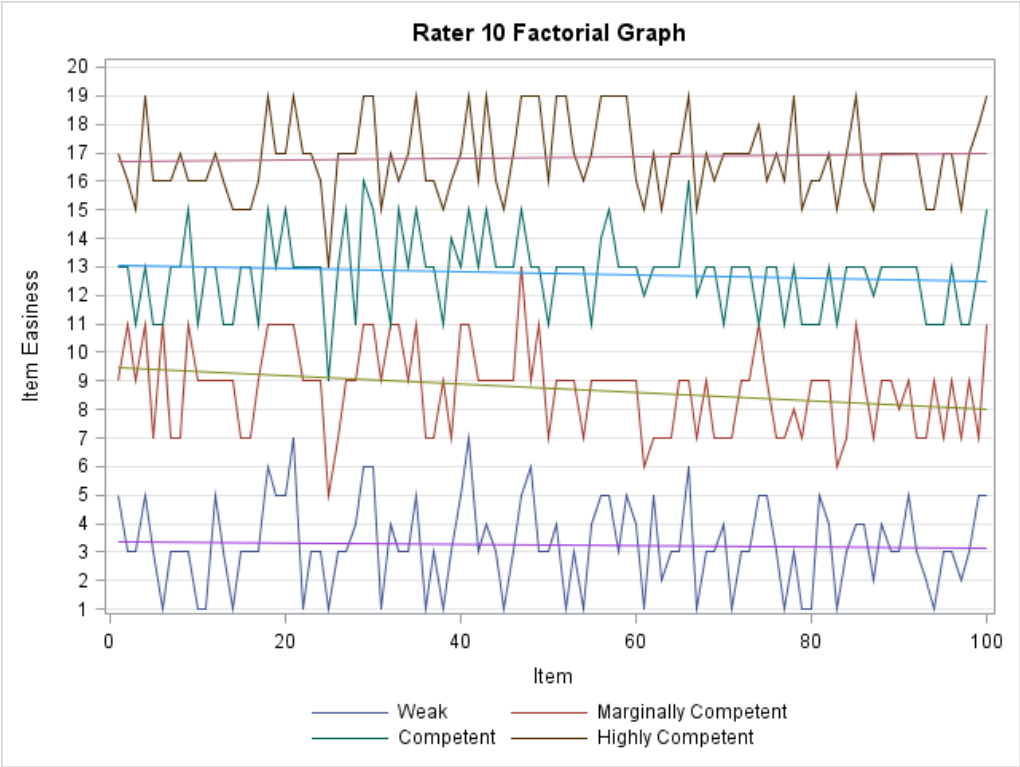


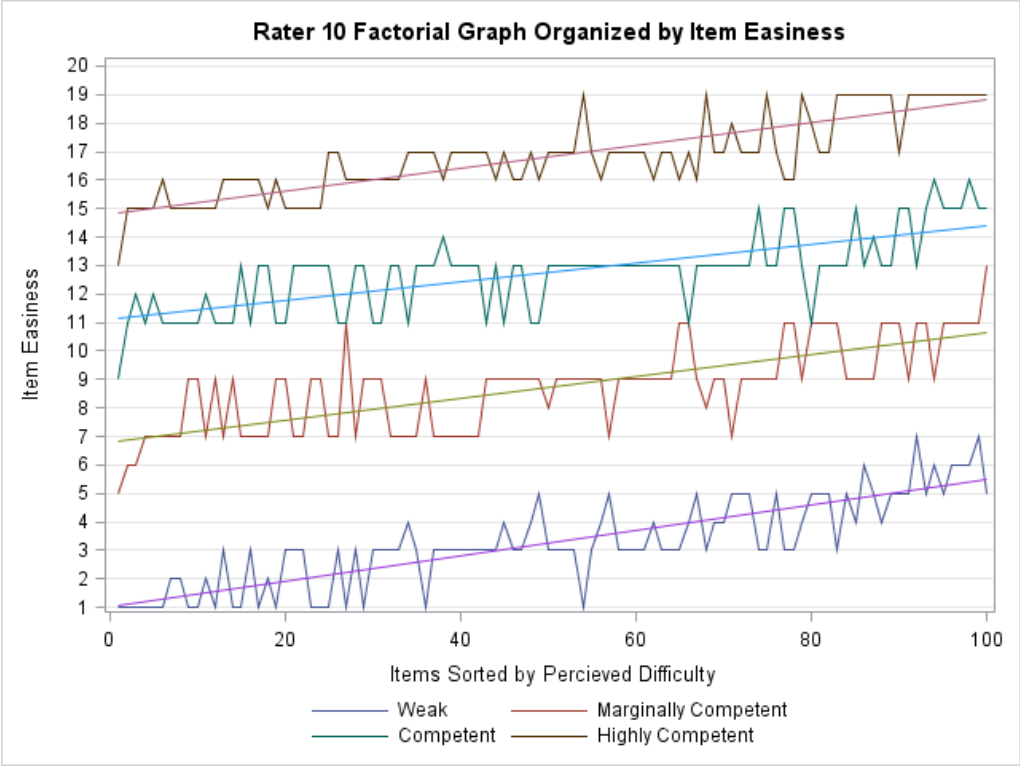
B.1.9 IIT Factorial Graph for Rater 9.



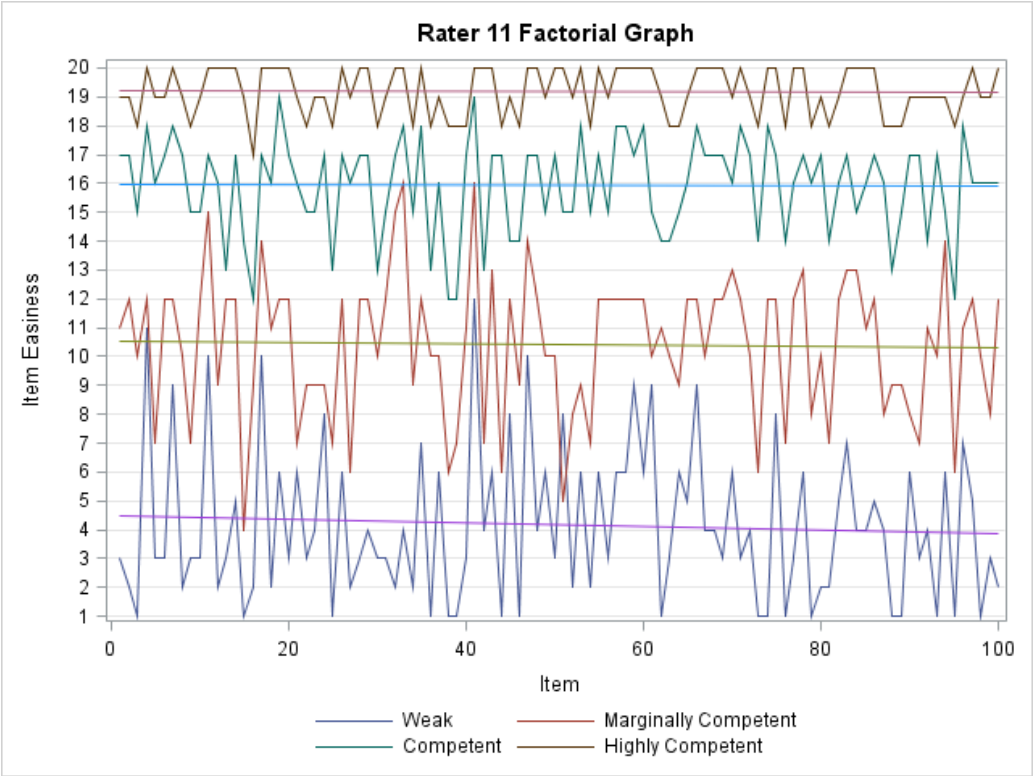


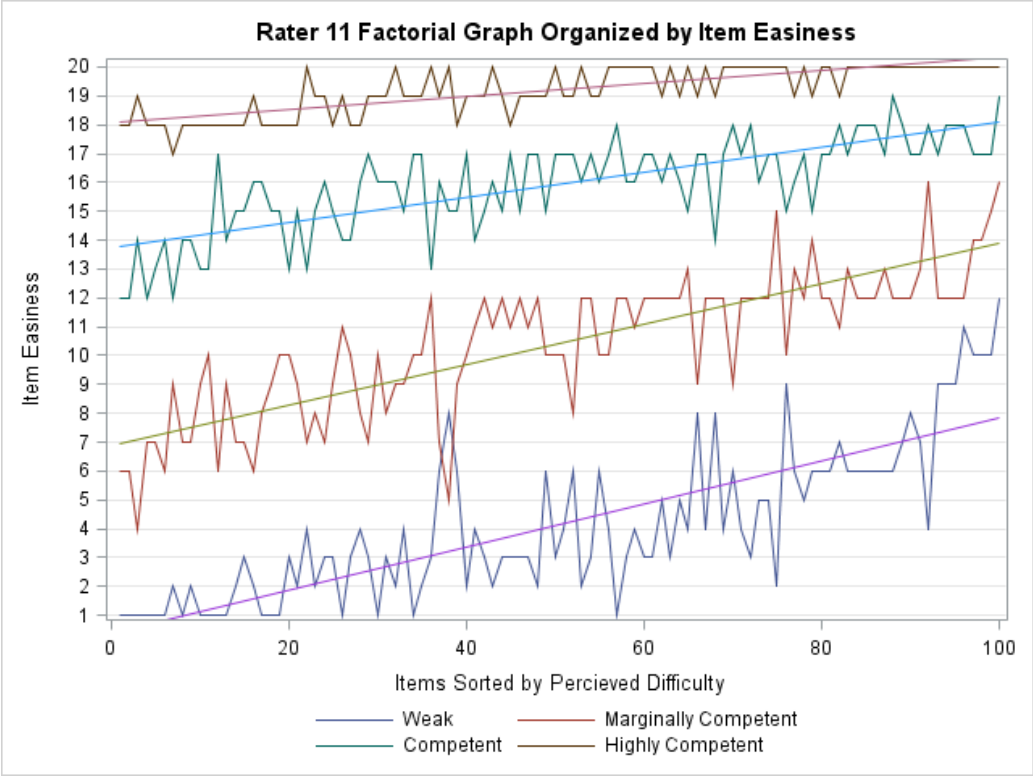
B.1.10 IIT Factorial Graph for Rater 10.



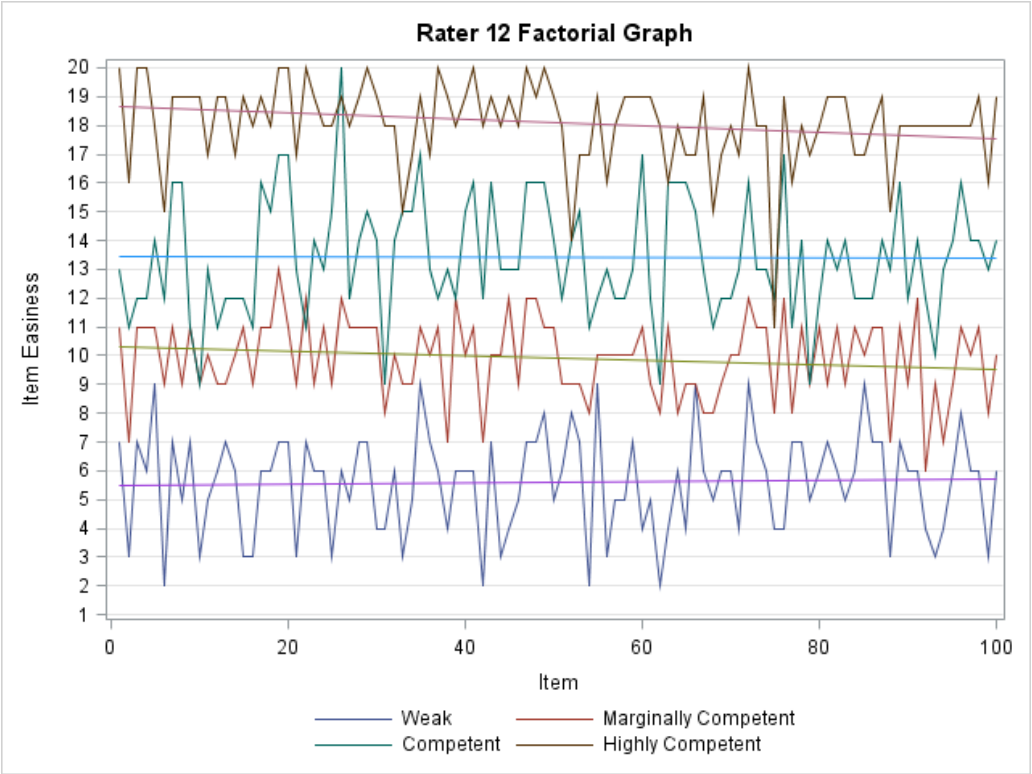


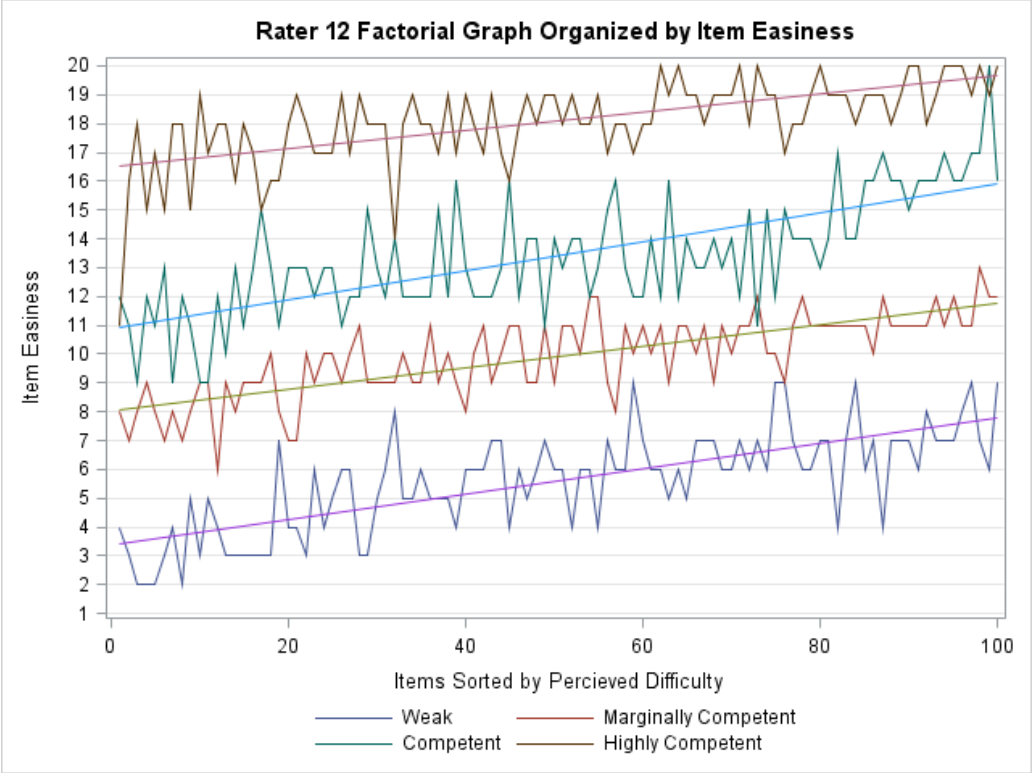
B.1.11 IIT Factorial Graph for Rater 11.





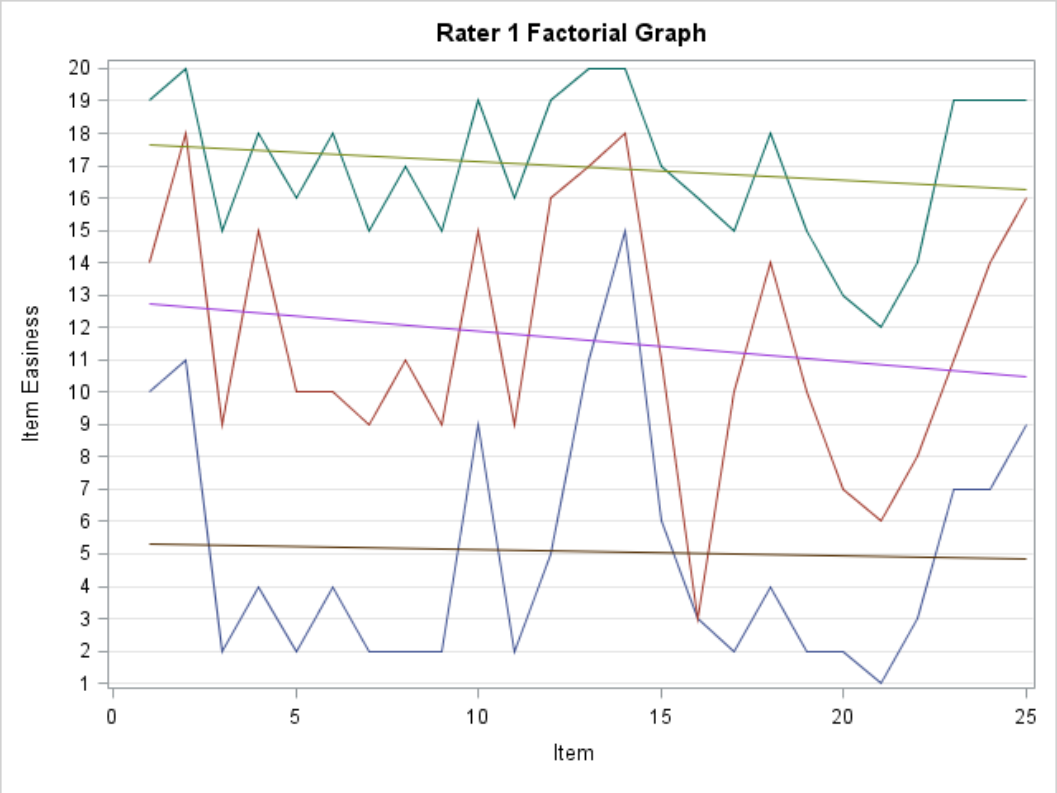
B.1.12 IIT Factorial Graph for Rater 11.

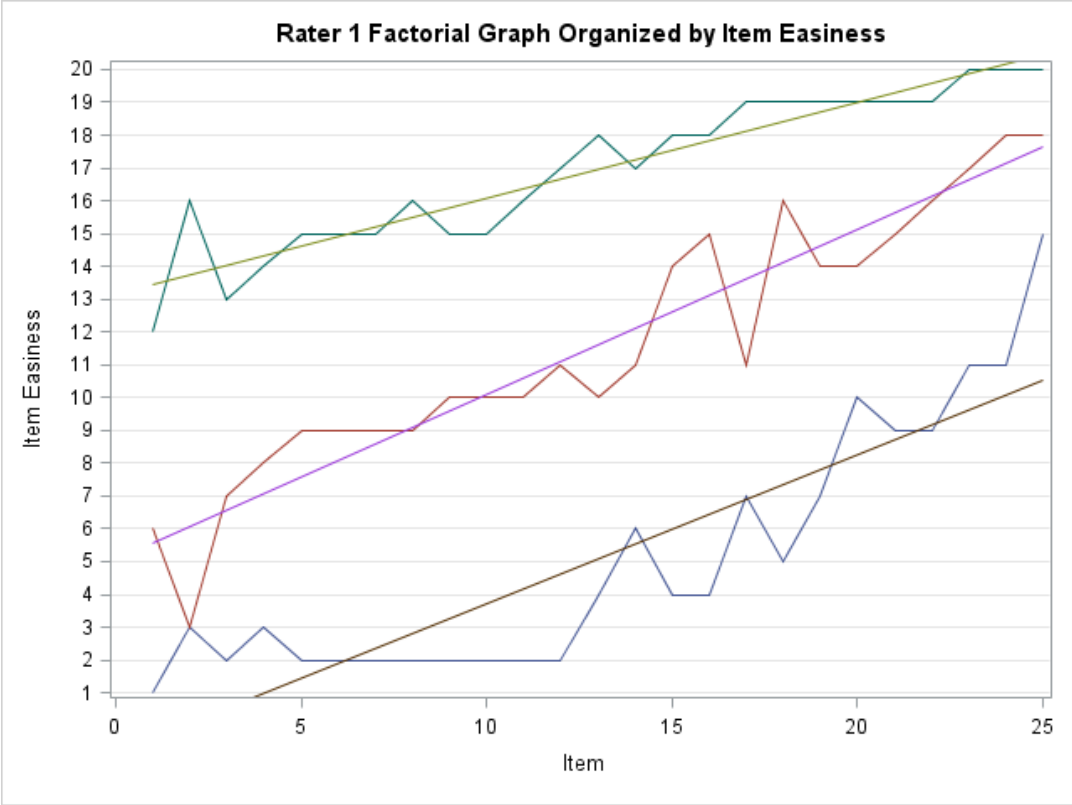




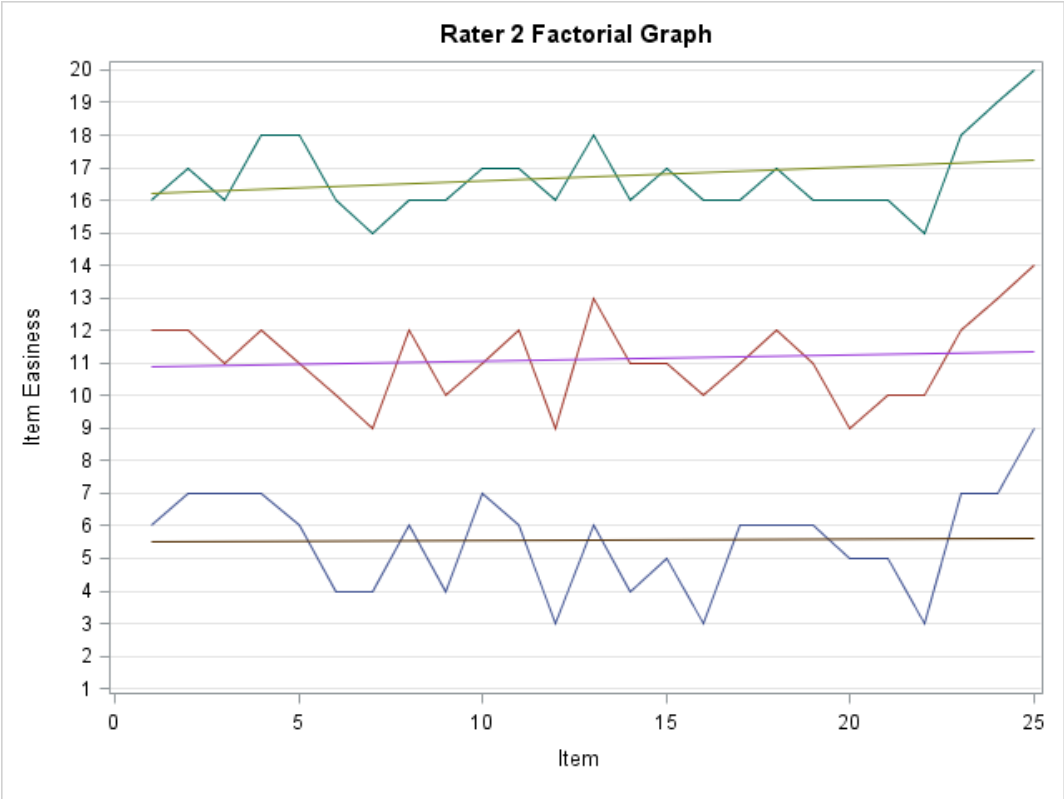
B.3 IIT Factorial Graphs For TIMSS Exam

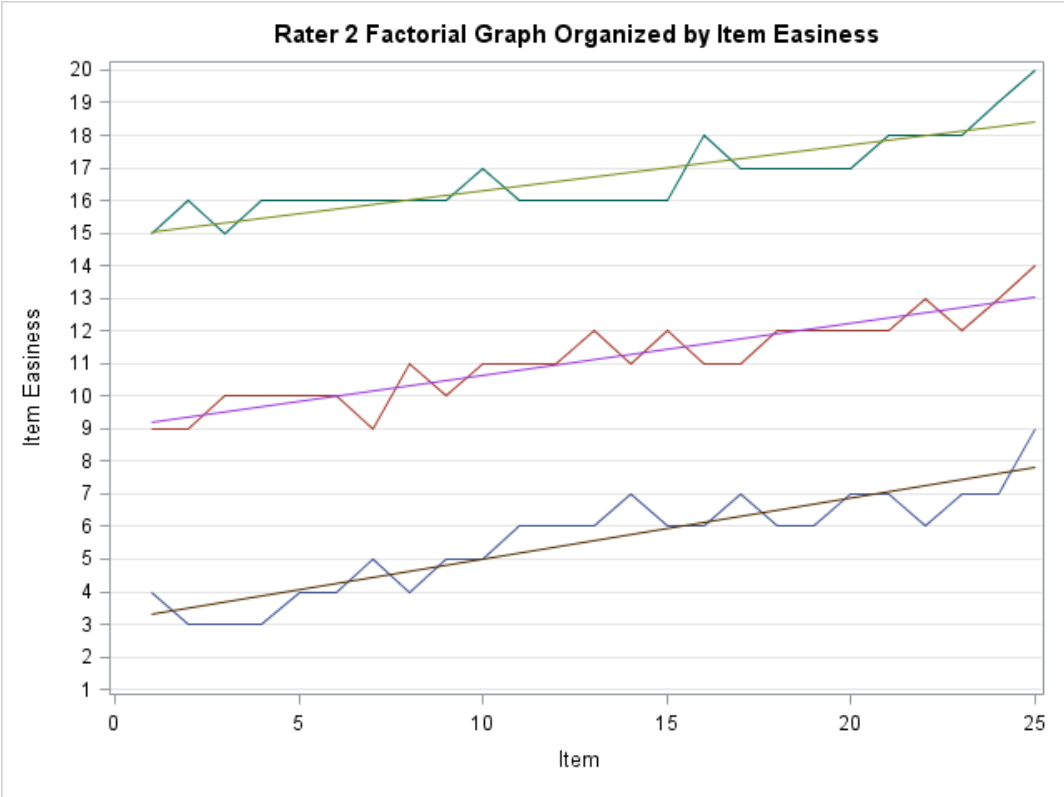
B.3.1 IIT Factorial Graph for Rater 1.



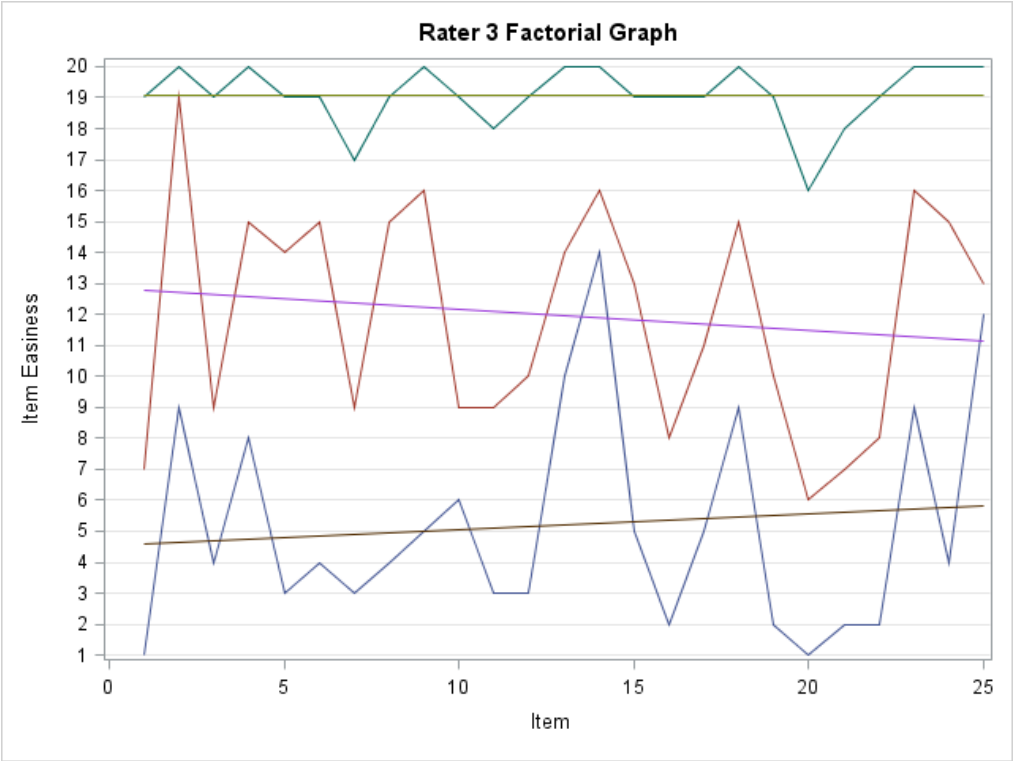

Rater 1 Factorial Graph Organized by Item Easiness
Item Easiness
Item
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
5
10
15
20
25

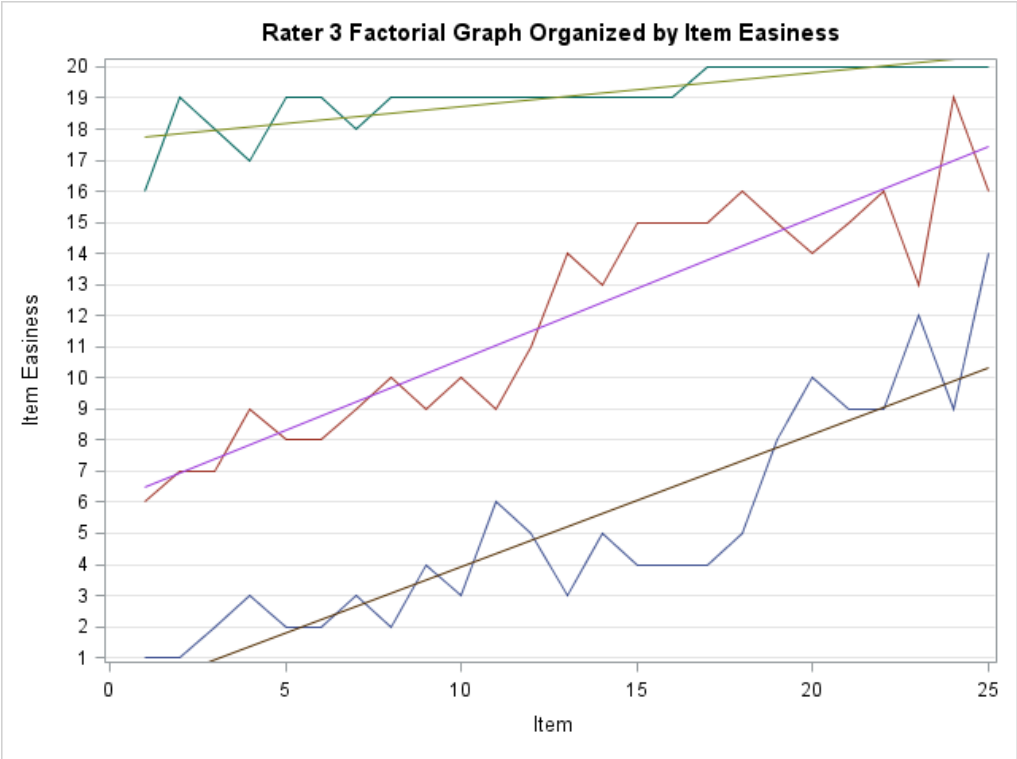
B.3.2 IIT Factorial Graph for Rater 2.



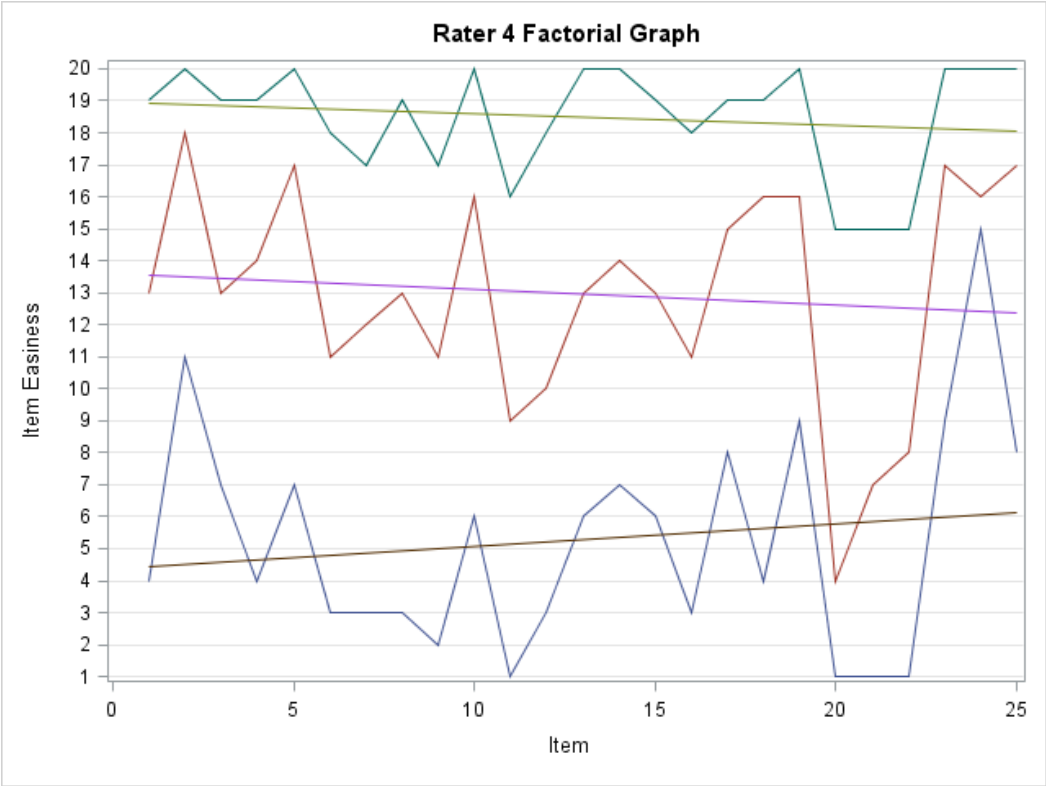


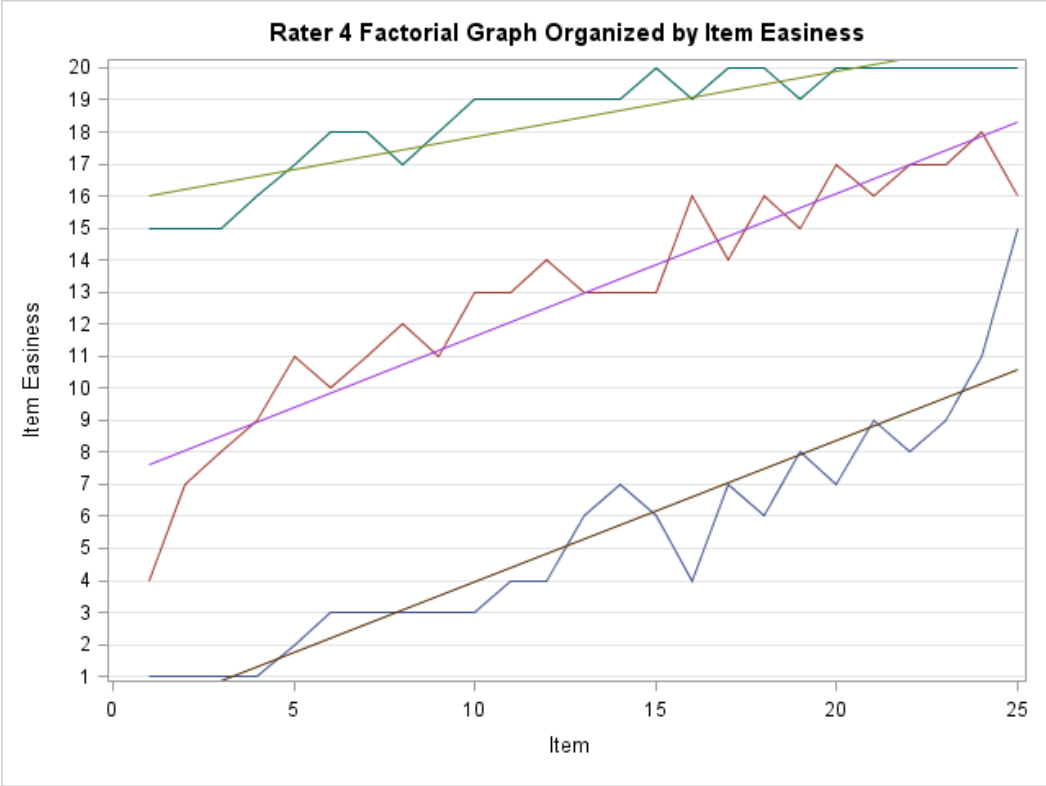
B.3.3 IIT Factorial Graph for Rater 3.



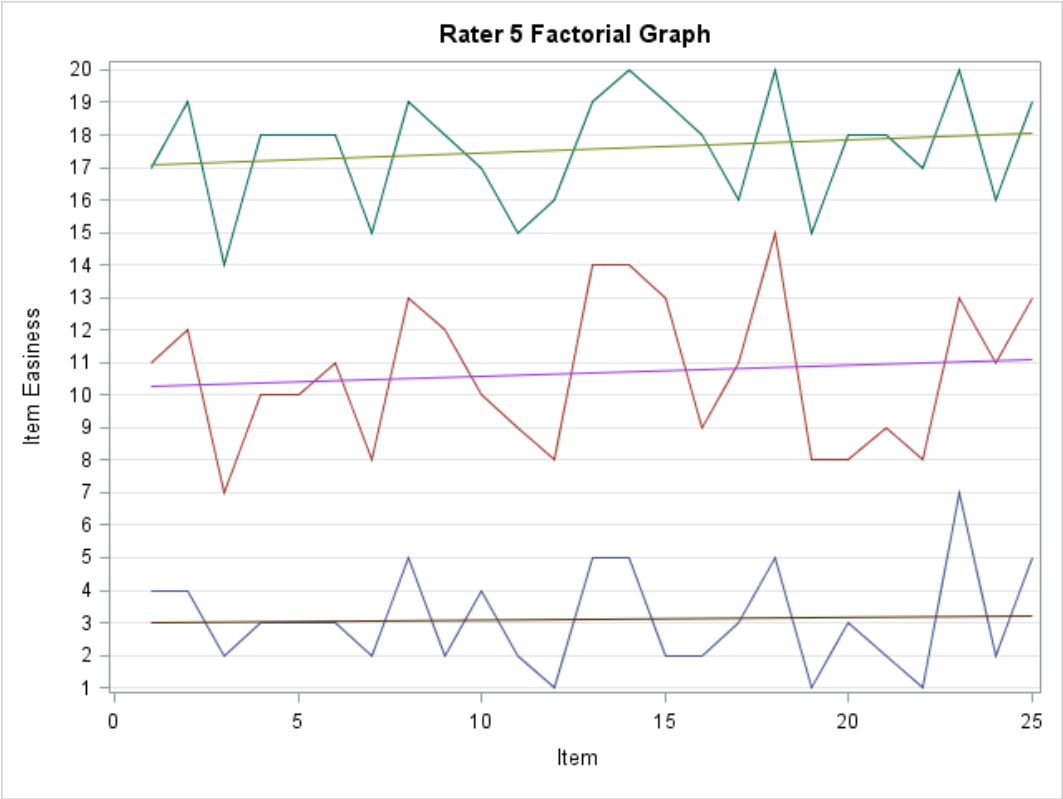


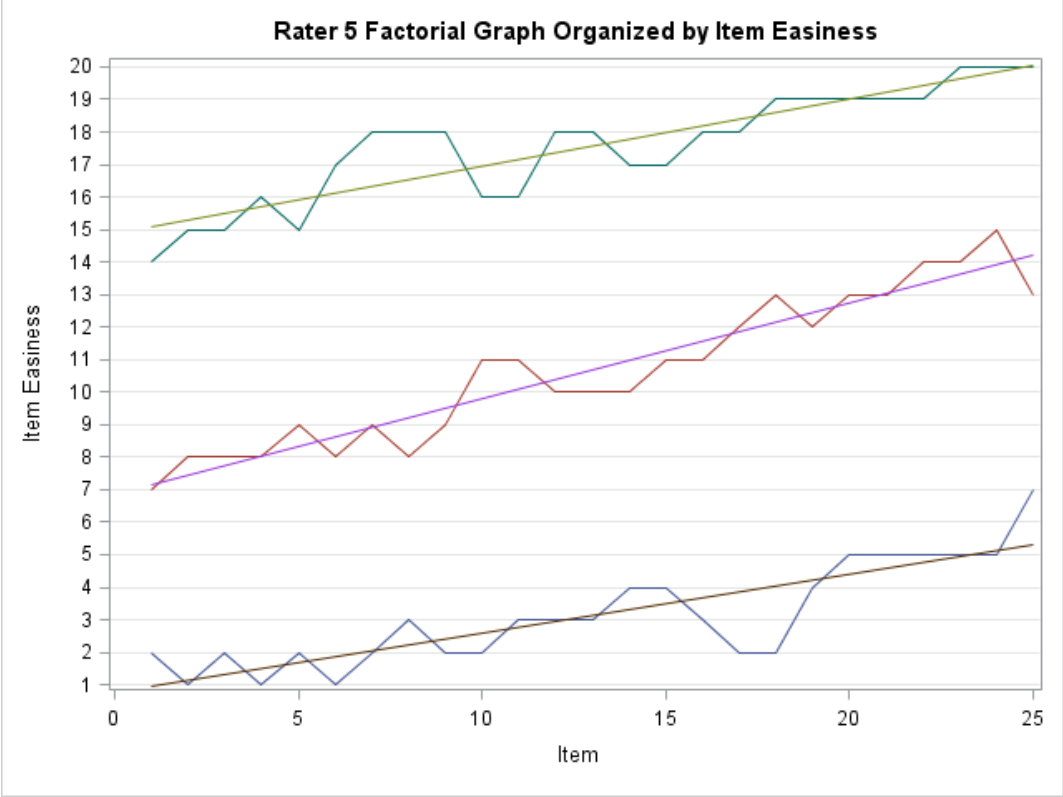
B.3.4 IIT Factorial Graph for Rater 4.



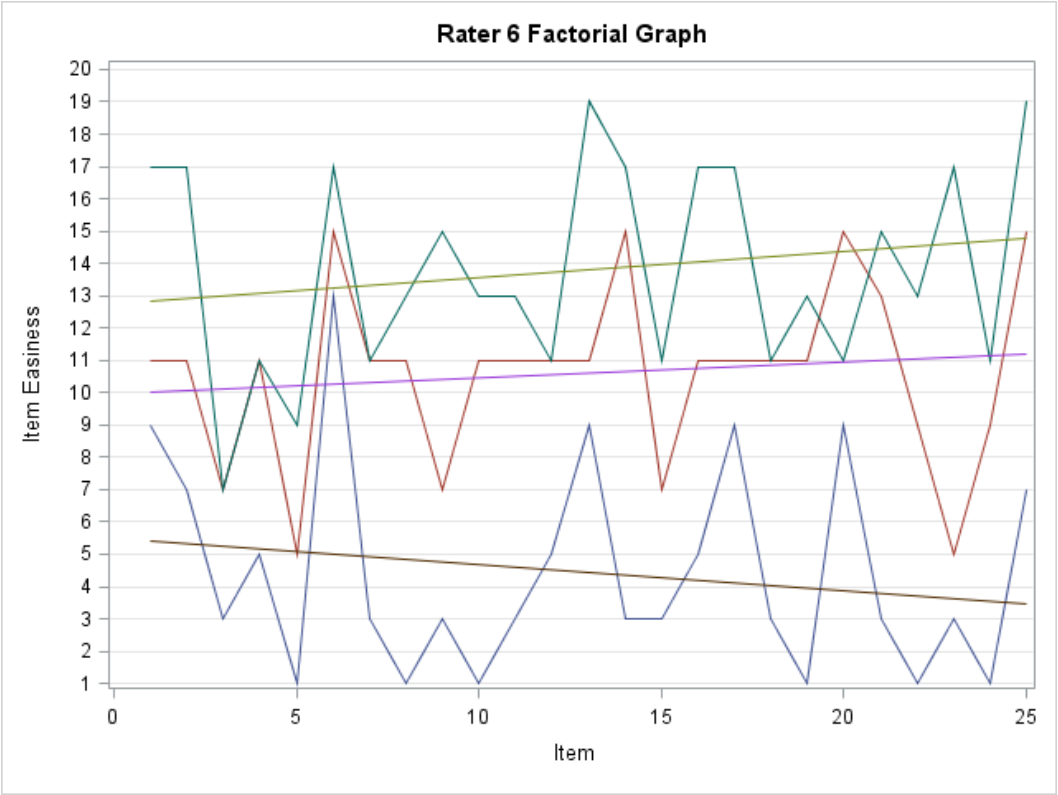


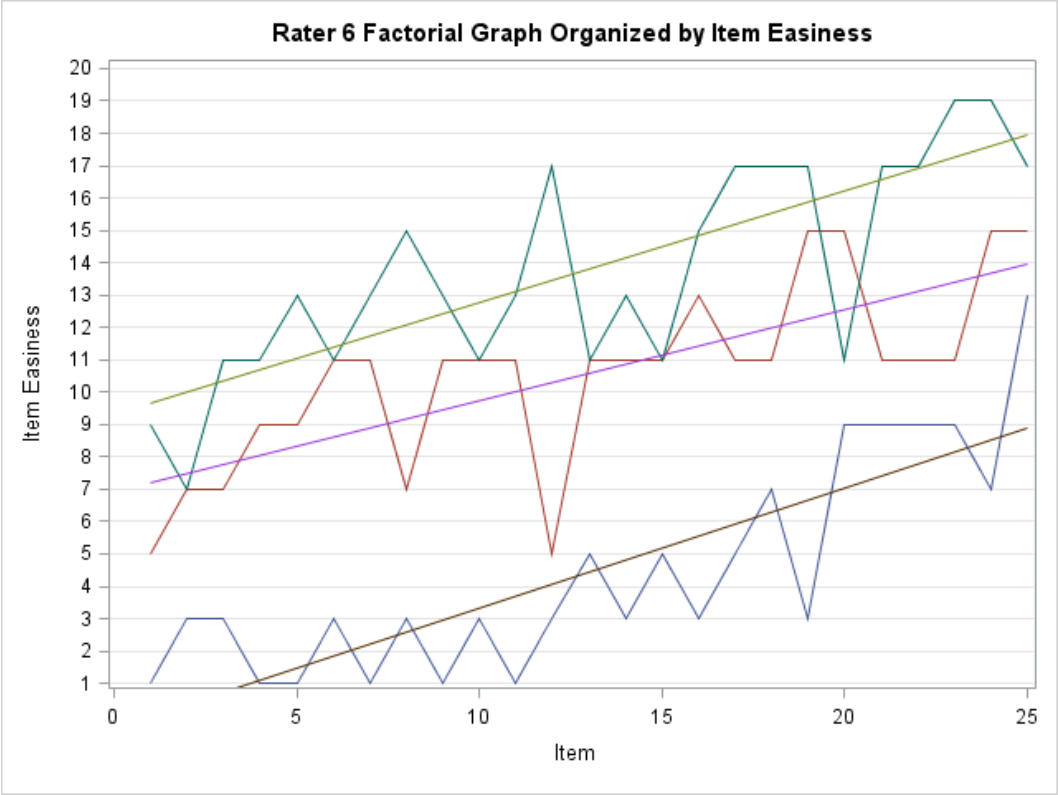
B.3.5 IIT Factorial Graph for Rater 5.



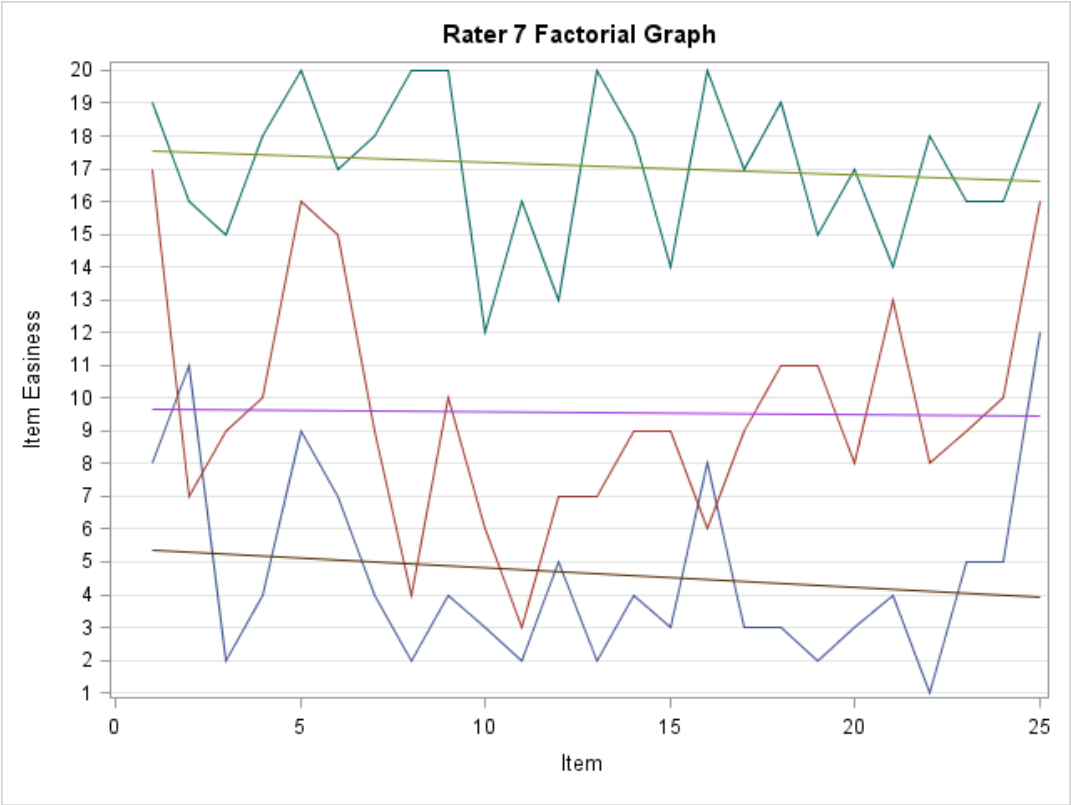


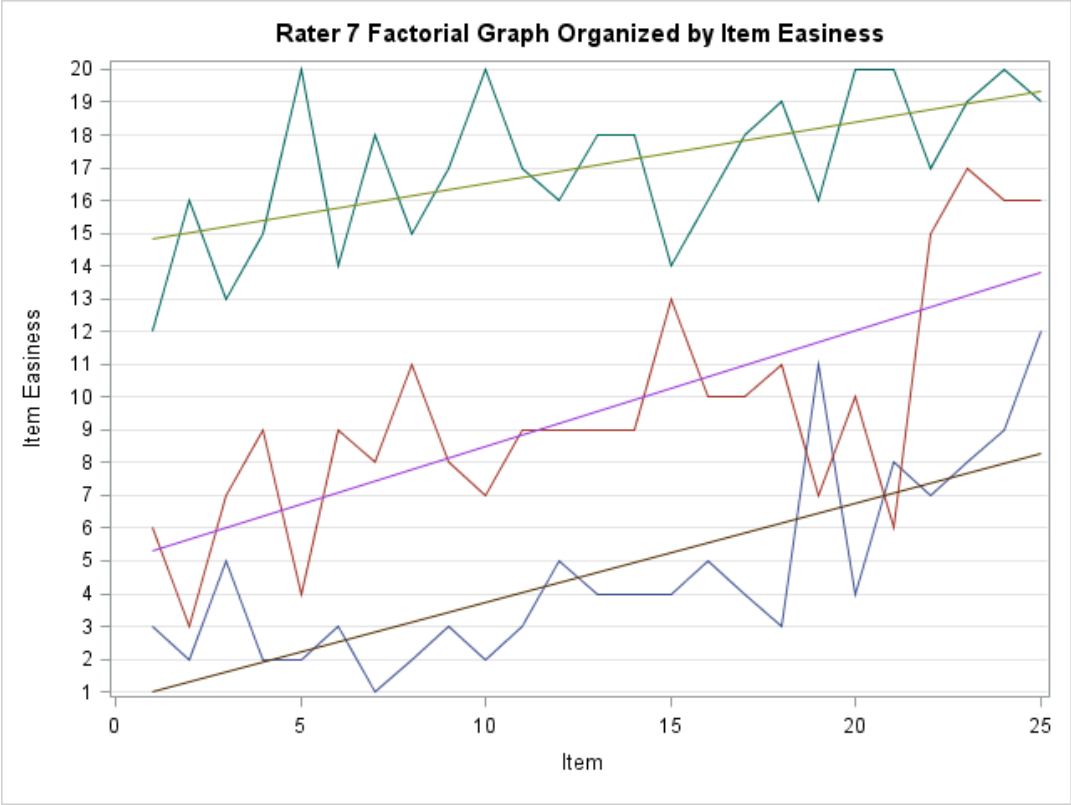
B.3.6 IIT Factorial Graph for Rater 6.



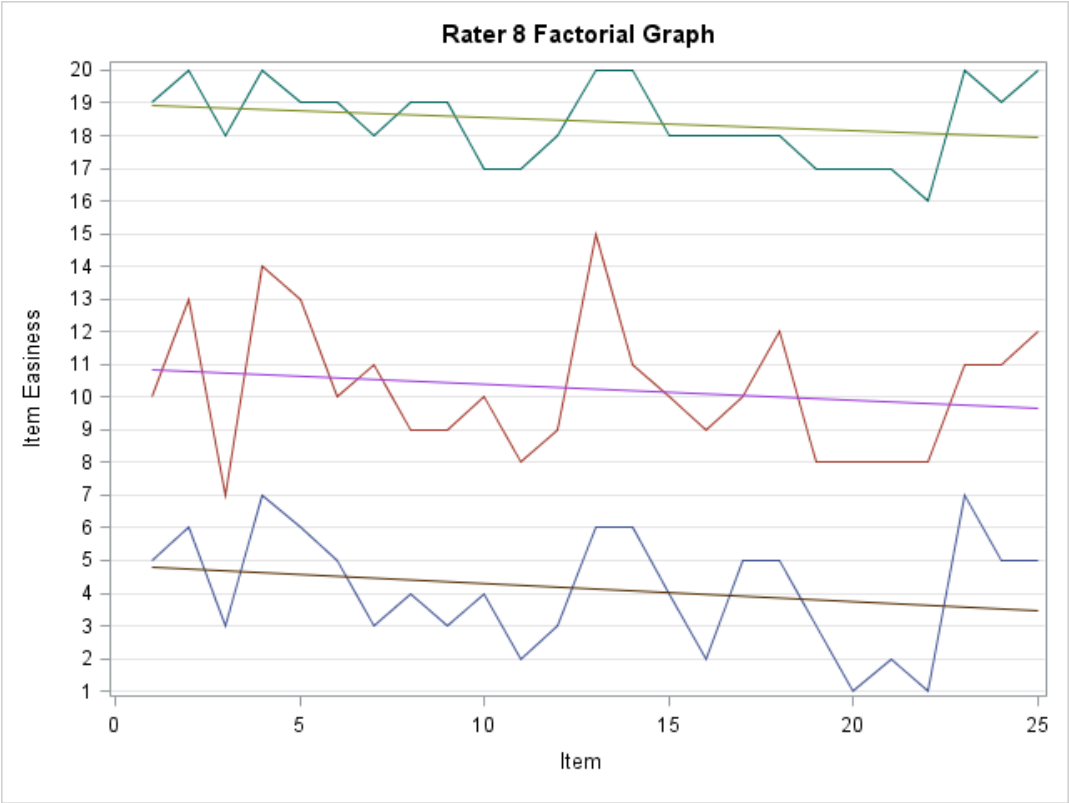


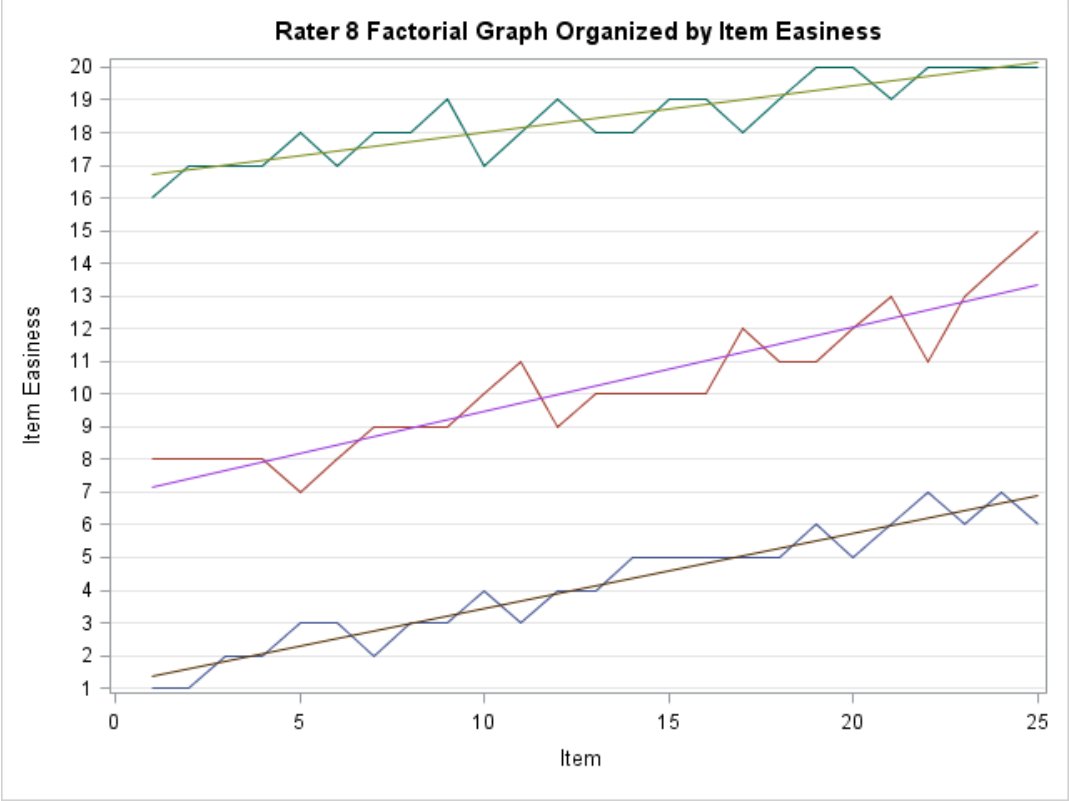
B.3.7 IIT Factorial Graph for Rater 7.



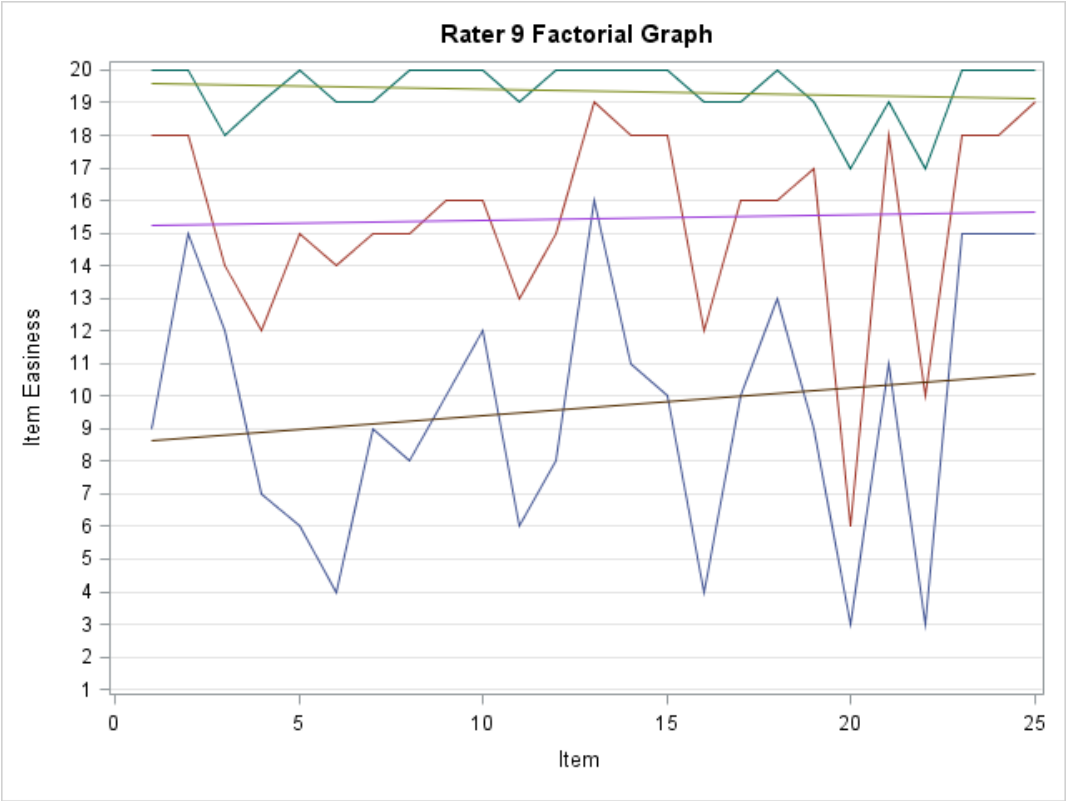


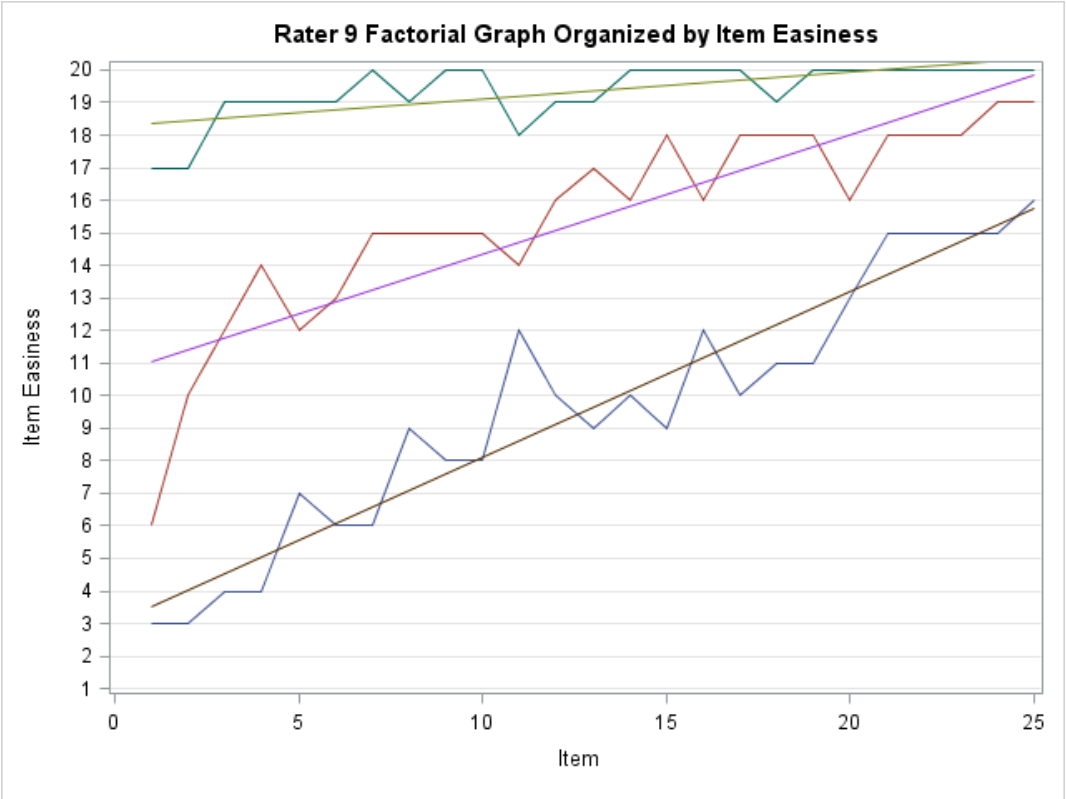
B.3.8 IIT Factorial Graph for Rater 8.



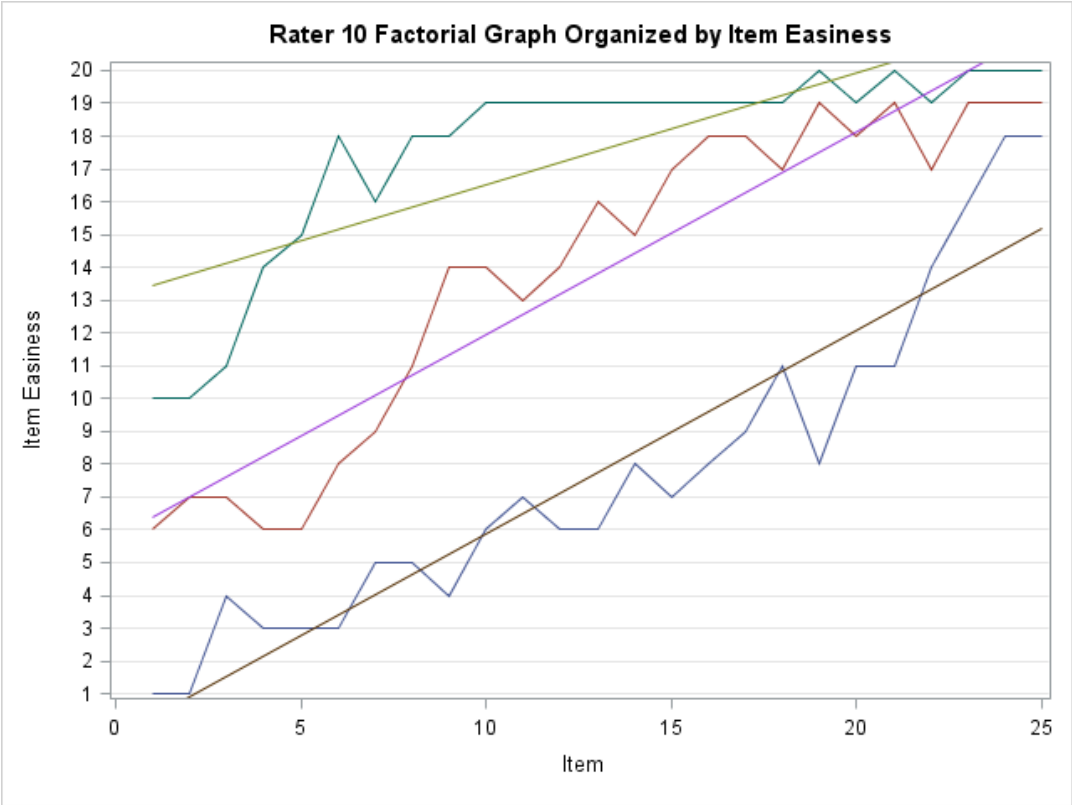


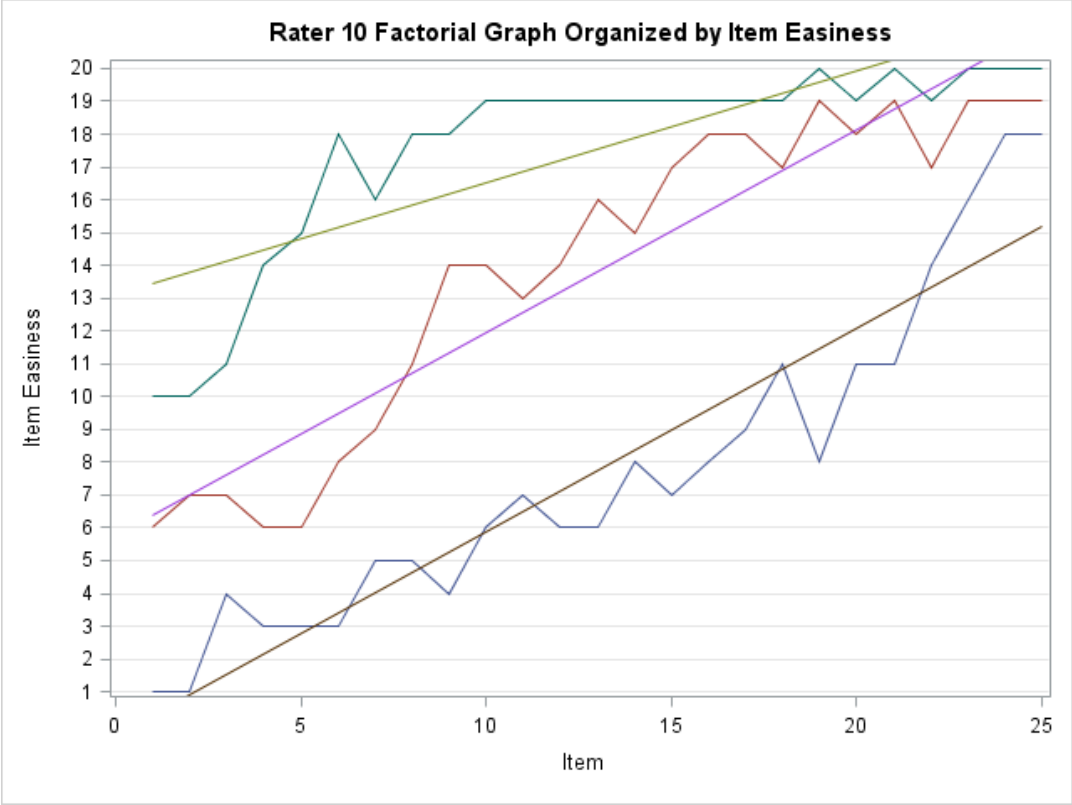
B.3.9 IIT Factorial Graph for Rater 9.





B.3.10 IIT Factorial Graph for Rater 10.





REFERENCES

Anderson, N. H. (1976). How functional measurement can yield validated interval scales of mental qualities. Journal of Applied Psychology, 61, 677-692.

Anderson, N. H. & Berkowitz, L. (1978). Cognitive theories in social psychology: Papers from the advances in experimental social psychology. Boston: Academic Press. ISBN 0-12-091850-1.

Anderson, N. H. (1981). Foundations of information integration theory. Boston: Academic Press. ISBN 0-12-058101-9.

Anderson, N. H. (1982). Methods of information integration theory. Boston: Academic Press. ISBN 0-12-058102-7.

Anderson, N. H. (1991). Contributions to information integration theory. Mahwah, N.J: Erlbaum. ISBN 0-8058-0836-1.

Anderson, N. H. (1996). A functional theory of cognition . Hillsdale, N.J: L. Erlbaum Associates. ISBN 0-8058-2244-5.

Anderson, N. H. (2001). Empirical direction in design and analysis . Hillsdale, N.J: L. Erlbaum. ISBN 0-8058-3978-X.

Anderson, N. H. (2008). Unified social cognition. Psychology Press. New York: Routledge

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement, 4th* ed. (pp. 508-600). Washington, DC: American Council on Education.

Atkinson, D. (2012). Moving forward: legal issues and considerations for standard setting in professional licensure and certification programs. In G. Cizek (Eds.), *Setting performance standards: foundations, methods and innovations* (pp. 503-534). New York: Routledge.

Berk, R. A. (1996). Standard setting: The next generation (where few psychometricians have gone before!). *Applied Measurement in Education*, 9, 215-235.

Brennan, R. L., & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. *Applied Psychological Measurement*. 4, 219-240.

Brown, W. J. (2001). Social, educational and political complexities of standard setting. In G. Cizek (Eds.), *Setting performance standards: concepts, methods and perspectives* (pp. 373-386).Mahwah: Lawrence Erlbaum Associates.

Brown, W. J. (2012). Moving forward: educational, social and population considerations in setting standards. In G. Cizek (Eds.), *Setting performance standards: foundations, methods and innovations* (pp. 571-580). New York: Routledge.

Bunch, M. (2012). Practical issues in standard setting. In G. Cizek (Eds.), *Setting performance standards: foundations, methods and innovations* (pp. 415-438). New York: Routledge.

Bourque, M. L. (2009). *A history of NAEP achievement levels: Issues, implementation and impact, 1989-2009.* Paper commissioned for the 20th anniversary of the National Assessment Governing Board.

Camilli, G., Cizek, G. J., Lugg, C. A.(2001). Psychometric theory and the validation of performance standards: history and future perspectives. In G. Cizek (Eds.), *Setting performance standards: concepts, methods and perspectives* (pp. 445-476).Mahwah: Lawrence Erlbaum Associates.

Carson, J. D. (2001). Legal issues in standard setting for licensure and certification. In G. Cizek (Eds.), *Setting performance standards: concepts, methods and perspectives* (pp. 427-444).Mahwah: Lawrence Erlbaum Associates.

Cizek, G. J. (2012). An introduction to contemporary standard setting: concepts, characteristics, and contexts. In G. Cizek (Eds.), *Setting performance standards: foundations, methods and innovations* (pp. 3-14). New York: Routledge.

Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement,* 30, 93-106.

Cizek, G. J. (2012). The forms and functions of evaluations of the standard setting process. In G. Cizek (Eds.), *Setting performance standards: foundations, methods and innovations* (pp. 165-178). New York: Routledge.

Cizek, G. J. (2001). Conjectures on the rise and call of standard setting: an introduction to context and practice. In G. Cizek (Eds.), *Setting performance standards: concepts, methods and perspectives* (pp. 3-18). Mahwah: Lawrence Erlbaum Associates.

Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement*., 30, 93-106.

Cohen, Jacob (1988). *Statistical power analysis for the behavior sciences* (2nd ed.). New York: Routledge

Cohen, A. S., Kane, M.T., & Crooks, T.J. (1999) A generalized examinee-centered method for setting standards on achievement tests. *Applied Measurement in Education*, 12, 343-366.

Egan, K. L., & Green, D. R. (2003, April). *Influences on judges' decisions.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Egan, K. L., Schneider, M. C., Ferrara, S. (2012). Performance level descriptors: history, practice and proposed framework. In G. Cizek (Eds.), *Setting performance standards: foundations, methods and innovations* (pp. 79-106). New York: Routledge.

Glass, G. V. (1978). Standards and criteria. *Journal of* e*ducational Measurement*, 15, 237-261.

Hambleton, R. K., Pitoniak, M. J., Copella, J. M. (2012). Essential steps in setting performance standards on educational tests and strategies for assessing the reliability of results. In G. Cizek (Eds.), *Setting performance standards: foundations, methods and innovations* (pp. 47-76). New York: Routledge.

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. Cizek (Eds.), *Setting performance standards: concepts, methods and perspectives* (pp. 89-118). Mahwah: Lawrence Erlbaum Associates.

Hambleton, R. K. (1978). On the use of cut-off scores with criterion-referenced tests in instructional settings. *Journal of Educational Measurement*. 15, 277-290.

Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice*, 18(4), 5.

Hein, S. F., & Skaggs, G. E. (2009). A qualitative investigation of panelists' experience of standard setting using to variations of the bookmark method. *Applied Measurement in Education*, 22, 207-228.

Hein, S. F., & Skaggs, G. E. (2010). Conceptualizing the classroom of target students: A qualitative investigation of panelists' experience during standard setting. *Educational Measurement: Issues and Practice*, 29(2), 36-44.

Howard, Ian (2012). *Perceiving in Depth*. New York: Oxford University Press.

Huff, K., & Plake, B. S. (2010). Innovations in setting performance standards for k-12 test-based accountability. *Measurement: interdisciplinary Research and Perspectives*, 8, 130-144.

Impara, J. C. & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions of the Angoff standard setting method. *Journal of Educational Measurement*, 35, 69-81.

Jaeger, R. M. (1989). Certification of student competence. In R.L. Linn (ed.), *Educational measurement, third edition*. (pp. 175-218). Mahwah, NJ: Erlbaum.

Jaeger, R. M. (1990). Establishing standards for certification tests. *Educational Measurement: Issues and Practice*, 9, 15-20.

Jaeger, R. M. (1991). Selection of judges for standard setting. *Educational Measurement: Issues and Practice*, 10, 3-6, 10, 14.

Kane, M. T. (2001). So much remains the same: conception and status of validation in setting standards. In G. Cizek (Eds.), *Setting performance standards: concepts, methods and perspectives* (pp. 53-88). Mahwah: Lawrence Erlbaum Associates.

Kane, M. T. (1998). Choosing between examinee-centered and test-centered standard-setting methods. *Educational Assessment,* 5(3), 129-145.

Kane, M. (1995). Examinee-centered vs. task-centered standard setting. In *Proceedings of Joint Conference on Standard Setting for Large-Scale Assessments* (pp. 119-139). Washington,DC: National Assessment Governing Board and National Center for Education Statistics.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of educational Research,* 64, 425-461.

Lewis, D. M., Mercado, R. L. (2012). The bookmark standard setting procedure. In G. Cizek (Eds.), *Setting performance standards: foundations, methods and innovations* (pp. 225-254). New York: Routledge.

Lewis, D. M., & Mitzel, H. C., (1995) *An item response theory based standard setting procedure*. Symposium presented at the annual meeting of the California Educational Research Association, Lake Tahoe, NV.

Lewis, D. M., Mitzel, H. C., Green, D. R., & Patz, R. J. (1999). The bookmark standard setting procedure. Monterey, CA: McGraw-Hill

Loomis, S. C. (2012). Selecting and training standard setting participants: state of the art policies and procedures. In G. Cizek (Eds.), *Setting performance standards: foundations, methods and innovations* (pp. 107-134). New York: Routledge.

Mehrens, W.A., and Lehman, I.J. (1991). *Measurement and Evaluation in Education and Psychology*, (4th edn) Holt, Rinehart and Winston Inc: Orlando, Fl.

Mehrens, W. A., Cizek, G. J. (2012). Standard setting for decision making: classifications, consequences and the common good. In G. Cizek (Eds.), *Setting performance standards: foundations, methods and innovations* (pp. 33-46). New York: Routledge.

Melican, G. J., & Plake, B. D. (1985). Are correction for guess and Nedelsky's standard setting method compatible? *Journal of Psychoeducational Assessment,* 3,31-36.

Mills, C. N. (1995). Establishing passing standards. In J. C. Impara (Ed.), *Licensure testing: Purpose, procedures, and practices* (pp. 219-252). Lincoln, NE: Buros Institute of Mental Measurement.

Mills, C. N., Melican, G. J., & Ahluwalia, N. T. (1991). Defining minimal competence. *Educational Measurement; Issues and Practice,* 10(2), 7-10.

Mitzel, H. C., Lewis, D. M., Patz, R. J., Green, D. R. (2001). The bookmark procedure: psychological perspectives. In G. Cizek (Eds.), *Setting performance standards: concepts, methods and perspectives* (pp. 249-282).Mahwah: Lawrence Erlbaum Associates.

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425 (2002).

Peire, M. (2008). A guide to understanding and developing performance-level descriptors. *Educational Measurement: Issues and Practice*, 27(4), 15-29.

Plack, C. J. (2005). *The Sense of Hearing*. Routledge. ISBN 0-8058-4884-3

Plake, B. S. (1998). Setting performance standards for professional licensure and certification. *Applied Measurement in Education*, 11, 65-80.

Plake, B. S., Melican, G. L., & Mills, C. N. (1991). Factors influencing intrajudge consistency during standard setting. *Educational Measurement: Issues and Practice*, 10(2), 15-25.

Plake, B. S. (2008). Standard setters: Stand up and take a stand! *Educational Measurement: Issues and Practice*, 27, 3-9.

Plake, B. S., Cizek, G. J. (2012). Variations on a theme: the modified angoff, extended angoff and yes/no standard setting methods. In G. Cizek (Eds.), *Setting performance standards: foundations, methods and innovations* (pp. 181-200). New York: Routledge.

Phillips, S. E. (2001). Legal issues in standard setting for k-12 assessments. In G. Cizek (Eds.), *Setting performance standards: concepts, methods and perspectives* (pp. 411-426).Mahwah: Lawrence Erlbaum Associates.

Phillips, S. E. (2012). Legal issues for standard setting in k-12 educational contexts. In G. Cizek (Eds.), *Setting performance standards: foundations, methods and innovations* (pp. 535-570). New York: Routledge.

Raymond, M. R., Reid, J. B. (2001). Who made thee a judge? selecting and training participants for standard setting. In G. Cizek (Eds.), *Setting performance standards: concepts, methods and perspectives* (pp. 119-158). Mahwah: Lawrence Erlbaum Associates.

Reid, J. B. (1991). Training judges to generate standard setting data. *Educational Measurement: Issus and Practice,* 10(2), 11-14.

Shepard, L., Glaser, R., Linn, R., & Bohrnstedt, G. (1993). *Setting performance standards for student achievement.* Stanford University, Stanford, CA: National Academy of Education.

Shepard, L. A. (1995). Implications for standard setting of the National Academy of Education evaluation of National Assessment of Educational Progress achievement levels. In *Proceeding of the Joint Conference on Standard Setting for Large-Scale Assessments, Volume II* (pp. 143-160). Washington, DC: US Government Printing Office.

Shepard, G., Glaser, R., Linn, R., & Bohrnstedt, G. (Eds.). (1993). *Setting performance standards for student achievement.* Washington, DC: National Academy of Education.

Skorupski, W. P. (2012). Understanding the cognitive process of standard setting panelists. In G. Cizek (Eds.), *Setting performance standards: foundations, methods and innovations* (pp. 135-148). New York: Routledge.

Thorndike, R. L. (Ed.). (1971). *Educational measurement*. *second edition*. Washington, DC: American Council on Education.

Weiss, D. J. (2006). *Analysis of variance and functional measurement: A practical guide.* New York: Oxford University Press.

Werner, E. (1978) *Cutting scores for occupational licensing tests: Manual of considerations and methods,* Sacramento, CA: California Department of Consumer Affairs.

Webb, L. C., & Fellers, R. B. (1992) Setting the standards for passing the registration examinations. *Journal of the American Dietetic Association*, 93, 1409-1411.

Zieky, M. J. (2001). So much has changed: how the setting of cutscores has evolved since the 1980s. In G. Cizek (Eds.), *Setting performance standards: concepts, methods and perspectives* (pp. 19-52). Mahwah: Lawrence Erlbaum Associates.

Zieky, M. J. (2012). So much has changed: an historical overview of setting cut scores. In G. Cizek (Eds.), *Setting performance standards: foundations, methods and innovations* (pp. 15-32). New York: Routledge.