USING RASCH MEASUREMENT THEORY TO EVALUATE THE PSYCHOMETRIC QUALITY OF A FINANCIAL RISK TOLERANCE SCALE

by

JORGE RUIZ-MENJIVAR

(Under the Direction of John E. Grable)

ABSTRACT

The development, evaluation, and improvement of financial risk tolerance (FRT) assessments are pursued goals among scholars and professionals, particularly in the fields of financial planning and consumer economics. Accurate and consistent FRT measures are important for financial planning and counseling professionals committed to advising consumers and researchers dedicated to understanding and predicting individuals' financial decisions and behaviors under risk. In an effort to foster continuing discussions on ways to measure FRT with greater reliability and validity, and to advance FRT assessment in general, this research presents a novel paradigm to measurement with the expectation that the procedure will lead to the development of new measures, and the evaluation, and refinement of existing FRT measures. This dissertation purposed two main objectives. The first is to introduce Rasch Measurement Theory as a theoretically strong and probabilistic model, and one that may serve as an alternative to Classical Test Theory for the measurement of FRT. A thorough description of this model detailing its robust theoretical assumptions, analysis about its advantages, and discussion on its utility for improving FRT measures is presented. The second purpose is to demonstrate an

application of Rasch Measurement Theory by using a FRT measure in the context of financial planning, and consumer economics. By using a psychometric analysis based on Rasch Measurement Theory, the scale properties were evaluated and refined. This yielded an improved, robust, and psychometrically sound version of the Grable and Lytton's (1999) FRT scale.

INDEX WORDS: Financial risk tolerance, Scale, Assessment, Rasch Measurement Theory, Psychometrics.

USIING RASCH MEASUREMENT THEORY TO EVALUATE THE PSYCHOMETRIC QUALITY OF A FINANCIAL RISK TOLERANCE SCALE

by

JORGE RUIZ-MENJIVAR

B.S., University of New Orleans, 2011

M.S., University of Florida, 2013

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2016

USIING RASCH MEASUREMENT THEORY TO EVALUATE THE PSYCHOMETRIC QUALITY OF A FINANCIAL RISK TOLERANCE SCALE

by

JORGE RUIZ-MENJIVAR

Major Professor:	John E. Grable
Committee:	George Engelhard, Jr.
	Swarn Chatterjee
	Sophia T. Anong

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2016

ACKNOWLEDGEMENTS

During my three years in the program, I have had the opportunity to meet, work, and collaborate with some of the most talented and brightest minds of our field. Thanks to each one of you for influencing and shaping my interests and education, and making my time here at the University of Georgia (UGA) truly memorable.

I would like to give my sincere thanks to my committee chair, Dr. John E. Grable. Thank you for your patience, friendship, unconditional support, and mentorship during these three years. This dissertation would have been impossible without your true dedication and devotion. Words are not enough to express my deep gratitude.

In addition, I would like to extend my earnest appreciation to the rest of my committee members. Dr. George Engerlhard, Jr., I cannot begin to thank you for introducing me to the "beauties" of Measurement Theory. Your classes, discussions, and mentorship have been inspiration for this project, and the many more to come. I knew I was in the "right" place, with the "right" person when I first came to Measurement class in Fall 2014. Dr. Sophia T. Anong, thanks for your friendship, support, and enthusiasm. The seed about the usage of scaling tradition methods for this dissertation was first planted in your "Research" class. And last but not least, I am thankful with Dr. Swarn Chatterjee for the invaluable input for this dissertation. I am truly grateful for your time and help for the completion of this project. I look forward to collaborating with each one of you in the future!

Moreover, I would like to take the time to express my deep appreciation to the

faculty and staff of the Department of Financial Planning, Housing and Consumer Economics (FHCE), as well as the College of Family and Consumer Sciences (FACS). Special thanks go to Dr. Sheri Worthy for trusting in me and for always creating opportunities that have enhanced and shaped my academic career from multiple angles. Also, I would like to take a minute and thanks Dr. Andrew "Andy" T. Carswell, a friend and the best "study abroad" colleague you could ever ask. I am grateful for allowing me to be part of such an exciting program in Costa Rica. You will be missed in CR this year! Another person in FHCE who I would like to express my gratitude is Dr. Robert B. Nielsen. Thanks for your friendship, and for the time and guidance along this journey. What fond memories I have of all those Thursday afternoon conversations during my first year at UGA.

An important element of these three years in Athens is the friendships I have fostered. I would like to thank to the FHCE graduate students (former and current) who have been always there to "fight this battle" together, side by side. Special thanks must go to Judith Aboagye, Lu Fan, Sae Rom Chung, Dr. Wookjae Heo, Ji-Young Jung, Narang Park, Dr. Kristi-Warren Scott, Kimberly Watkins, Lini Zhang, and Haidong Zhao. Thanks for your collegiality, encouragement, and unconditional support.

I would like to also thank Ilyar Heydari-Barardehi. Thanks your friendship and brotherhood, and for repeatedly proving yourself. This journey would have been "Zafran-less" without you. I am eternally grateful to you.

Finally, I would like to thank my family. Brent R. Carr (and Toby, Hunter, Babü, Miška, and Nikolai), this expedition would have not been successful without your inspiration, help, and direction. Infinite paragraphs of appreciation would not be enough

to thank you! Last but not least, I would like to express my thankfulness to rest of my family: my mother (Alva de Ruiz), father (Jorge Ruiz Barrera), siblings (Erika, Alba, Ricardo, Fernando, and David), Brad and Emily Judy, and Luis and Douglas. Thank you all for your love and trust.

"Finally, from so little sleeping and so much reading, his brain dried up and he went completely out of his mind." Miguel de Cervantes in Don Quixote.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

The development, evaluation, and improvement of financial risk tolerance (FRT)[1] assessments are ongoing efforts among scholars and professionals, particularly in the fields of financial planning and consumer economics. Accurate and consistent FRT measures are important for both financial planning and counseling professionals committed to advising consumers and researchers dedicated to understanding and predicting individuals' financial decisions and behaviors under risk.

The present study aims to foster the ongoing discussion about ways to measure FRT with more reliability and validity. A novel theoretical and measurement paradigm, Rasch Measurement Theory, is presented in detail. This research provides arguments and statistical evidence to demonstrate how some of the limitations inherent to measurement methods previously used in financial planning and consumer economics (e.g., Classical Test Theory) can be addressed by using this alternative measurement technique. Furthermore, an example of the application of this measurement model to a popular FRT instrument (i.e., Grable and Lytton's [1999] scale) is presented. The end product of this research is the presentation of a refined, improved, and psychometrically sound version of this scale. Ultimately, this dissertation serves as an educational piece with discussions

[1] FRT refers to an individual's attitude towards financial risk and to the willingness to undertake more financial risk for the potential of obtaining higher returns (Nobre & Grable, 2015).

and implications that invite exploration and consideration of different measurement aspects for FRT.

The Importance of Measuring Financial Risk Tolerance

Financial risk tolerance in the profession

In the financial planning profession and securities advice industry, increasing attention has been given to the measurement and assessment of clients' FRT. For example, in 2010, the Financial Industry Regulatory Authority[2] released an official notice that the Security Exchange Commission (SEC) had approved the adoption (with implementation date of July 2012) of a new rule (Rule 2111) governing client's suitability (FINRA, 2011; 2012a; 2012b). A particular aspect of this rule governs the customer-specific suitability obligations expected of a broker-dealer, such that the associated person should have an adequate fund of knowledge to "have a reasonable basis to believe that a recommended transaction or investment strategy involving a security or securities is suitable for the customer, based of the information obtained through reasonable diligence…to ascertain the customer's investment profile" (FINRA, 2014). Specifically, FINRA's 2111 rule cited FRT as a key element that should be assessed when constructing a customer's investment profile.

Similarly, the Certified Financial Planning (CFP) Board of Standards, Inc.[3] recently issued a resource publication outlining updated learning objectives for curriculum development as related to FRT assessments. Examples include helping a

[2] FINRA is a nonprofit organization and independent regulator that oversees nearly 4,000 brokerage firms and more than 600,000 registered securities representatives doing business with the U.S. public (FINRA, 2016).

[3] CFP Board does not explicitly spell out the assessment of FRT as part of their *Practice of Standard, Code of Ethics or Rules of Conduct.* However, Practice of Standard 300-1, Rule of conduct 4.1 and 4.4 allude to this practice to some extent (CFP, 2016).

client to identify FRT levels when developing and analyzing a clients' portfolios during the investment planning process, and considering FRT as a factor for selection of suitable retirement investments (CFP, 2015). The overarching purpose in the issuance of such guiding principles within the profession is to ensure that financial and investment recommendations match a client's goals through alignment with the client's investment profile factors, such as FRT, investment experience, and time horizon, among others.

Thus, as the measurement of FRT becomes more prevalent within the profession—especially when advisors and financial planners are presented with the task of creating investment plans for clients—the evaluation of the quality of the instruments employed to assess FRT will garner greater attention. Ultimately, practitioners are interested in using metrics that consistently and accurately measure risk tolerance and any other relevant element of a client's investment profile, so that their financial advice and strategies are suitable, pragmatic, and abide to the standards stipulated by the profession regulators.

Financial risk tolerance in academia

Over the years, the academic literature has documented the significant role that FRT has played in shaping financial decision-making under uncertainty, such as investing and asset allocation (Barsky, Juster, Kimball, & Shapiro, 1997; Grable, 2000; 2008; Grable & Joo, 2004; Grable & Lytton, 1999; 2001; Hanna & Lindamood, 2005; Mittra, 1995; Sung & Hanna, 1996). Hence, in academia, researchers have paid special attention to the valid and reliable assessment of FRT, as well as to its relationship and interaction with other variables (e.g., life cycle variables, demographics, and financial variables such as horizon and financial position). Scholars working, specifically in the areas of financial

planning and consumer economics, have devoted effort to refining FRT estimation. Examples of such efforts include the revision of FRT's definition, the differentiation of this construct from other risk-related concepts (e.g., risk capacity, risk preference, risk need), the identification and understanding of the different potential domains from which FRT might be composed, and ongoing development and improvement of instruments (e.g., questionnaires and tests used to assess FRT) (Grable, Archuleta, & Nazarinia, 2010; Grable & Lytton, 2001; Hanna, Gutter, & Fan, 2001; Hanna & Lindamood, 2004; Nobre & Grable, 2015). In fact, in terms of the formal development and refinement of FRT instruments, several academic endeavors have been completed in the last 25 years (Brayman et al., 2015; Kuzniak, Rabbani, Heo, Ruiz-Menjivar, & Grable, 2015). For instance, scholars have developed instruments using theoretical approaches (i.e., Classical Test Theory) and also empirical perspectives; and they have evaluated the properties (e.g., items selection, word choice, questionnaire length) and quality of existing instruments. Moreover, they have assessed the correlation between these instruments and reference behaviors, such as actual portfolio allocation.

Ultimately, researchers have aimed for the development, and have continued targeting the development, of new metrics and improve existing tools that allow a more accurate and consistent measurement of this elusive construct. The importance of better and more stable measures for FRT lies in the fact that such instruments are used to model and predict human behavior, specifically financial decision-making processes—from everyday money decisions to financial choices after unexpected events such as a financial crisis. The end product of such academic endeavors not only benefits the academic

community, but also provides practitioners with instrumentations that assess FRT in a more effective, efficient, scientific, and objective manner.

How is Financial Risk Tolerance Measured?

Different approaches have been offered for the measurement of FRT in academia and in the profession. For the most part, quantitative techniques, both theory and non-theory grounded have been predominantly used in research settings (e.g., scales and indices). Per contra, in the profession, non-theoretical and qualitative (e.g., subjective evaluation via interviews) approaches are still customary (Brayman et al., 2015; Carr, 2014). In recent decades, though, the use of either mixed methods (i.e., a combination of qualitative and quantitative techniques), or solely quantitative systems, have gained usage among practitioners.

Specifically, in academia, rating scale questionnaires have been primarily used when assessing FRT. Examples of these are the Grable and Lytton (1999) FRT Scale (GL-FRT), the Barsky, Juster, Kimball, and Shapiro (1997) instrument, and the popular and parsimonious Survey of Consumer Finances (SCF) single item FRT question. In essence, researchers have aimed to quantify the spectrum of FRT by assigning scores (that are often translated into nominal classifications) to particular response categories of a question(s).

For some of these questionnaires, especially those developed with a theory of measurement in the background, the scoring systems (e.g., cut-off values and category response level ordering for questions) have been established via Delphi panels composed of experts in the field (i.e., academicians, policy makers, and practitioners). Additionally, in theoretically sound instruments, the question content and category response level

ordering have been based on area related theories. In the case of FRT, finance and economic theories have been employed—from theories where the agent is deemed to be rational (e.g., Utility Theory-based models, and the Capital Asset Pricing Model) to theories in which cognitive biases and psychological elements of human behavior are present (e.g., Prospect Theory, Modern Portfolio Theory)

Existing FRT instruments utilize the following four major types of questions: (a) choice of type of investments and financial products selected, (b) questions on risk/reward, (c) a combination of the aforementioned type of questions, and (d) reported actual behavior or observed actual behavior (Hanna, Gutter, & Fan, 2001; Shelbecker, Roszkowski, & Cutler, 1990). Examples of actual questions included in existing instruments are presented here to illustrate the different types of questions used in practice.

The widely used one-item SCF[4] measure evaluates FRT with the following choice of type of investment and financial product question:

"Which of the following statements…comes closest to the amount of financial risk that you are willing to take when you save or make investment?

1. Take substantial financial risk expecting to earn substantial returns.
2. Take above average financial risks expecting to earn above average returns.
3. Take average financial risks expecting to earn average returns.
4. No willing to take any financial risks."

[4] It is questionable onto whether a one single item test can comprehensively measure and capture the complexity and elusiveness of FRT (Bonoma & Schlenker, 1978; Gilliam, Chatterjee, & Grable, 2010; Grable and Schumm, 2010; Roszkowski, Davey, & Grable, 2005). A discussion on this issue is presented later in this chapter.

The scoring system traditionally used for this question rates respondent's FRT level as "substantial" if response choice 1 is selected, "above average" if 2, "average" for 3, and "low risk" if 4 is chosen.

The following is an example of a "risk/reward" hypothetical question used in the Barsky et al. (1997) instrument:

A. "Suppose that you are the only income earner in the family, and you have a good job guaranteed to give you your current (family) income every year for life. You are given the opportunity to take a new and equally good job, with a 50-50 chance that it will double your (family) income and a 50-50 chance that it will cut your (family) income by a third. Would you take the new job?"

B. "Suppose the chances were 50-50 that it would double your (family) income and 50-50 that it would cut it in half. Would you still take the new job?"

C. "Suppose the chances were 50-50 that it would double your (family) income and 50-50 that it would cut it by 20 percent. Would you take the new job?"

The scoring system for this instrument generally entails a two-stage process in which the first stage in the scoring process relies on the answer indicated to question "A." A participant who accepts the job proposed in "A" is then asked question "B." However, if the subject declines the job offered in "A," then he or she is asked "C." The second stage of the scoring process serves to assign the FRT level for each respondent. An individual who indicates "yes" to A and B is deemed to exhibit "very high-risk"

tolerance. If the subject selects "yes" for A, and "no" for B, then a level of "high risk" is concluded. "The low risk" level is assigned to a subject whose response pattern is "no" to A and C. Finally, the FRT level for a subject who answers "no" to A, but "yes" to C is categorized as "very low-risk tolerant."

The 13-item GL-FRT scale uses a combination of the previously mentioned types of questions.[5] For instance, an example of a "risk and reward" question reads as follow: "Given the best and worst case returns of the four investment choices below, which would you prefer?

a.$200 gain best case; $0 gain/loss worst case

b. $800 gain best case; $200 loss worst case

c. $2,600 gain best case; $800 loss worst case

d. $4,800 gain best case; $2,400 loss worst case"

A hypothetical question in which the selection of an investment vehicle is used to evaluate FRT asks:

"Suppose a relative left you an inheritance of $100,000, stipulating in the will that you invest ALL the money in ONE of the following choices. Which one would you select?"

a. A savings account or money market mutual fund

b. A mutual fund that owns stocks and bonds

c. A portfolio of 15 common stocks

d. Commodities like gold, silver, and oil

Finally, some researchers have argued that risk-related behaviors exhibited when making a decision under uncertainty can be utilized as substitute indicators for FRT

[5] The complete list of questions and suggested scoring system for the GL-FRT is presented in appendix A of this manuscript.

(Kimball, Sahm, & Shapiro, 2008; Schooley & Worden, 1996). In the literature, the actual allocation of one's investment portfolio (e.g., percentage of assets allocated in stocks, bonds, mutual funds, certificate of deposit, etc.) has been used as a proxy for FRT. While, actual behaviors (either self-reported or observed if available) are used as a reference to establish validity and evaluate the predictive power of instruments (namely "concurrent validity" in psychometrics), from a measurement perspective, such behaviors do not necessarily equate to the construct of interest. For instance, in the context of risk-taking attitudes, the allocation of assets is seldom a function of a client's FRT exclusively. Other explicit or implicit risk factors might be present, such as risk capacity (i.e., the ability to undertake a loss from a particular decision under risk [Nobre & Grable, 2015]), risk preference (i.e., the overall feeling an investor has towards the selection of an investment choice over another, regardless of whether the feeling is based on subjectively or objectively [Nobre & Grable, 2015]), and additional variables such as age, previous financial experience, financial knowledge, or financial planning and professional assistance. Generally speaking, it is the case that from a measurement and theoretical perspective, attitudes are predecessors of reference behaviors. And for that reason, scores from an attitudinal instrument (e.g., FRT scale), for example, should be linked and used to predict reference behaviors (e.g., actual asset allocation) (Messick, 1995). Hence, caution should be exercised with any loose treatment of attitudes and their reference behaviors as being interchangeable entities.

Within the profession, non-theoretical qualitative approaches are still prevalent in practice. For instance, it is not uncommon to encounter personal financial planners and counselors assessing a client's FRT based solely upon subjective judgments formed

during the interview process. Snelbecker, Roszkowski, and Cutler (1990) explained that some of this overconfidence exhibited by advisors who rely only on subjective judgments might originate from the belief that years of experience in the industry, training, and education are sufficient elements to make informed assessments of their client's FRT. From a theory of measurement perspective, this approach is both questionable and problematic, as exclusively using subjective judgments, with no theoretical framework, possesses a higher degree of inconsistency in the inter and intra-assessment of a client's FRT. At the same time, this approach might potentially lead to inaccurate, presumptuous, and biased conclusions; hence, such practices are generally discouraged (Carr, 2014; Roszkowski, Davey, & Grable, 2005).

Other practitioners employ a combination of subjective judgments and more objective instruments (e.g. questionnaires) to derive customers' FRT levels. Quantitative (either theory or non-theory based) approaches to measure FRT have gained popularity in recent decades among some professionals in the field. For example, some have started to use available FRT instruments produced by researchers, while others have developed their own built-in-home questionnaires that match their needs (e.g., brevity and parsimony, inclusion of questions that ask only about products offered by the company, etc.).

Interestingly, many of these built-in-home questionnaires are premised on the assumption that the inclusion of variables such as gender, age, income, profession, race, and education are needed to compute and originate conclusions about a client's FRT (Grable & Lytton, 1999; Roszkowski & Grable, 2005; Roszkowski, Snelbecker, & Leimberg, 1993; Shelbecker, Roszkowski, & Cutler, 1990). Likewise, this technique has

also been reported as a common practice among those who subjectively assess FRT (Roszkowski & Grable, 2005). It is important to note that this contentious method of including non-risk related variables to assess FRT is actually troubling. When viewed with a theory of measurement, this technique poses a major threat to the validity of the instrument—specifically, it violates the *unidimensionality* aspect of validity (Messick, 1995).

In addition, for some, the use of questionnaires is simply a tool used to meet a fiduciary or suitability standard established by governing and credentialing boards and for establishing physical evidence by gathering enough data from the client to make financial recommendations. Yet, the reliance of subjective judgments to measure FRT is still far greater than the trust of the instruments alone (Hanna, Finke, & Waller, 2008; Roszkowski & Grable, 2005; Van de Venter, & Michayluk, 2007).

Finally, at this point, the reader might have inductively noticed that there is not a single standardized manner or at least commonly approved aspects/guidelines to measure FRT that has been agreed upon within or between academia and the profession (Bouchey, 2004). From a measurement perspective, this poses a challenge in terms of consistency and comparability of scores or nominal attributed level of FRT obtained from different measures. Efforts and initiatives to develop FRT standardized measures and assessment practices have been in place for more than a decade (Grable & Lytton, 1998; 1999; Kuzniak et al., 2015; Roszkowski, Davey, & Grable, 2005), yet the adoption of such scales has been slow. The present study can be perceived as a follow up of to previous initiatives to foster the ongoing discussion about more efficient and effective approaches and methods to measure FRT. At the same time, this study provides specific arguments

that explain alternative theoretical views and methods that allow for a more granular assessment of FRT, while also addressing some of the current problems present in the measurement of this construct.

Can Financial Risk Tolerance Be Measured More Effectively?

The absence of measurement theory in financial risk tolerance instruments

As mentioned earlier in this discussion, several of the questions', scoring systems, and instrument cutoff values included in some of the existing FRT measures have been formulated using risk-related theories, such as economic theory and prospect theory. Without doubt, the consideration of such theories is an important aspect in the scale development and evaluation process, yet the use of these content related theories is just one of the theoretical aspects to consider in the pursuit of stable measures.

In fact, the relevant elements of the measurement in a scale's construction and refinement (e.g., scores consistency and accuracy, practices on selecting cut-off values for scoring, score meaning, test utility, among others [Messick, 1995]) are guided by a theory of measurement from various logic and mathematical models developed in the discipline of Test Theory (also known as psychometrics). Test Theory is crucial in that it lays the foundation for measurement. Stevens' (1946) conceptualization of measurement—a commonly accepted definition—is the assignment of a numeral to an object or event according to rules. Tongerson (1958) and Lord and Novick (1968) improved this definition by positing that measurement should pertain to the properties of an object or subject of interest, rather than the object or subject in and of themselves.[6]

[6]Properties are the concern of measurement (Pfanzagl, 1968). In geology, a researcher says she wants to measure the volume of a metamorphic rock, and not the rock itself.

Second, it is the field of psychometrics from which arises a general framework for the various steps involved in the cyclical measurement process that allows assessment from early stages in scale development and evaluation to the continuous maintenance phases of the instrument. Third, the discipline devotes efforts to determine the effects of prevalent measurement problems in scale construction and evaluation and attempts to develop novel approaches and methodologies to overcome or at least mitigate some these measurement challenges (Croker & Algina, 2006). Finally, the field of psychometrics seeks to improve test construction and evaluation practices using enhanced techniques to operationalize variables of interest, and uses robust analytic methods to test the accuracy and reliability of tests. In summary, the field of psychometrics seeks a more intelligent use of such instruments in decision-making (e.g., the consideration and implication of score meaning and consequences of test use[7]).

Unfortunately, a shared, and limiting, characteristic among several existing FRT instruments is the absence of a formal theoretical measurement background in the scale construction, evaluation, or maintenance process. Take, for example, the FRT item from the Survey of Consumer Finances—a dataset sponsored by the Federal Reserve Board. This measure has often been used in finance and cited in economics research. In fact, researchers often prefer this measure as it offers a more parsimonious manner for measuring FRT. Additionally, the preference and adoption of this instrument (i.e., The Survey of Consumer Finances) may be linked to the fact that it is the only FRT question

---

Thus, two different objects might become equivalent if consideration is constrained to one property: All metamorphic rocks (regardless of the shape) with the same volume.

[7] See Messick (1995) for a complete discussion on the consequences of score meaning and test utility.

continuously included in a nationally representative survey of American households and consumers (Grable & Schumm, 2010).

However, over the years, researchers have noted that this item might not be a "good proxy for people's true risk aversion" (Chen & Finke, 1996, p. 94),[8] and that it may not capture the complex nature of FRT (Gilliam, Chatterjee, & Grable, 2010; Grable & Schumm, 2010). A study conducted by Grable and Schumm (2010) revealed equivocal results from the psychometric analysis performed on this instrument.[9]

Additionally, this item lacks a test theory background. The project director of the SCF, Arthur Kennickell, reported that the SCF risk tolerance question was developed by the New York Stock Exchange. He noted that there was no attempt to measure validation when the question was first included in the SCF in 1983 (Yao, Hanna, & Lindamood, 2004). Thus, despite its wide acceptance in both academia and the profession, from a measurement perspective, this instrument lacks the statistical and theoretical sophistication, and potentially lacks the quality controls, needed when using a measurement theory.

In summary, an important first step in the improvement of FRT measures is the incorporation of measurement theory and psychometric analyses starting from the development stages and through the constant evaluation and maintenance of such measures.

[8] Risk aversion is defined as the inverse of FRT (Nobre & Grable, 2015).
[9] The estimated reliability of the SCF item was most likely between $\alpha = .52$ and $\alpha = .59$ (Grable & Schumm, 2010).

Using test-score tradition versus scaling tradition in financial risk tolerance measures

However, not all of the existing FRT measures have neglected the incorporation of measurement theory in the development and evaluation stages. Consider existing FRT measures that have utilized models and techniques that fall under the umbrella of the paradigm referred to as Test-Score Tradition. Table 1 shows some of the different models included under such a paradigm.

An example of this is the private and patented FRT instrument provided by the Australian based firm *FinaMetrica*, which financial advisors and institutions have extensively used. On the firm's website, it is reported that their risk profiling system possesses measurement properties that exceed the generally accepted psychometric standards (FinaMetrica, 2015). The patent of its automated assessment of FRT reveals that *FinaMetrica* based its current FRT system on the original version of the Survey of Financial Risk Tolerance[10] (SOFRT) developed by Dr. Michael Roszkowski (1992). The latter was created employing Test-Score tradition models, such as Classical Test Theory (CTT) analysis.

[10] The questions on the SOFRT are varied in nature, including: preferences for different investment vehicles, expected returns, reactions to sample portfolios, life style characteristics, probability and payoff preferences, preferences for guaranteed versus probable gambles, minimal required probability of success, and minimal return required to undertake a risk.

[11] PCA is used to simplify the structure of a set of variables. The ultimate purpose of PCA is to reduce the number of observed variables to a smaller number of components (Johnson & Wichern, 2007).

Table 1.

Measurement Theories Classified into Two Research Traditions (Engelhard, 2013)

Test-Score Tradition	Scaling Tradition
Key Models: 1. Classical Test Theory (CTT) 2. Generalizability Theory (GT) 3. Factor Analysis (FA) 4. Structural Equation Modeling (SEM)	**Key Models:** 1. Psychophysical Model (PM) 2. Absolute Scaling (AS) 3. Item Response Theory (IRT) 4. Non-Parametric Item Response Theory (NIRT)
Essential Features: Test-score focus, Linear models, Focus on test scores and the estimation of error components.	**Essential Features:** Item-person response focus, Non-linear models, Focus on modeling the responses of persons to items.
Key Theorists and models: Spearman (CTT) Kuder and Richardson (CTT) Cronbach and his colleagues (GT) Spearman (FA) Thurstone (FA) Jöreskog (SEM)	**Key Theorists and models:** Thorndike (PM) Thurstone (AS) Birnbaum (IRT) Rasch (IRT) Guttman (NIRT) Lazarsfeld (NIRT) Mokken (NIRT)
Theory intro practice: Brennan (GT) Jöreskog (SEM)	**Theory intro practice:** Lord (IRT: Birnbaum) Wright (IRT: Rasch)

Another popular scale and well-established instrument in the field of financial planning that has incorporated measurement theory in its development is the GL-FRT scale. Grable and Lytton first published the scale in 1999 and described the steps taken in the scale's creation and evaluation, yet did not formally express the particular measurement paradigm used therein. They did, however, mention using PCA[11] as a technique to assess dimensionality. Regardless, their reported psychometric analyses and

evaluations corresponded particularly well to Classical Test Theory analyses. It is important to mention that, unlike the *FinaMetrica* instrument, the GL-FRT is a publicity available scale. As a result, the GL-FRT scale has been widely used both by numerous researchers in the field, as well as, by small financial planning firms and independent advisors (Kuzniak et al., 2015). To the knowledge of this author, no current instrument that is being used in the field of financial planning or consumer economics has been developed or evaluated using other models in the scaling tradition (e.g., Item Response Theory and Rasch Measurement Theory).

Test-Score tradition, and specifically CTT analysis, remains the dominant approach for developing and evaluating measurement data in both FRT research, and the field of financial planning and consumer economics in general. In part, CTT's popularity and common usage arises from CTT being simpler conceptually and theoretically, as well as in practice (Liu, 2010). However, such simplicity comes with a trade off; CTT is subject to limitations that stem from weak theoretical assumptions. Consider two examples of the limitations embedded within CTT. One clear limitation is in its poor precision for estimating a construct of interest (e.g., FRT) at the individual level, for instance. This limitation arises because CTT assumes a constant measurement error for every subject within the population. Another limitation of CTT (and of models in the Test-Score Tradition) is that it uses only the total items sum score in the analysis; thereby, failing to account for specific attributes (to be described later) of each question included in the instrument. Note that specific emphasis is given to CTT in this project, as it currently remains the most dominant measurement methodology used in the development of FRT measures and in financial planning and consumer economics. A

more thorough description of the assumptions and limitations of CTT is offered in Chapter Two.

Fortunately, the discipline of psychometrics has made major advances in the last five decades. A new measurement paradigm and alternative methods (i.e., Scaling Tradition), compared to the older CTT and other test-score tradition models, have been offered. Two popular models from the Scaling Tradition are Item Response Theory (IRT) and the Rasch Measurement model—often grouped together and broadly referred to as Modern Psychometrics or Modern Measurement Theory (See Table 1 for a detailed list of other models included in the Scaling Tradition). These have been extensively used in different academic and professional disciplines, from educational and applied psychology to behavioral health and clinical sciences such as medicine and nursing (Engelhard, 2013). Modern Psychometric Theory's adoption in many fields lies in that the fitting of these probabilistic models possesses strong measurement assumptions (e.g., Hambleton & Swaminathan, 1984; Lord & Novick 1968; Wainer & Thissen 2001). And, importantly, the modern era of psychometrics has allowed for the liberation of measurement from limitations found within CTT. To this extent, IRT and The Rasch model offer advantages to the more traditional methods of instrument development and evaluation. For example, the new models allow for separation and individual analysis of the following: items (e.g., questions used in the instrument), persons (e.g., subjects responding the questionnaire), and ability (e.g., latent constructs of interest, such as financial risk tolerance, financial education, and financial well-being). Moreover, consider IRT and Rasch model's advantage over CTT in error, as the former use individualized standard errors of measurement for ability measures. In essence, then,

these techniques provide for a finer detection granularity upon inspection of measurement elements and issues that threaten reliability, validity, and the stability of measures, which otherwise might not be identified from a broader perspective such as in the coarser granularity seen in a total score. Thus, the application of modern psychometric techniques to the measurement of the elusive construct of financial risk tolerance can enhance accuracy, and reliability for estimates, not only at the total test score and group level, but also at the individual item and person level. Thus, these microscopic analyses, and evaluations from a stronger theoretical basis, may shed novel insights and enhance deliberation on measurement issues that may hone techniques leading to improved measurement, including that of FRT assessment.

Finally, and at this point it is pivotal to note that, this project focuses extensively on the application of the Rasch model as this was the model chosen for the analysis. Thus, exclusive attention to this model is paid hereafter. A detailed description of the history and evolution, assumptions, and components of the Rasch Model is provided in Chapter Two. Additionally, an in-depth comparison between CTT and the Rasch Model is presented in the following chapter.

Purpose of This Study

This dissertation was developed with two main objectives. The first was to introduce Rasch Measurement Theory as a theoretically strong and probabilistic model, and one that may serve as an alternative to CTT for the measurement of FRT. A thorough description of this model, theoretical assumptions, advantages, and utility for improving FRT measures is presented. The second objective was to demonstrate an actual application of Rasch Measurement Theory by using a popular FRT measure in the field

of financial planning and consumer economics. Specifically, the Grable and Lytton (1999) Financial Risk Tolerance scale (GL-FRT) was evaluated in this study. By using a psychometric analysis based on the Rasch Model, the scale properties were evaluated and refined. This yielded an improved, robust, and psychometrically sound version of the GL-FRT.

Framework Used in this Study

For this project, the Scaling Tradition paradigm was used to evaluate FRT. Specifically, Rasch Measurement Theory was selected for the theoretical framework. One main reason for this choice was that Rasch Theory proposes a set of probabilistic models developed for the purpose of describing response patterns of respondents to individual items. The Danish mathematician, Georg Rasch developed this basic model in his pioneering work in Rasch (1960 [1980]; 1961). In essence, the Rasch model proposes that a given construct of interest (FRT in this case) can be thought of as a line or a continuum upon which items (e.g., questions included in the scale), and subjects/objects (e.g., respondents) can be mapped. Figure 1 illustrates this idea. Specific properties of items and subjects (namely item and person parameters, respectively) are evaluated to determine their location along the continuum. For items, the difficulty (also called *endorsability* when measuring attitudinal constructs) is used to locate the item on the line; whereas, for respondents, the ability (or exhibited level of the construct of interest) is used. A detailed description of the framework is presented in Chapter Two.

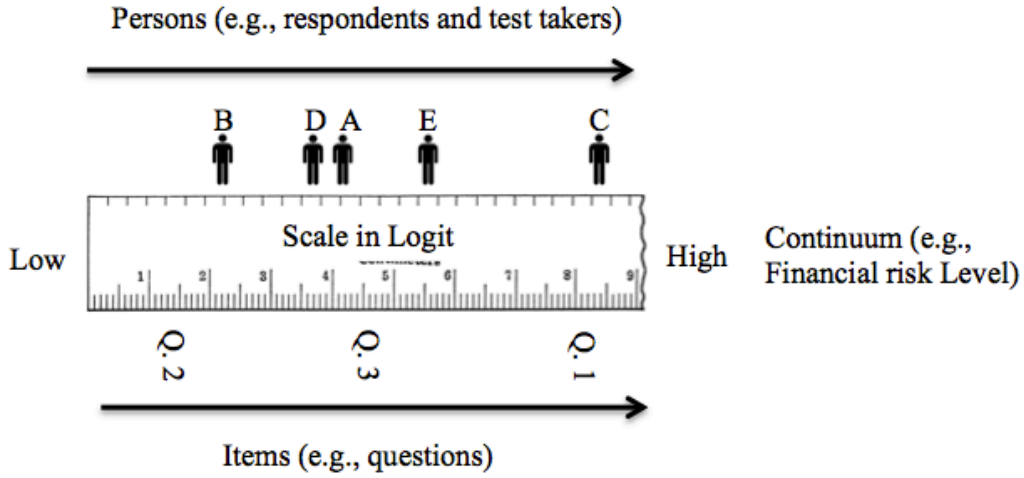


Figure 1. A Graphical Representation of the Main Idea Proposed by Georg Rasch: A Ruler (Continuum) Common to Persons and Items.

Significance and Contribution of This Research

This study was undertaken to advance the research on FRT assessment through the introduction of a novel measurement paradigm with the expectation that it will lead to adoption in the development, evaluation, and refinement of other FRT measures. The end product of this research is an improved version of a FRT measure exhibiting a strong psychometric basis, and one that is readily available to researchers and financial professionals. The form of this scale allows for a more systematic and consistent measurement of FRT across samples, studies, and programs.

Finally, the contribution of this project is not exclusive to the FRT literature, as the utilized measurement framework and model are applicable for the field of financial planning and consumer economics. This research can be conceived of as a stepping-stone for researchers in these fields to learn and apply the underlying methodology to other instruments that assess latent constructs of interest (e.g., financial well-being and financial stress). The use of modern psychometrics in these fields should encourage a

reconsideration of how measurement issues can be explored, from a different perspective that permits finer granularity and potentially greater specificity.

Summary of Methodology

For the Rasch Measurement analysis in this project, the GL-FRT scale—an extensively used FRT measure in the field of financial planning and consumer economics (Kuzniak et al., 2015)—was employed. This particular FRT instrument was selected for the following reasons: (a) it was developed using a psychometric theory (i.e., CTT), unlike some other existing instruments, such as the Survey of Consumer Finances FRT item on FRT or the Barsky et al., (1997) measure, which both lack theory of measurement background; (b) the GL-FRT measure is a publicly available instrument with readily available scoring guidelines for users, whereas other instruments created and refined with the use of test theory (i.e., CTT), such as *Finametrica* FRT instrument, are not free to the public; and (c) the author of this project had convenient access to a large, multi-year dataset with responses from a demographically rich sample to the particular instrument.

The data for this research were obtained from a repeated cross-sectional data collection project hosted by Rutgers New Jersey Agricultural Experiment Station. From year 2007 to 2014, Rutgers collected over 2000,000 responses to the GL-FRT scale via an open-access site (http://njaes.rutgers.edu:8080/money/riskquiz/). The sample frame used for this project came from responses collected from January 2013 through December 2013. This period was selected, as year 2013 was the most recent period with submitted responses throughout the entire year. Responses for year 2014 were available until the month of June. Including only valid and completed responses, the sample frame

for the particular period was composed of 25,079 subjects. Further, this sample was delimited to responses on subjects 25 years of age or older. Then, the final sample used was comprised of responses of 11,906 subjects.

Since the inception of the initial, basic Rasch Model, additional extensions have been developed (e.g., Rasch rating scale model, partial credit model, and many-facet model). For the purpose of this study, the Wright and Masters (1982) Partial Credit Model was employed. It is important to remember that the number of response categories across all the items on the GL-FRT scale is not consistent. For example, question 1 of this scale has four response categories, whereas question 4 has only three response categories. The partial credit model allows each item or questions to have its own response category structure; thus, this model was suitable for the scale analyzed in this research.

Delimitations

In this project, subjects younger than 25 years of age were excluded from the analysis. The rationale for this decision was based on the fact that several of the questions in the GL-FRT scale are content specific for the choice of making formal investments via financial assets (i.e., stock, bonds, money market) in organized (exchange) markets. The literature suggests that young adults typically do not have sufficient wealth to formally invest in securities markets (Campbell, 2006; Constantinides, Donaldson, & Mehra, 2002; Guiso, Haliassos, & Jappelli, 2002; Van Rooij, Lusardi, & Alessie, 2011). Additionally, the majority of young professionals start to actively participate in such markets only once they have access to a retirement portfolio—typically through an employer (Bassett, Fleming, & Rodrigues, 1998; Madrian & Shea, 2000; Mullainathan & Thaler, 2000) such that the exclusion seemed prudent for generalizability purposes.

Organization of Remainder of Dissertation

The remainder of this dissertation is organized in the following manner. Chapter Two provides a literature review that describes in detail the main aspects of the theory used in this study (i.e., Rasch Measurement Theory). The discussion reviews previous psychometric works that have been completed on the measurement of FRT. Special attention was directed at the research done for the FRT scale utilized for the analysis (i.e., the GL-FRT scale) demonstrating that it is psychometrically sound. Chapter Three describes the research method used to complete this study—the Rasch Measurement Model, and more precisely the Partial Credit Model. The results of the statistical analyses are presented in Chapter Four. Finally, Chapter Five culminates in a discussion of the results, implications for research and practice, limitations of the present study, and recommendations for future research in the area of FRT measurement.

CHAPTER 2

LITERATURE REVIEW

For organizational purposes, this chapter is divided into the following three main sections: (a) *What is Financial Risk Tolerance?* (b) *A Review of Theories of Measurement: Classical Test Theory and Rasch Measurement Theory,* and (c) *The Grable and Lytton (1999) Financial Risk Tolerance Scale*. The first section of this chapter offers a review of the characterization of financial risk tolerance (FRT) and other risk-related concepts used in the field of financial planning and consumer economics. The second section describes two theoretical frameworks (i.e., Classical Test Theory and Rasch Measurement Theory) used for the measurement of unobserved or indirectly measured variables (also called latent constructs), such as the attitudinal concept of FRT. Finally, the last section of this chapter presents the Grable and Lytton FRT Scale (GL-FRT) and the psychometric work previously completed with this particular scale.

Section I. What is Financial Risk Tolerance?

Financial risk tolerance (FRT) is a significant and influential factor in economic and financial decision-making under uncertainty (Grable, 2008). At the household level, FRT becomes an influencer on decisions, such as asset and retirement portfolio allocation, and strategies for achieving wealth, growth, and accumulation (Barsky et al., 1997). Typically, those individuals with higher levels of FRT obtain higher returns on their investments over longer periods of time (Grable, 2008). The literature also documents correlational relationships between FRT and other financial variables. For

instance, scores of those persons with higher levels of FRT are associated with greater net worth and income, higher levels of financial knowledge and financial satisfaction, and greater economic expectations (Grable, 2000; 2008; Grable & Joo, 2004; Hanna & Lindamood, 2005).

Overall, researchers and professionals tend to agree that FRT is an important aspect in the context of financial decision-making as it plays an associative and predictive role with regards to financial-risk taking behaviors. But there is ambiguity about what constitutes FRT, which gives rise to inconsistent treatment of FRT and dissonance about FRT. Much of this inconsistency is likely due to a relaxed use of different risk-attitude terms that are often used interchangeably. Multiple similar risk-related concepts compound this inconsistency. For instance, in practice and in research, it is not uncommon for similar, but not equivalent, terms for FRT, such as risk preference, risk need, or risk perception to be confused or mistaken for one another and treated interchangeably. Nobre and Grable (2015) noted this problem and offered an informative clarification of the terminology for widely used risk attitudes such as financial risk tolerance, risk aversion, and risk preference, among others.

From a measurement viewpoint, such terms, even when similar, should not be used interchangeably for assessment purposes if they have clearly defined differences. Strictly speaking, if this is done, then the same construct is not being measured. It is possible that these similar concepts may even be highly correlated; yet that still does not signify their equivalence. Liu (2010) explained that defining the construct of interest with specificity is one of the crucial initial steps in the measurement and instrument development process. Expanding on this idea is that latent variables may also be

multifaceted or multidimensional. Researchers should consider this aspect when defining a construct, as it will be influenced by the dimensionality or number of domains that are included. And once the stage of actual construct measurement begins, it is pivotal to identify and explicitly state the particular dimension(s) that is being measured. This influence will apply regardless of which process or theory of measurement (e.g., Classical Test Theory, Rasch Measurement Theory, Item Response Theory) is being used (Croker & Algina, 2006; Liu, 2010). For the construct of FRT, several domains have been identified. For example, the literature lists investment risk, risk comfort and experience, and speculative risk as dimensions of FRT (Grable & Lytton, 1999; Kuzniak et al., 2015). However, not all FRT measures identify or explicitly disclose the dimensions that are being assessed.

For measurement purposes, when similar concepts are confounded, or latent variables of interest are inconsistently defined, or there is a failure to identify further dimension(s), the accuracy, consistency, and particularly the level of comparability is threatened in measurement resulting from tools that claim to measure the same latent construct. Specifically within the context of FRT, Ruiz-Menjivar, Blanco, Çopur, Gutter, and Gillen (2014) examined the comparability of three popular FRT measures in the field of financial planning (i.e., GL-FRT; Hanna, Gutter & Fan's [2001] improved version of Barsky, Juster, Kimball, & Shapiro [1997] FRT measure; and the FRT item from the Survey of Consumer Finances). Their results showed doubt over whether or not the evaluated instruments were measuring the same dimension of FRT. A possible explanation offered for the observed discrepancy on scores across instruments was the disagreement of FRT definitions.

Definition of financial risk tolerance and other risk related attitudes terms

The following definitions of FRT and other risk attitudes are mainly based on Nobre and Grable (2015). Specific attention is devoted to some of the similar, but not equivalent, definitions of FRT: risk aversion, risk preference, risk perception, risk capacity and risk composure.

Broadly, Cordell (2001) defined FRT as the level of uncertainty an individual is willing to take when making financial decisions. The International Standardized Organization (2005) further defined FRT as the willingness to undertake a less desirable outcome in the pursuit of a more desirable outcome. Nobre and Grable (2015) added even more specificity and context by stating that FRT refers to an individual's attitude towards financial risk, and the willingness to undertake more financial risk for the potential of obtaining higher returns. For the purpose of this project, this latter definition is used.

The term FRT is less popular among researchers and professionals in the field of economics; instead, the preferred term is risk aversion, which can be considered the inverse of FRT when viewed from the perspective of an expected utility theory (Barsky et al., 1997; Nobre & Grable, 2015). Thus, FRT and risk aversion are continuums laid in opposite directions. For example, those individuals having a high level of risk aversion would also be considered as having a low level of risk tolerance, and vice versa.[11]

The following terms are risk-related concepts that, again, although similar to FRT are not equivalent to it. Thus, substituting FRT for any of these terms should be avoided (Carr, 2014; Nobre, & Grable, 2015).

[11] The main audience for the present project is researchers and professionals working in financial planning and consumer economics; hence, FRT is used instead of risk aversion.

Risk preference refers to the overall feeling that an investor has towards the selection of an investment choice over another; this is regardless of whether the feeling is subjective or objective (Nobre & Grable, 2015). In other words, risk preference entails the rank order of attractiveness given to a certain set of investment choices.

Risk perception, on the other hand, refers to the cognitive evaluation and perceived attributed risk of a certain set of investment choices or alternatives. The main difference between risk preference and risk perception is that the latter is purely based on subjective grounds (Nobre & Grable, 2015). It is worth noting that risk perception is more fluid and subject to change than is risk preference.

Another risk-related term is risk capacity. This can be thought of as the ability to withstand a potential loss derived from a risky behavior or risky choice (Nobre & Grable, 2015). Risk capacity is a relatively stable concept as it is calculated primarily from elements that are fixed (at least in the short run), such as income and net-worth. Concrete examples of risk capacity measures are financial leverage rations, liquidity ratios, and savings ratios.

Risk need refers to the level of risk needed to achieve a particular goal (Nobre & Grable, 2015). Much of risk need depends on the conditions needed for achieving a desired short-term or long-term goal; thus, risk need is fairly fluid. Different level of risk need can be set for a client at a particular point of time.

Finally, Carr (2014) proposed a new term, risk composure (also known as risk appetite). This refers to the tendency or inclination of exhibiting a particular and constant type of risk-taking behavior when making financial decisions under uncertainty. This is the status quo equivalent in the realm of risk attitude concepts.

Section II. A Review of Theories of Measurement: Classical Test Theory and Rasch Measurement Theory

In this section, a description of Classical Test Theory (CTT)—a popular approach and analysis for developing and evaluating FRT scales in financial planning and consumer economics—is first presented. The CTT model and its assumptions and limitations are described. Then, Rasch Measurement Theory—a modern measurement paradigm and alternative method CTT—is explained. The explanation will entail Rasch Measurement Theory's background, its mathematical conceptualization, and its assumptions are presented. Finally, this section culminates with a comparison of CTT and Rasch Measurement Theory.

Classical Test Theory

The existence of errors in measurement, the notion and categorization of errors as random or systematic, the conventionalization of correlation, and a method for indexing errors were the main ideas that led to the birth of Classical Test Theory (CTT) (Traub, 1997; Croker & Algina, 2006). Charles Spearman, an English psychologist, is generally considered the founder of CTT. For more than a decade, Spearman presented logical and mathematical arguments that test scores were fallible measures limited by human traits. This led to his introducing the relationship between observed scores and true objective scores (i.e., unobservable true ability or a true score) by accounting for a random error term (Spearman 1904; 1907; 1913). After several repeated efforts to explain this relationship and other related issues, Spearman finally laid the foundation for CTT (Croker & Algina, 2006; Traub, 1997). And later, this theory and its model were further

explained by authors such as Guilford (1936), Gulliksen (1950), Lord and Novick (1968), and Magnusson (1967).[12]

In essence, the relationship between observed scores and true scores as proposed by Spearman is represented as follows:

$$X_i = T_i + e$$

Where X_i represents the observed score for person i, T_i represents the true ability or true score for person i, and e represents a systematic error (randomly distributed among all subjects within the population [i= 1, 2, 3,…,N]). In other words, CTT posits that a test score is a random observation of a person or subject's true ability. Since only X_i can be observed, though the goal is an estimation of T_i, X_i is used as a proxy. The precision and accuracy of this proxy, X_i (an estimation for T_i), depends much on the quantification and magnitude of e. There are many methods describing how to estimate and minimize e for a target population (e.g., reliability coefficient estimation procedures: test-retest methods, inter-rater reliability methods, and internal consistency methods). Regardless of the technique(s) used to estimate e, CTT relies on following key assumptions:

a) X_i is a total score, or the sum of all individual item scores. For example, if a FRT test consists of 3 items with a Likert-type scale ranging from 1 to 5, and the client response is 4, 5, and 3, respectively for each question; then, the X_i would be 12 (4+5+3).

b) e is randomly distributed among all subjects within a population. Hence, because e is estimated for the entire population, e is the same for all individuals in that population.

[12] The current conceptualization of CTT was codified by Novick (1966) and described in texts such as Lord and Novick (1968), and Allen and Yen (2002).

c) T_i (or true score) is represented in a range, and is not a fixed score. Therefore, T_i can only be described in terms of probability—not certainty. For instance, T_i can be expressed as a 95% confidence interval: $T_i = X_i \pm 1.96\ \sigma_{xx}$, where σ_{xx} is the standard error of measurement for the test. The value of error can be estimated depending on the source of error. Examples of procedures used for the estimation of coefficients for evaluating internal consistency or item reliability (that is, the degree to which different test items aiming to measure the same construct actually produce similar results) are Cronbach's alpha, split-half reliability, Kuder-Richarson (KR_{20}) method of rational equivalence, Cohen's Kappa coefficient, and others.

At this point, it is helpful to introduce two central concepts in CTT—reliability (also referred to as consistency or precision) and validity (accuracy). Messick (1981) explained that are crucial prerequisites for justifying the inferences to be drawn from test scores, and for defending the selection of a particular measurement test over alternative ones. In CTT, reliability is used to evaluate the consistency of X_i; while validity qualitatively examines whether or not X_i accurately represent T_i. Though CTT is by definition and fundamentally speaking a theory of reliability (the main purpose of CTT is to evaluate and develop the reliability of a test), this theory still has important implications for test validation: a test that is not reliable cannot be valid. Reliability is a necessary but not sufficient requirement for having validity (Boyle, 1991).

Reliability

Reliability refers to the repeatability of a measurement, or the degree to which test scores are a function of systematic sources of variance rather than error variance

(Thorndike & Hagen, 1961). In other words, reliability is concerned with the precision of of calculating test scores (X_i); specifically, with the consistency or reliability of scores across repeated application of the test. In general, tests with acceptable levels of reliability often have a lower degree of error than those instruments with less reliability.

The CTT approach to reliability is to quantify the errors arising from numerous sources, such as when there are different test forms, raters, or differences in test scores and items. In fact, CTT classifies reliability into different forms such as interrater reliability (i.e., the consistency of scores assigned by two independent raters), equivalent form reliability (i.e., comparability of scores from two distinctive versions of the same test), and test-retest reliability (i.e., the consistency of scores over time). In addition, if the concept of equivalent form reliability is extended to items, then an additional form of reliability is derived: internal consistency or inter-item reliability. This is in contrast to test formats as used in the equivalent forms.

But how are these forms of reliability quantified? Reliability is estimated using correlation coefficients, namely, reliability coefficients. Pearson correlations (for interval/ ratio level data) and Spearman correlations (ordinal level data) are reported for such reliability coefficients. Reliability coefficients range from 0 to 1. [13] A coefficient of value 1 indicates perfect reliability, whereas a reliability coefficient of 0 specifies that the test scores obtained on repeated administrations (that in the case of a test-retest reliability coefficient) are entirely unrelated, and thus unreliable.

For interrater reliability, a correlation coefficient is estimated. For example, this estimate may be between two sets of scores assigned by two different raters to the same

[13] For simplicity, negative correlations are ignored in this discussion

group of test takers or subjects. Similarly, test-retest reliability is estimated with correlation coefficients. But in this case, the estimation is between two sets of scores obtained by the same group of test takers for two sequential administrations of the same test. And for internal consistency, there are different procedures for quantifying errors. Consider the following comparison between the Kuder-Richardson Formula 20 (KR_{20}) and Cronbach's alpha.

KR_{20} is an estimation procedure for the calculation of the degree of internal consistency among a set of multiple-choice type of questions or dichotomous questions. However, a more widely used and accepted reliability measure for internal consistency is Crobach's alpha—or simply, the alpha coefficient. In fact, within the context of financial planning, Cronbach's alpha is certainly the most popular and frequently reported indicator of reliability among academicians and researchers (Bond & Fox, 2007; Nunnally, 1967). This is understandable considering the greater versatility that Cronbach's alpha affords when compared to the more one-dimensional limitations that are seen with the KR_{20} formula. In essence, the alpha coefficient is a more generalized form of KR_{20}, and is one that is more suitable for different item formats (multiple choice, constructed-responses, Liker-type scale, dichotomous, etc.). In short, an alpha coefficient is the averaged correlation coefficient among all possible pairs of items of a test. The mathematical representation of Cronbach's alpha is as follows:

$$\alpha = \frac{K}{K-1}\left[1 - \frac{\sum_1^k \sigma_{i^2}}{\sigma_{x^2}}\right]$$

Where *k* is the total number of items on the test, σ_{i^2} represents the squared standard deviation or variance of subjects' scores for item j (j=1, 2,..,N), and σ_{x^2} is the squared standard deviation or variance of subjects' scores for the entire test. Thus,

Cronbach's alpha can be thought of as the percentage of variance in a sample of respondents' scores due to the covariation among items produced by subjects' true abilities (Liu, 2010). The alpha coefficient ranges from 0 to 1, where a coefficient closer to 1 is indicative of good internal reliability (i.e., possibly reflecting greater intercorrelation of test items). Yet, this calls into question actually how near to 1 the coefficient must be to conclude that it is actually good or has an acceptable reliability. There are several guidelines and cutoffs for assessing whether or not a coefficient indicates good reliability or poor reliability. Despite these guidelines, it should be noted that there is no universally accepted cutoff for assessing reliability, and often the "gold standard" varies with varying disciplines.

Nunnally (1967) stated that a reliability coefficient of 0.90 and above is deemed acceptable, especially when the instrument is employed to make decisions about individuals. Following the recommendations in Saad, Carter, Rothenberg, and Israelson (1999), researchers in the area of financial planning and consumer economics have generally accepted reliability coefficients of 0.70-0.79 as acceptable. In a broad sense, research in financial planning and consumer economics has adopted the standards from the field of psychology where a coefficient alpha of at least 0.70 is deemed to be a good indicator. Moreover, scholars in financial planning often cite Boyle (1991) to argue that generally, scores below $\alpha = 0.70$ are considered useful only in exploratory studies, while scores above $\alpha = 0.90$ are considered problematic due to item redundancy.

Validity

The other central concept in Classical Test Theory is validity—often thought as a synonym for accuracy. Validity refers to the notion of how well the suggested use of the

test scores compare to and conform with the actual intended purpose of a test. Essentially, a test might be valid for one particular purpose, yet be unsatisfactory or invalid for other purposes. Thus, validity offers some insight into how satisfactory or unsatisfactory a test might be when used for a given purpose.

It is worth noting that a test that is not reliable cannot be valid. Reliability is a necessary but not a sufficient requirement for having validity (Boyle, 1991). This is a crucial point to understand, as utilizing a test even with the highest possible reliability will not automatically generate validity. On the other hand, validity leads to reliability (Croker & Algina, 2006)

The process of establishing validity for a particular instrument is called validation. It is important to note that when assessing a measurement's validity, one is validating test scores rather than the items of measurement or the instrument itself (Hattie, Jaeger, & Bond, 1999). Cronbach (1971) explained validation as that evidence collection process that supports the type of inferences that will be drawn and derived from test scores (X_i). In preparation of a validation study, the desired inferences must first be delineated. Once these inferences are specified, an empirical study is then orchestrated to collect evidence of the utility of the observed scores for such identified inferences (Croker & Algina, 2006). The validation process yields studies that can be classified into three main types: content validation, criterion-related validation, and construct validation.

The main objective of a content validity study is to evaluate whether or not the items (i.e., questions) included in a test effectively and accurately represent the construct of interest that is intended to be measured. Examples of the steps that Croker and Algina (2006) suggest considering in a content validity study are "defining the domain of interest

for the latent variable, selecting a panel of qualified experts in the construct domain, [and] providing a structured framework for the process of matching items to the construct domain of interest" (p. 218). It is standard for a panel of content experts to gather together to review whether or not a preselected set of test domains represents adequate coverage of the construct domain that is being measured. Similarly, experts may decide to revise items or questions in a test to determine whether or not they represent a suitable sampling, and are, indeed, a true bank of indicators to measure the construct of interest. Here the relative nature of validity can be seen.

A second type of study for establishing validity involves evaluating criterion-related validity. This refers to the process of collecting and providing evidence of an association between a test score and a criterion measure (e.g., relevant reference measure or behavior/performance). The first step in a criterion-related validation typically involves the clear identification of a criterion (or reference) variable, and an available and valid measure(s) for such a variable. Then, the correlation between the scores obtained from the criterion measure, and the scores obtained from the instrument under validation is computed. When there is a statistically significant relationship, this evidence indicates that there is support for the establishment of validity for the measure under study. Within criterion-related validity, two distinct categories of validation evidence exists: predictive validity, and concurrent validity. If the scores from the criterion measure are collected after the collection of scores from the instrument under validation, then the criterion-related evidence is referred as predictive validity. Conversely, if the data from both measures is collected at nearly the same time, then the criterion-related evidence is denoted as concurrent validity.

Similar to criterion-related validity (and more specifically concurrent criterion-related validity) are convergent validity and divergent/discriminant validity. Convergent validity is simply concurrent validity that has a positive, statistically significant correlation between the scores of the criterion measure and the scores from the measure under validation. In contrast, divergent/discriminant validity refers to concurrent criterion-related validity when there is a statistically significant, negative correlation between the criterion measure and the measure under validation. Consider the case for FRT. Typically, criterion-related validity is assessed by comparing how well test results conform to actual risk-attitude behaviors. An example of one such behavior is investment in risky assets. A measure for this reference or criterion behavior (investment in risky assets) could be the percentage of equity allocation over total assets in a client's portfolio. Thus, in the process of validation, the scores from the test under validation would be compared with the percentage of investment in equities. Based on financial risk theories, one would expect that individuals who exhibit high-risk tolerance on the test would hold riskier assets (i.e., more stocks or high risk securities than bonds or more conservative risk securities). If such as positive correlation is observed, then there is evidence to support convergent criterion-related validity for the measure that has been established (assuming that the data for both the reference behavior and the measure were collected at or about the same time). And, to reiterate an earlier point, this validation assumes, and requires, test reliability.

And finally, a last validation study for consideration is in establishing construct validity. Construct validity refers to the notion that the test is successful in capturing its intended phenomena. Multiple methods for assessing construct validation exist (e.g.,

differentiation between groups, common factor analysis[14], and a Multitrait-Multimethod matrix) (Croker & Algina, 2006; Campbell & Fiske, 1959).[15] In financial planning and consumer economics, factor analysis (and more precisely exploratory factor analysis) has been the popular method for assessing construct validity (Grable & Lytton, 1999; Grable, Archuleta, & Nazarinia, 2010). This is performed to evaluate the matrix with the correlations among items (i.e., questions in an instrument). The main idea is to identify factors based on distinctive correlation patterns. Each factor (ξ_i, i=1, 2,.,N) represents a set of items that correlate highly. These high correlations are thought to measure some latent trait or dimension of the construct of interest.

Issues to consider when using Classical Test Theory

CTT is practical and uses a conceptually simple framework in measurement; however, its lower level of technical complexity comes with a set of limitations (Liu, 2010). The following are the main limitations and critiques of CTT in the extant literature: (a) T_i is dependent on X_i; (b) e is averaged over a population and assumed the same for every subject; and (c) neither T_i nor X_i is interval. Understanding limitations or searching for areas for improvement or refinement are necessary part of scientific endeavors and this certainly holds true for measurement. Searching for improvement upon limitations naturally inherent within CTT has been done herein, and below these

[14] Note that common factor analysis should not be confused with Principal Component Analysis. The following is a comprehensive list of papers that elaborate on the differences between principal component analysis and factor analysis: Bentler & Kano, 1990; Ford, MacCallum & Tait, 1986; Gorsuch, 1990; Loehlin, 1990; MacCallum & Tucker, 1991; Mulaik, 1990; Fabrigar, Wegener, MacCallum, & Strahan, 1999; and Suhr, 2009).

will be briefly explained, along with their implications such that areas of refinement can then be considered.

Issue of T_i being dependent upon X_i. The mutual dependence of test scores on items (e.g., the questions used in a FRT test) needs consideration. By T_i being dependent on X_i, there is the implication that T_i for a particular subject is linked to the entire measurement instrument via observed scores (X_i). So, any changes in an instrument, either by removing or adding items, then results in changes in T_i. As this is counterintuitive that the T_i (true scores or true ability) for subjects should change from one test to another—assuming both tests are valid (Bond & Fox, 2007; Harvey & Hammer, 1999; Liu, 2010; Lord & Novick, 1968; 1980), then this cues a potential area for improvement. It is desirable to be able to remove and add items to an instrument without affecting T_i.

This issue of true test scores being linked to observed scores also implies that items' statistics would become sample dependent. Yet, this limitation can be improved upon, as it is actually the inverse of the relationship noted above that would be expected. That is, it would be desirable for an instrument to measure true scores regardless of which subjects are tested or the subsequent scores that are observed. For example, reliability estimates tend to be higher in heterogonous groups than in homogenous groups, so having a versatile instrument is beneficial, as the properties of the true score should not change based upon who is being measured. In fact, an ideal instrument should be stable across different groups.

To elaborate, consider the following example to help illustrate this limitation. It comes from the physical sciences, and is based on temperature and a thermometer's

scale. Imagine a large bucket of water sitting on a table in a room with constant ambient temperature. The temperature of the water in the bucket has some true value, measured in degrees Celsius. The temperature of the water in the bucket may be considered an intrinsic value here. It would not be expected that the temperature of the water would change or be altered by merely the type of thermometer being used (that assuming the thermometers are valid instruments to measure temperature). Then, both an alcohol-thermometer and mercury (Hg)-thermometer should be equally handy at revealing a true score (e.g., temperature) intrinsic to the water at a given time. And, if the scale was changed from Celsius to Fahrenheit on the instruments, the temperature of the water in the bucket is still not expected to vary. The same expectations would hold if these same thermometers were used to measure other liquids, such as a bucket full of oil or a bucket full of liquid mercury. It is desirable that the thermometers measure the true temperature of the various liquids being tested. The measures (e.g., thermometer readings) represent properties (e.g., temperature) of various objects/subjects (e.g., water, mercury, oil).

The thermometer is analogous to a test or tool in psychological measurement, and the type of thermometer (e.g., alcohol or Hg) is analogous to various psychological tests that might be used to do the measuring. The subject matter or object (e.g., water, oil, and mercury in buckets) whose properties (i.e., temperature) are being queried is analogous to individual subjects in the psychological test. As with the above example, it would not be expected that the individuals' properties would vary or change dependent upon which psychological test is administered.

Consider the following illustration to further elaborate upon improvements for the context of FRT. Assume there are two validated tests measuring FRT, A and B, and each

test has merely two questions each having rating scales from 1 to 4, where 1 indicates low FRT and 4 indicates high FRT. Test A contains items that are very easily endorsed, or agreed with. So, it should be almost expected that test A might yield a score that assumes that a subject has a willingness to take risk even when the subject is not located on the higher end of the FRT domain (e.g., an inflated risk). Due to the ease of e*ndorsablity* of questions in Test A, a client with an actual low-risk tolerance (T_i) could potentially still select 4 on the scale. In this scenario, the easy to endorse risk questions lead to observed scores (X_i, the test scores for FRT) that will be artificially inflated. The opposite would hold true for a scenario where Test A uses items that are very difficult to endorse, and could produced an artificially deflated risk tolerance. Consequently, if T_i is linked to X_i, and in this sense dependent upon it, the true scores (T_i) are then attributed risk tolerance levels arising from the test itself. To further consider how this can be problematical, consider the following.

Now, if test B is administered to the same client, and such a test contains hard to endorse questions or hard to agree questions, say much harder than test A, then the observed test scores would be expected to be lower than that which was seen in test A. If Test B did yield a lower score, the level of risk tolerance (T_i) assigned to this particular person would be lower as well. Thus, in this example, the level of risk tolerance (T_i) for this client would change based on which test is administered. As the previous example showed with the alcohol thermometer and mercury thermometer, the intrinsic temperature of water was not expected to change. But, here, depending on which test is used, Test A or Test B, different values of T_i are obtained. Intuition suggests that this should not be the case for an ideal instrument, including those calculating FRT. Hence,

an improvement for FRT will be to unlink the two variables such that T_i is not dependent upon X_i. By doing so, the statistical basis for calculating FRT will be improved.

Issue of *e* being assumed constant for each subject in the population. CTT assumes the error of an entire population is a particular value, and this value is assigned to all individuals within the population. This is an efficient method, especially when calculating very large samples. However, the tradeoff for this efficiency is loss of ability to individualize the error for each subject. This holds true regardless of the instrument used that generates the error. Individual subjects have a unique value or true score, and an observed or measured value having its own unique error. So, an improvement in calculating FRT, would be to take this individualized error into account. In addition, some subjects or respondents may react differently to the same instruments or tests, or some subjects could be affected uniquely by the test itself; and all of these effects may not be uniform across all subjects. At an extreme, some subjects may be more reactive or volatile to the test itself. As such, each subject's unique affectedness could lead to a potentially large range or variability; and, a variation that may not necessarily be uniform. This would also lead to potentially large variation in error associated with each of these values. Can it be assumed that the test has a flat error for each individual based only on the individual's score?

In the thermometer example, the alcohol-thermometer and Hg-thermometer both have their own unique built in errors specific to them. Perhaps the thermometer's glass case is imperfectly shaped such that its volume is not uniform throughout its length causing the internal alcohol or mercury to rise unevenly and create error in measurement. Perhaps the imperfections in the glass are numerous, small, and random; it might be

possible the errors here are flat or could be assigned even across the measures. It could be that one or two places along the thermometer are severely warped, and these errors might have large variation. Our flawed thermometer here might even reliably or repeatedly inflate the temperature measurement.

Or, perhaps the hash marks for the temperature scale are not painted perfectly horizontal on the thermometer and do not line up uniformly along the chamber. Here the measure of error would be calculable, and even corrected if the angle of the hash were known. If the hash marks were all inconsistently painted on haphazardly, the error, becomes harder to predict. It might even yield consistent temperatures of say, 80 degrees Fahrenheit, across numerous buckets of water and oil when the true temperature is 79 for all of them. But suppose our thermometers are designed to measure temperatures from 75-80, and can measure into tenths of a degree. Here a small flaw in the thermometer would lead to a large variation in measurement and error in the same scenario.

Consider another phenomena of measurement and error that is less intuitive. Imagine a 5 gallon bucket of water, and a standard drugstore bought thermometer; and suppose the thermometer is warm, and the water is cold. Dipping the thermometer into the bucket is not likely to change the water's temperature, or at least doing so would likely be negligible or undetectable given the instrument used for the measurement. But, now assume the bucket were smaller, say the size of a tiny thimble holding only one drop of cool water, and assume our same thermometer was very warm. Here, the thermometer, itself, can cause a quite noticeable change the temperature of water due to heat exchange. Or, in other words, the test instrument itself is causing a change in the subject that is

being tested. The test itself leads to biasing the results that were meant to be tested (i.e., the original T_i before effected by test measurement).

Complicating matters, consider how unique subjects may react differently to a test. For example, both water and mercury each have unique heat exchange properties with mercury being > 100 times more thermally conductive (e.g., the amount of heat that can pass through it over time) and water having 50% more heat capacity (e.g., the amount of heat needed to raise its temperature). In the thimble example above, thimbles of mercury and water would each experience variation in the degree to which they were affected by the thermometer—that is, each might have not only variation in temperature, but might have a different error, or a variation in their respective margins of error, and even other variations (such as at different ambient temperatures or barometric pressures). The same holds true for testing abstract subjects with test instruments. It has been posited that the idea of *e* being common to all subjects within the population is counterintuitive and potentially harms the precision of measures for different subjects or respondents (Hambleton & Swaminathan, 1984; Liu, 2010). This is true in the case of CTT as well, and attempting to account for these individual errors may lead to improvement in measurement. And, here, attempting to overcome this limitation may lead to more precision in the measurement of FRT.

Issue that neither T_i nor X_i is interval, but ordinal. Inferential statistics, such as a *t* test or *F* test, assumes that data to be analyzed are measured at the interval level (Mitchell & Jolley, 2012). Thus, theoretically and strictly speaking, this aspect of CTT limits the application of inferential statistics to ordinal measurement data based on CTT. And any strict conclusions that can be made about ordinal scores will be limited to rank

order only. In practice, though, researchers treat measurement data, based on CTT, as it were interval data rather than ordinal data, in an effort to apply inferential statistics. It is important to note that doing so results in a treatment yield that may decrease in statistical power such that the test may not reject the null hypothesis when it should, and thus comes with the problem of increasing the probability of a Type II error.[16]

For example, assume a four-item FRT test, in which each item with a response format of five hypothetical statements about willingness to take risk are presented (for coding and analysis purposes, the first category is coded as 1 and it is indicative of very low FRT, 2 low FRT, 3 average, 4 high, and 5 very high FRT). The possible total test score ranges from 4 to 20. The cutoff levels determined by a panel of experts is as follows: 4-8, very low risk tolerance; 9-12 low risk tolerance; 13-16, high-risk tolerance; and 17-20 very high-risk tolerance. Further, suppose the test is administered to four clients, A, B, C and D. The FRT test scores for the clients are 18, 15, 10, and 7 respectively. If the inferential statistical assumptions of interval data are not violated, the only conclusion that can be reached here is exclusively in terms of rank order given by the scores. For example, A scored higher than B, C and D; or D scored lower than all three others. Rank order has proven itself exceedingly useful. However, because this is ordinal data, one cannot subtract the raw test results and obtain a difference between client A and B such that A-B = 3 and expect this 3 to assist in way other than in the ordering (hence, ordinal numbers) of the clients. The value of 3 in that instance has no intrinsic value. The same is true of a value of 3 obtained through subtraction of C and D.

[16] A line of research has suggested that the departure from the "intervalness" requirement does not appear to substantially impact Type I and Type II errors (Binder, 1984; Jaccard, & Wan, 1996; Labovitz, 1967; 1970; Zumbo, & Zimmerman, 1993). This issue remains a topic of debate in social sciences research and practice.

And, it is not possible to compare the two values of the 3 for any use other than in ordering.

Rasch Measurement Theory

Problems and limitations with basic models in science have led to the revised conceptualization and development of improved models that attempt to diminish shortcomings and maximize potential outcomes in measurement. The CTT model has been widely used due to its simplicity, but CTT's assumptions, especially for testing, are somewhat problematic. This has led to the development of alternative measurement models—and to what Brennan (1998) referred to as to the liberalization of CTT. In the last five decades, there has been a steady trend for the psychometric discipline to continue to try to offer both an increasingly pragmatic view (i.e., The Scaling Tradition view) and a model that is theoretically stronger (e.g., Item Response Models (IRT) and Rasch Models) when targeting measures of latent constructs. Rasch Measurement Theory, and its probabilistic model, is one such alternative.

The Danish mathematician, Georg Rasch developed the basic logistic model in his pioneering work (1960/1980). Essentially, the Rasch model proposes that a given construct of interest (e.g., FRT) can be thought of as a linear continuum. This allows for a common measure to be applied to items (e.g., questions included in the scale) and subjects/objects (e.g., respondents)—both of which are mutually independent. Specific properties of items and subjects (namely, item and person parameters, respectively) are evaluated to determine their location along a continuum. For item parameter, the property evaluated is the difficulty of items. Particularly, when measuring attitudinal constructs, as FRT, item difficulty is called item *endorsability.* This refers to item location on the

continuum. For example, a very hard to endorse item would be located on the high area of the continuum or line. For interpretation purposes, it can be said that a person with higher ability (level of FRT) is more likely to endorse or agree with the response category indicative of high FRT for this particular high-to-endorse item. Thus, *endorsability* is used to locate an item somewhere upon the line. For respondents, the subject's ability, or *agreeability* of selecting a particular response choice for items, is used to also place the respondent somewhere upon this line.

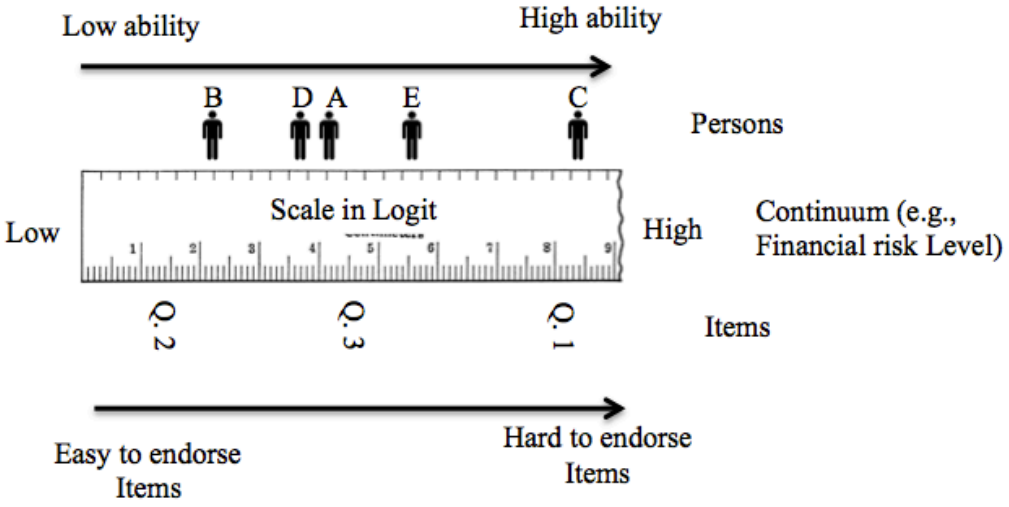


*Figure 2***.** Graphical Representation of the Main Idea with Parameters Proposed by Georg: A Ruler (Continuum) Common to Persons' abilities and Items' *endorsabilities*.

For a better grounding on the foundation of Rasch Model, visualize the following example. Imagine a 3-question instrument that measures one specific dimension of FRT. Each question has 4 response categories, 1 being indicative of a low FRT and 4 being indicative of a high FRT The test is administered to five subjects (A, B, C, D and E). Figure 2 illustrates the idea that the respondents are placed along the continuum based upon their abilities, and items based upon their respective *endorsabilities* or difficulties.

As shown in Figure 2, FRT is measured in some analogous universal metric of distance, such as centimeters (or logits—explained later). The ruler allows for measurement in only one direction (i.e., it may only increase or decrease in such direction) Note, upon this line, continuum, or *uni-dimesion*, or however it might be called, one will find both subjects and the items. Moreover, the line is shared—or common—to both subjects (e.g., persons who completed the administered test) and items (e.g., questions within the test). Each person is mapped along the continuum based upon particular exhibited levels of ability. Ability in this particular example refers to the level of risk tolerance displayed by each subject. Recall that this ability is determined based on the *agreeability* to select one response choice over another for a particular question. Respondents with a higher level of risk tolerance will be located on the higher end of the continuum or the line.

For this example, person C has a high level of FRT, and is mapped accordingly, as are the others. Similarly, each item of the instrument is mapped along the continuum based upon the item's particular level of difficulty or *endorsability*. Again, recall the interchangeability of the terms difficulty and *endorsability*. Because the example used the attitudinal construct of FRT *endorsability* is the proper term here.

From Figure 2, it can be seen that question 1 is considered hard to endorse. Those persons possessing higher ability or a higher level of FRT have a greater probability of answering this question such that it reflects where they are positioned on the ruler. Thus, person C has a higher probability of answering question 1 in a way that reflects a greater willingness to take financial risk. On the other hand, question 2 sits much further to the left, in the region where the easy to endorse items are found. This can be interpreted as following: even a person with low ability or a lower level of FRT—Person B for

example—would have a high probability of agreeing with or selecting the response categories indicative of more willingness to take financial risk. In this sense, person B has a higher probability of selecting response category of 3 and 4; this would not be considered an inconsistent response pattern as it might be in CTT. Finally, an aspect worth noting is that items and persons are mutually independent and the relationship between both is used only to create a line. Thus, the estimation of subjects' ability or FRT, in this case, is not dependent on the items presented to them. Similarly, the location of the items on the continuum, whether they are easy or hard to endorse, does not depend on the particular ability exhibited by the sample utilized. This is a major advantage of Rasch and (IRT models in general) over CTT.

As mentioned earlier, Rasch models belong to a set of probabilistic models developed for the purpose of describing response patterns of subjects to individual items. From the Rasch perspective, the probability of a particular respondent answering an item in a particular fashion is determined by the difference between the ability or agreeability, and difficulty or *endorsability* measures. The higher the respondent's agreeability for an item, the higher will be their probability to respond to the question (i.e., to select the response category that reflects a higher point on the line, for example). Likewise, the harder it is to endorse a question (i.e., when the item is located in the higher region of the continuum), the lower will be the likelihood for a respondent to select a response category indicative of the higher region (i.e., less *endorsability* for the higher region of the latent construct being measured).

An important term within this framework is Rasch calibration or modeling. This refers to the application of the Rasch model to a set of items that distinctively define a

stable line for the construct of interest. Rasch calibration can be thought of as analogous to creating or constructing a yardstick or meter stick (or in Rasch terms, a logit scale), similar to the previous example with the ruler. Once a set of evaluated items has been calibrated to define a linear measure, these questions can then be used with any sample within the population to estimate the ability measures for each surveyed subject. Unlike CTT, the ability measures obtained with Rasch are (a) truly interval, (b) independent from the set of items used in the test,[17] and (c) produce individual errors for each subject. Thus, Rasch Measurement theory overcomes the three limitations mentioned above in the earlier CTT section.

In order to explain the properties of the Rasch Measurement Theory, a more detailed description is offered below. As a start, the mathematical representation of the basic dichotomous Rasch model[18] will be presented. The probability of response X_{ni} = 1 (the probability of selecting the response category indicative of high risk tolerance when high risk tolerance is coded 1 and low-risk tolerance 0) can be represented as follows[19]:

$$P_{ni1}(X_{ni} = 1 | \theta_n, \delta_i) = \frac{e^{(\theta_n - \delta_i)}}{1 + e^{(\theta_n - \delta_i)}} \qquad (1)$$

Where P_{ni1} is the probability of person *n* with ability θ_n answering X_{ni} = 1 for item *i* with difficulty or *endorsability* δ_i.

Conversely, the probability of response X_{ni}=0 is as follows:

$$P_{ni0}(X_{ni} = 0 | \theta_n, \delta_i) = \frac{1}{1 + e^{(\theta_n - \delta_i)}} \qquad (2)$$

[17] Once a set of items has been calibrated to define a stable linear measure, respondents do not necessarily have to exactly answer the same items (an item or subset of items for reference will be needed only) to obtain stable and comparable measures.
[18] For simplicity purposes, the dichotomous (X= 0 or 1; or equivalent to true/false, yes/no options or 2 response categories indicative of high/low FRT for a question) is used in this illustration.
[19] The following formulas were obtained from Engelhard (2013) and Liu (2010).

Where P_{ni0} is the probability of person *n* with ability θ_n answering X_{ni}=0 for item *i* with difficulty or *endorsability* δ_i.

Now, these expressions will represent the likelihood or odds of person *n* answering item *i* as X_{ni}=1. Note that the likelihood or log-odds for an event is the ratio of the probability of an event happening (P_{ni1}) over the probability of not happening (P_{ni0} or $1 - P_{ni1}$). Thus,

$$Odds = \frac{P_{ni1}}{P_{ni0}} = \frac{\frac{e^{(\theta_n - \delta_i)}}{1 + e^{(\theta_n - \delta_i)}}}{\frac{1}{1 + e^{(\theta_n - \delta_i)}}} = e^{(\theta_n - \delta_i)} \quad (3)$$

Further, if *L* represents the natural logarithm of the likelihood or odds, which is called the logit (or log-likelihood, or log-odds), it is possible to obtain the following:

$$L = \ln \frac{P_{ni1}}{P_{ni0}} = (\theta_n - \delta_i) \quad (4)$$

Equation (4) shows the mathematical representation model of the dichotomous Rasch Model in logit. Thus, the logit (i.e., the unit of measurement for the ability of a person and for *endorsability* of the item) for a person *n* to answer X_{ni}=1 for question *i* is simply the difference between θ_n and δ_i, where θ_n is the ability of person *n*, and δ_i is the *endorsability* of item *i*. The larger the difference between θ_n and δ_i, the higher the likelihood that person *n* will respond X_{ni}=1for item *i*.

Important properties can be described about θ_n (ability parameter) and δ_i (item parameter), as seen in Equation (4). First θ_n and δ_i are on a true interval scale (i.e., logit units), which, in fact, highlights a solution to one of the limitations of CTT when applying the Rasch model to measurement data. Both the ability and item parameters possess the property of linearity, which allows for making direct comparisons between ability measures, or between item measures independent of one another.

This can be shown as follows for ability parameters:

$$L_1 - L_2 = (\theta_1 - \delta_i) - (\theta_2 - \delta_i) = \theta_1 - \theta_2 \quad (5)$$

And as follows for item parameters:

$$L_1 - L_2 = (\theta_n - \delta_1) - (\theta_2 - \delta_2) = \delta_2 - \delta_1 \quad (6)$$

Equation (5) demonstrates that the difference in ability measures is determined by the differences in log-odds, regardless of the item parameter (i.e., difficulty or *endorsability* of item *i*). In Equation (6) the same notion is seen, only in terms of item parameters. Examining these two equations as written, it is readily noticeable that there is a mutual independence relationship that exists between the item and ability parameters. In Rasch language, this is called the item invariance property for the item parameter, and the person invariance property for the ability parameter.

A second important property of θ_n and δ_i is that both are latent variables; or restated, they are not raw data nor direct observations as seen in CTT. In Rasch modeling, the only direct observations are the scores obtained from the X_{ni}'s (i.e., the response patterns). As noted in Birnbaum (1968), the raw scores are sufficient statistics to estimate the parameters for both θ_n and δ_i. These parameters may be computed using various approaches. For example, conditional maximum likelihood estimation (CMLE) is a method specific to the Rasch model (Baker, 2010). This method was developed by Andersen (1972; 1973). In brief, CMLE is applied to item difficulties first. Once the items are calibrated, the ability parameter estimate for each raw score can be obtained by using maximum likelihood via an iterative process, typically the multiparameter Newton-Raphson method (Baker & Kim, 2010). In that sense, the item parameter estimates are then those that maximize the likelihood of the data observed in the raw scores for

persons. Because person abilities are conditioned out of the item difficulty estimation, ability estimates are called incidental parameters, and item difficulties are structural parameters (Linacre, 2004). Another method is the Joint Maximum Likelihood Estimation (JMLE), also known as Unconditional Maximum Likelihood Estimation (UCON). Wright and Panchapakesan (1969) formulated and proposed JMLE as a procedure for sample-free item analysis (i.e., item invariant parameter estimation or item measure estimates that are not dependent on the sample utilized for the estimation). Wright and Panchapakesan posited that the estimation of the item and ability parameters occur when the observed raw score for the parameter is congruent with the expected raw score. Unlike CMLE, estimates for both ability and item are obtained simultaneously.

Unidimensionality

The property of linearity for the parameters θ_n and δ_i was discussed earlier. This property conveys a major assumption made within Rasch modeling; that is, *unidimensionality*. In this sense, the item and ability estimates are *unidimensional* measures that increase (or decrease) in one direction only. This notion of *unidimensionality*, or the progression of a line in only one direction, is crucial to note because it serves as a theoretical foundation for the creation of stable and well-marked "yardstick" or *unidimensional* continuum for the construct of study. Thus, a measurement scale (from the Rasch perspective) describes one and only one dimension or attribute of the construct that it is being measured. In terms of FRT, for instance, it is important that one dimension of the construct, say speculative risk or investment risk, be evaluated and modeling separately. Note though, that the instrument (e.g., survey) can contain more than one scale

Requirements for invariant measurement

Before discussing the requirements for invariant measurement, it is first important to describe the idea of invariant measurement. Engelhard (2013) explained that invariant measurement refers to the philosophical approach to measurement where scales have desirable properties, such as the ability to independently map items and persons on a continuum, the independence of item from persons, and vice versa. Rasch Measurement Theory provides a workable framework that, under appropriate conditions, can yield an invariant measurement. According to Engelhard, in order to develop useful, stable objective measures[20] that approximate ideal-type scales that adequately present measures of the latent variable and constructs, the following main requirements must be met:

1. "The measurement of person must be independent of the particular items that are being used for it measurement" (p. 14). This is referred to as item-invariant measurement of persons.
2. "A more able person must always have a better chance of success on an item than a less able person" (p.14).
3. "The calibration of items must be independent of the particular persons used for calibration" (p. 14).
4. "Any person must have a better chance of success on any easy item than on a more difficult item" (p.14).
5. "Item and person must be simultaneously located on a single underlying latent variable" (p.14).

[20] The Institute for Objective Measures (IOM) defines "objective measurement" as the repetition of a unit amount that maintains its size, within an allowable range of error, no matter which instrument, intended to measure the variable of interest, is used and no matter who or what relevant person or thing is measured (IOM, 2000).

Section III: The Grable and Lytton (1999) Financial Risk Tolerance Scale

Background

Grable and Lytton's (1999) financial risk tolerance measure (GL-FRT) is a thirteen item scale, mixed assessment of risk tolerance that contains questions on "guaranteed versus probable gambles, general risk choice, choice between sure loss and sure gain, risk as related to experience and knowledge, risk as a level of comfort, speculative risk, prospect theory, and investment risk" (p.174) (See Appendix A for the complete instrument and scoring instructions). The GL-FRT covers three main dimensions of the FRT construct: investment risk, risk comfort and experience, and speculative risk.

The GL-FRT has been available at no cost through Rutgers New Jersey Agricultural Experiment Station since 2003 (https://njaes.rutgers.edu/money/riskquiz/). Since the year of its inclusion on the web until mid 2014, more than 200,000 consumers, educators, and researchers have used the scale. Likewise, financial and investing planning firms have adopted the scale to assess their clienteles' FRT (Kuzniak et al., 2015). Domestic and international academicians in the area of financial planning and consumer economics have also used the GL-FRT. The scale has been formally referenced in more than two hundred research publications (Google, 2014).

Grable and Lytton first published the scale in 1999. In their paper, the authors described the steps taken in the scale's creation and evaluation. Though they did not formally acknowledge the measurement paradigm used in that paper, their reported psychometric analyses and evaluations corresponded particularly well to Classical Test Theory analyses. Grable and Lytton (1999) acknowledged that their scale was developed

through guidance provided from MacCrimmon and Wehrung (1986). They followed aspects of these guidelines that were used to ensure a proper assessment of FRT by using a combination of simple and complex hypothetical based questions and avoidance of redundant items. Their purpose was to provide a content tool that was comprehensive yet parsimonious in terms of length.

Reliability

The reliability for the G&L-FRT scale has been measured using Cronbach's alpha. In Grable and Lytton (1999), the alpha reported was of 0.75. Based on Peterson (1994), this level of reliability is consistent with the accepted cutoff for alpha coefficients in psychological and marketing studies. Peterson (1994) noted that the average reported Cronbach's alpha in the psychological and marketing literature ranges from 0.76 to 0.77. In 2004, Yang evaluated the reliability of the GL-FRT instrument using two samples: one composed of college students and another one comprised of adults. The author noted that for both samples, the Cronbach's alpha coefficient was greater than $\alpha = 0.70$. Yang concluded was that the GL-FRT worked well, overall, for both younger and older respondents. More recently, Kuzniak and associates (2015) reviewed the GL-FRT in its fifteenth anniversary. The reported Cronbach's alpha during this time was 0.77; ranging from 0.73 to 0.90 across different groups.

Validity

Content-related validity

In the initial steps of the development of GL-FRT, Grable and Lytton (1999) reported selecting more than 100 items related to the "willingness to take risk" from the literature. In an effort to develop a measure that would comprehensively capture the FRT

construct, the authors grouped items into different categories such as, guaranteed gambles versus probable gambles, general risk choice, choice between sure loss and sure gain, risk as experience and knowledge, risk as a level of comfort, speculative risk, prospect theory, and investment risk. Although, it is not clear whether o these items were reviewed by a panel of experts at any point before or after the development of the scale, the GL-FRT has been referenced and adopted for use by many academicians and professionals in the field of financial planning and consumer economics. This, in itself, might be considered as qualitative evidence of content-validity.

Criterion-Related Validity

In 2003, Grable and Lytton presented a criterion-related validity study. In particular, the authors aimed to establish concurrent criterion-related validity. Convergent validity was established by documenting a statistically significant and positive correlation between scores from GR-FRT and the reference measure of equity ownership. Similarly, evidence for divergent, discriminant concurrent-related validity was shown by presenting a negative significant correlation between GL-FRT scores and fixed-income/cash ownership. Furthermore, Grable and Schumm (2010) reported a statistically significant correlation of 0.60 between the GL-FRT score and scores on the FRT item from the Survey of Consumer Finances (SCF). In that same year, Gilliam, Chatterjee, and Grable (2010) conducted a concurrent criterion-related validity test of the GL-FRT by correlating the GL-FRT measure with both the SCF's item on risk, and with the ownership of risk investment assets. Consistent with previous results, Gilliam and associates found evidence for convergent criterion-related validity of the GL-FRT

measure. Finally, in a more recent study, Kuzniak et al. (2015) presented new evidence for criterion-related validity.

Construct-Related Validity

In 1999, Grable and Lytton conducted a principal component analysis (PCA)[21] to determine the number of factors captured by the scale. They concluded that three extracted factors (i.e., investment risk, risk comfort and experience, and speculative risk) adequately measured dimensions of the multifaceted FRT concept.

Summary of Chapter Two

This chapter has provided a review of the definition of FRT and other risk-related concepts utilized in the area of financial planning and consumer economics. Furthermore, it has provided a description of the properties, mathematical models and assumptions for two widely used theoretical frameworks (i.e., Classical Test Theory and Rasch Measurement Theory) in the measurement of latent constructs, such as FRT. Lastly, this chapter presented a detailed description of the GL-FRT scale and the psychometrics work previously completed with such. The remainder of this dissertation is organized as follows: Chapter Three describes methodology selected to complete this project. In this case, it is the Partial Credit Model, an extension of the basic Rasch Measurement model. Chapter Four presents the results from the Rasch Measurement analysis. And finally, the

[21] Both PCA and exploratory factor analysis (EFA) are used to simplify the structure of a set of variable; these are data reduction techniques. However, the ultimate purpose of these substantially differs. The objective of PCA is to reduce the number of observed variables to a smaller number of components. Conversely, the purpose of FA is to identify the number of underlying constructs or factors. For the construct validity evidence studies, typically, EFA is preferred over PCA (Fabrigar, Wegener, MacCallum, & Strahan, 1999).

last chapter of this dissertation presents the implications, limitations, and future research recommendations derived from this project.

CHAPTER 3

METHODOLOGY

Chapter Three describes the following aspects: (a) the instrument, dataset, and sample utilized within this dissertation; (b) Rasch Measurement model applied for the data analysis; and (c) the analysis plan providing a step-by-step report of the procedures undertaken to complete this project.

Instrument, Dataset, and Sample

Instrument

This project used the Grable and Lytton (1999) financial risk tolerance (GL-FRT) scale, an extensively used financial risk tolerance (FRT) measure in the field of financial planning and consumer economics (Kuzniak et al., 2015). The GL-FRT scale consists of 13 multiple-choice items (or questions) that assess FRT especially focusing on the following three main domains: investment risk, risk comfort and experience, and speculative risk (Grable & Lytton, 1999). The scale uses questions (some in the format of hypothetical scenarios such as questions 11 and 12) on general risk choice, guaranteed versus probable gambles, and choice between losses and gains.

The response format for these questions consists of a set of response choices. Each choice is meant to represent a particular area along the FRT continuum. In fact, the response choices are presented in ascending or descending order along a particular continuum designed to reflect a pre-determined rank. The scoring system established by Grable and Lytton (1999) is available in Appendix A of this dissertation. For example,

question 1 asks the following: "In general, how would your best friend describe you as a risk taker?" The response choices for this particular question are presented as follows: (a) real gambler, (b) willing to take risks after completing adequate research, (c) cautious, (d) real risk avoider. The scoring system for response choice for this particular question is 4 points if (a) is selected, 3 if (b), 2 if (c), and 1 if (d) is chosen. Thus, the order of the response choices for these questions follows the decreasing rank and scoring system.[22]

In addition, while the majority of questions included in the scale are presented with four response choices (i.e., questions 1, 2, 3, 6 7, 8, 11, and 13), there are some questions that, instead, are offered with only three (i.e., questions 4, 5, and 12) or two response choices (i.e., questions 9 and 10) (See Appendix A for complete list of questions and response choices). Take for example, Question 4, which asks the following: "In terms of experience, how comfortable are you investing in stocks or stock mutual funds?" For this question three response choices are available: (a) not at all comfortable, (b) somewhat comfortable, (c) very comfortable, instead of four response choices as in the above example in Question 1.

It is important to note that this particular FRT instrument was selected for this project for three reasons. The first is that this scale was developed using a psychometric theory, namely, Classical Test Theory (CTT), unlike some other existing instruments, such as the Survey of Consumer Finances FRT item or the Barsky et al. (1997) FRT measure. This contrast allows for comparison and illustration of one theory versus another, each founded on a different theoretical perspective. The second reason for

[22] Note that for analysis purposes, all 13 items were coded congruently to increase (or decrease) in one direction only. In particular, question 1 was reverse coded to match other items coding direction.

selecting this scale is that the GL-FRT measure is a publicly available instrument with readily available scoring guidelines for users, whereas, other instruments created and refined with the use of test theory (i.e., CTT), such as the *Finametrica* risk tolerance instrument, are not freely offered to the public. The third reason was access to a large and robust dataset with responses from a demographically rich sample.

Dataset and Sample

The data for this research were obtained from a repeated cross-sectional data collection project hosted by Rutgers New Jersey Agricultural Experiment Station. From year 2007 to 2014, Rutgers collected responses from over 2000,000 individuals to the GL-FRT scale via an open-access site (http://njaes.rutgers.edu:8080/money/riskquiz/). The sample frame used for this project came from responses collected from January 2013 through December 2013. Year 2013 was selected, as it was the most recent period with submitted responses for the entire year.[23] Including only valid and completed responses, the sample frame for that particular period was comprised of 25,079 respondents.

A further step was taken in delimiting the sample to subjects who were of age 25 or older. The justification for this action was grounded on the fact that many of the questions included in the GL-FRT scale are content specific for the choice of making formal investments via financial assets (stock, bonds, money market) in organized (exchange) markets. In this regard, the finance and economics literature has documented that, in general, young adults would likely not yet have sufficient wealth to actively invest in these markets (Constantinides et al., 2002; Campbell, 2006; Guiso et al., 2002; Van Rooij et al., 2011). Hence, the exclusion of this cohort seemed prudent for

[23] For example, responses for year 2014 were available only until the month of June.

generalizability purposes. The final sample, then, was comprised of responses of 11,905 respondents.[24]

Description of Sample

As shown in Table 2, the sample was over-represented by male respondents (61.6%); however, this is not surprising given the general tendency of men to exhibit investing behavior (Grable, 2000). In terms of age profile, more than 50% of the respondents in the sample were between age 25 and 44, while about 6% of the sample were past retirement age. More than half of the sample indicated being married (57.6%). Regarding educational attainment, the sample was over-represented by subjects with post-secondary degrees. Approximately 36% of respondents selected "bachelor degree" as the highest degree of education attained; 35% reported having completed a "graduate or professional degree." Household income patterns showed that an over representation of wealthy household. About 37% of those in the sample reported household income of $100,000 or more. Table 2 also provides data related to financial decision-making as reported by respondents. The majority of respondents (84.2%) indicated making their own investment decisions. The remainder (17.6%) reported that they rely on the advice of another person, such as a stockbroker, or financial planner, when making investment decisions. Mean and standard deviation risk scores for the GL-FRT scale can be found in columns 3 and 4 of Table 2, respectively. The mean scale score, across the sample, was 28.39 (SD = 5.20). Specific mean scores for each category of respondent characteristics are also shown. Interestingly, those with some high school education or less had a higher

[24] For analysis purposes, it is important to note that a total of 154,765 responses were used (i.e., number of subjects [11,905] times the number of items included in the GL-FRT scale [13 items]).

average FRT score (mean = 30.17; *SD* = 8.61) than the other educational attainment categories. Further, consistent with findings with in the FRT literature, males had a higher mean FRT score (mean = 29.39; *SD* = 5.13) compared to their counterpart (mean = 26.80; *SD* = 4.83).

Table 2.

Descriptive Statistics for Respondents and the Scale (N=11,905 respondents)

Variable	Scale Data		
	Percent	Mean	SD
Risk Score		28.39	5.20
13	0.2%		
14	0.1%		
15	0.2%		
16	0.6%		
17	0.7%		
18	1.1%		
19	1.5%		
20	2.0%		
21	2.6%		
22	3.5%		
23	4.4%		
24	5.1%		
25	6.1%		
26	6.7%		
27	7.7%		
28	8.0%		
29	7.7%		
30	7.3%		
31	7.2%		
32	6.6%		
33	5.5%		
34	4.1%		
35	3.2%		
36	2.3%		
37	1.7%		
38	1.0%		
39	0.9%		
40	0.5%		
41	0.3%		
42	0.2%		
43	0.1%		
44	0.2%		

45	0.1%		
46	0.1%		
47	0.2%		
Gender			
Female	38.4%	26.80	4.83
Male	61.6%	29.38	5.13
Age			
25 to 34	41.7%	28.67	5.06
35 to 44	20.3%	28.61	5.20
45 to 54	17.3%	28.47	5.21
55 to 64	14.5%	27.60	4.86
65 to 74	4.6%	27.14	4.53
75 and Older	1.6%	28.01	7.64
Marital Status			
Never Married	23.6%	28.83	5.37
Living with Significant Other	7.9%	28.69	4.93
Married	57.6%	28.27	5.06
Separated or Divorced	7.9%	27.95	5.27
Widowed	1.8%	26.55	6.33
Shared Living Arrangement	1.2%	29.25	5.69
Education			
Some High School or Less	1.1%	30.17	8.61
High School Diploma	4.2%	27.43	5.94
Some College	14.5%	27.58	5.21
Associate's Degree	9.1%	27.45	5.37
Bachelor's Degree	36.1%	28.58	4.93
Graduate or Professional Degree	35.0%	28.83	4.98
Household Income			
Less than $25,000	7.2%	28.45	5.70
$25,000 to $49,999	17.1%	27.59	5.30
$50,000 to $74,999	21.6%	27.80	5.31
$75,000 to $99,999	17.4%	28.27	4.88
$100,000 or more	36.7%	29.15	5.00
Decision Making			
Make Own Investment Decisions	82.4%	27.87	5.53
Rely on the Advice of Professional	17.6%	27.77	5.12

Rasch Measurement Model: Partial Credit Model

Since the inception of the initial, basic Rasch Model, additional extensions have been developed (e.g., Rasch rating scale model, partial credit model (PCM), and many-facet model). For the purpose of this project and its analysis, Wright and Masters' (1982)

PCM was applied to the measurement data obtained from the administration of the GL-FRT instrument. PCM allows each item or question to have its own response category structure (e.g., two items with different response formats such as a Likert-type question and a open-ended question or two items with an unequal number of response categories). This is in contrast with the other extensions of Rasch, such as Rating scale or the basic dichotomous model, that assume the same structure for all items or questions in the instrument (e.g., multiple choice questions with the same number of response categories or all questions have an open-ended response format).

As mentioned previously, the response categories across all multiple-choice questions in the GL-FRT were not consistent. As noted earlier, some questions have four response categories options (i.e., questions 1, 2, 3, 6 7, 8, 11, and 13), while others have three (i.e., questions 4, 5, and 12), or two response category choices (i.e., questions 9 and 10). Due to differences in response category structure, PMC was deemed suitable for the scale analyzed in this research.

Equation 7 shows the mathematical representation of the Partial Credit Model:

$$L_n = \ln \frac{P_{ni1}}{P_{ni0}} = \theta_n - \delta_i - \tau_{ix} \qquad (7)$$

Where,

P_{nix} is the probability that *n*, on item *i,* would be observed (or would respond) in category *x,*

θ_n indicates the ability of person *n*,

δ_i denotes the *endorsability* of item *i,*

τ_{kx}, is a threshold to being observed in, or *endorsability* of, category *x* relative to category *x-1* for item *i*.

In the context of FRT, P_{nix} is the probability of person *n* selecting category *x* for question *i* of the GL-FRT scale. The ability parameter, θ_n indicates the ability of person *n* from the delimited sample. In other words, it indicates the FRT level exhibited (in logits) by person *n*. The item parameter, δ_i, refers the *endorsability* of item *i.* This is simply the location of item *i*, on the continuum (in logits). The tau, τ_{ix} represents the category coefficient location, which is interpreted as the difference in endorsability (or location on the FRT continuum) between adjacent categories for a particular item *i*. The purpose of τ_{ix}, then, is to allow for category coefficients to vary across items, unlike in other models (e.g., dichotomous Rasch model or rating scale model) were the category coefficient locations are fixed (e.g., Engelhard, 2013; Engelhard & Wind; Liu, 2010; Wright & Masters, 1992).

Analysis Plan

In this project, Rasch Measurement analysis was conducted in order to evaluate the psychometric properties of the GL-FRT scale using the Rasch Measurement Theory. Specifically, Winsteps Version 3.91.1 (http://www.winsteps.com/winsteps.htm) was used to conduct the Rasch modeling. A series of steps were taken to assess the quality of the GL-FRT scale; specifically, the following quantitative criteria were examined: (a) the *unidimensionality* of the scale, (b) monotonicity of response categories and category usage, (c) item fit, (d) category coefficient order, and (e) person reliability. The decision to evaluate these five elements was based on several recommendations and guidelines provided by Linacre (1999; 2000) and Engelhard (2002; 2013). In addition to the quantitative criteria listed above, the following graphical displays were carefully examined as part of the analysis procedures: (a) Wright Variable Map; (b) category

probability functions; (c) test information function, item information, and category information curves. Table 3 presents a summary of the guidelines for the sequence of steps taken in the analysis of this project. Each of these aspects is described below.

Table 3.

Guidelines for Examining the Psychometric Quality of the GL-FRT scale

Quantitative Evaluation Criteria	Graphical Evaluation Criteria
a) *Unidimensionality* b) Monotonicity of response categories and category usage c) Item fit c) Category coefficient order d) Person reliability	a) Wright variable map b) Category probability functions c) Test information, item and category information

Note. These guidelines were adapted from Engelhard (2002; 2013) and Linacre (1999; 2002).

Quantitative evaluation aspects

a) *Unidimensionality* of the scale: A key assumption of Rasch Measurement Theory is that a measurement scale describes, one, and only latent trait. In terms of the GL-FRT, it was important to determine whether the current version of the scale was truly a measure of one trait, that of FRT. The Rasch residuals of the measure were evaluated to examine the dimensionality of the GL-FRT.

b) Monotonicity of response categories and usage category: A monotonic progression of scale categories is expected in a Rasch analysis. Average person ability estimates in each category are often used as evidence of monotonicity. This evaluation process was followed for the assessment of the GL-FRT.

Another aspect evaluated was the frequency and distribution of observed category usage. Categories with fewer than 10 observations represent challenges for

interpretation of scales (Linacre, 2002). Thus, it was necessary to evaluate the need to collapse or eliminate these categories. Furthermore, for categories with zero observations, steps were taken to determine whether or not this was an artifact of structural or incidental zeros.

c) Item fit: Rasch analysis provides fit statistics for items that allow comparison of randomness observed in the data with randomness expected by the model. Different fit statics are available when conducting a Rasch analysis. Two commonly used fit indices are mean square residuals (MNSQ) and standardized square residuals (ZSRD). True to their name, the former are squared residuals; the latter are normalized *z*-scores of the residuals. Furthermore, MNSQ and ZSRD are presented as INFIT (weighted means) and/or outfit (simple arithmetic means) statistics. For this project, outfit MNSQ were used when evaluating the fitness of the items to be included in the GL-FRT. Outfit MSNQ were especially useful because of their sensitivity to outliers.

d) Category coefficient order: The thresholds between responses for a particular item should monotonically increase in value from low to high (Andrich, 1978a; 1978b). In this study, if there was not a linear progression between the particular thresholds for an item, then the category order functioning of the item was deemed inadequate (Linacre, 2002; Engelhard, 2002).

e) Person reliability: Similar to CTT, in Rasch analysis an analogous reliability coefficient was examined; this was the person reliability coefficient. Unlike Cronbach's alpha that is widely used in CTT, the person reliability coefficient in Rasch analysis does not include extreme or perfect scores in the computation.

Graphical evaluation aspects

a) Wright Variable Map: A unique feature of a Rasch analysis is the Wright variable map—a graphical representation of items and persons on the continuum. The variable map is a helpful visual aid for observation of the distribution of items along the continuum. A good measure should be able to target the intended population by matching the *endorsability* of items with the ability of a person.

b) Category probability functions: These are a visual representation of the probabilities relationship between *endorsability* for each of the categories for an item and the ability location of a subject on the latent construct. The expectation was that each category for an item functions in such a way that increasing subject location on the latent variable was associated with an increasing probability of a category being endorsed.

c) Test information function, item information, and category information curves: The total information function curve can be interpreted as a reliability indicator of the instrument. In other words, it indicates the regions along the continuum where the test measures with more precision. The inverse of the test information function is the standard error curve. Item information and category information represent the same idea of measurement precision, but in terms of item and category, respectively.

Finally, the evaluation of each of these aspects served as a quality control check to identify both the strengths of the scale and any areas needing improvement. The ultimate goal of this analysis was to create a stable measure with a set of Rasch calibrated items such that later, this measure could be used with persons as subjects. Note that the

particular emphasis of the analysis in this project was on items, rather than persons. Although the specific aspects evaluated in this analysis can be a general representation of criteria for scale evaluation using Rasch Measurement Theory, the theory is not limited to these criteria exclusively. Several others aspects (e.g., item differential functioning and person fit) of the theory's framework allows for a finer granularity in the evaluation of a scale of interest for the purpose of establishing evidence of invariant measurement.

Person's Ability Estimation and Person's Fit Statistic Tool

The last part of the methodology deals with the creation and development of a tool to estimate a person's ability (or in other words, a person's FRT, measured in logits), and to diagnose a person's misfit for the GL-FRT scale using the set of Rasch calibrated items. Windows Excel Version 14.6 was utilized to generate a template of the modified GL-FRT tool for use by end-users, consumers, financial planners, policy makers, or researchers whereby the scores from their answers would be automatically scaled using Rasch measurement theory.[25] Specifically, this Excel template allows for the estimation of a FRT measure (in logits) by entering the raw score (or answers to each question in the scale). Using the scoring system and cut-off values pre-determined by Grable and Lytton (1999),[26] the automatic tool produces the FRT logit score. It also generates the categorical level of FRT associated with that score (i.e., low risk tolerance, below-average risk tolerance, average/moderate risk tolerance, above-average, and high risk tolerance). This feature will be of great use for interpretability purposes. In addition to the

[25] Specifically, the tool utilized Joint Maximum Likelihood estimation (with 9 iterations) to estimate person's ability, standard errors, and the fit statistics.

[26] The cut-off values proposed by Grable and Lytton (1999) for the categorization of FRT using their scale is as follows (in raw scores): 18 or below, low FRT (i.e., conservative investor); 19-22, below-average FRT; 23-28, average/moderate FRT; 29-32, above-average FRT; and 33 and above, high FRT (i.e., aggressive investor).

FRT measure estimation, the tool will automatically compute a standard error and fit statistic for each person (i.e., raw scores representing the answers to the question in GL-FRT) such that unreasonable or inconsistent response patterns would be reported. In other words, when use of the fit statistic reports that a person's ability is a misfit, then, one possible interpretation is that the scale respondent is inconsistently answering questions in the domain of FRT. For example, the user may have selected response options indicative of high-risk FRT for some questions, and selected response categories indicative of low FRT for other questions. Remember, because Rasch assumes *unidimensionality* in terms of domains, it is expected that a probabilistic degree of consistency occurs across items. For interpretation purposes, a misfit should be thought as a warning that indicates, further assessment on the person's FRT is required.

Summary of Chapter Three

This chapter has reviewed the research methodology selected for the completion of this project. Specifically, a description of the instrument used (i.e., GL-FRT), dataset and sample, and the particular Rasch model employed in the analysis (i.e., Partial Credit Model) has been presented. Additionally, the plan of analysis detailing the aspects to evaluate psychometric quality of the GL-FRT scale has been delineated. The remainder of this project is structured as follows: Chapter Four reports the findings from the Rasch Measurement analysis. And, Chapter five presents the implications, limitations, and future research recommendations relevant to this dissertation.

CHAPTER 4

RESULTS

The results from the two-stage analysis are presented in this chapter. The initial stage involved the Rasch analysis of the original version of the GL-FRT scale (i.e., the analysis of all thirteen questions). After a careful evaluation of preliminary results, two particular items (questions 9 and 10) were identified and diagnosed as poorly fitting items for the tested model (i.e., a partial credit model [PCM] with all thirteen questions). Then, the second stage analysis examined the revised version of the GL-FRT in which the aforementioned items were removed from the instrument. For organizational purposes, this chapter is divided into three principal sections. The first section presents the results of a PCM analysis for the original 13-item scale. The second section reports the results of the application of PCM to a revised version of the scale. Finally, in the last section, the resulting parameters from the Rasch calibration of section two (i.e., items parameters) were used to develop a tool that generates individuals' abilities estimation (or in other words, the FRT exhibited by individuals), individualized standard errors, and fit statistics. An example with two pre-selected respondents from the dataset is used to illustrate the usage of this tool.

Model I: Partial Credit Model for Original 13-item Version of the GL-FRT

The initial analysis stage in this project was to evaluate responses to the original version of the GL-FRT scale. All thirteen items were analyzed using the PCM. As mentioned in the previous chapters, a set of quantitative elements were examined to

determine the psychometric quality of the scale; these were the *unidimensionality* of the scale, monotonicity of response categories and category usage, item fit, category coefficient order, and person reliability (Linacre, 1997; 1999; 2000). Additionally, following the recommendation of Engelhard (2002; 2013) visual displays were also reviewed as part of the analysis procedure. Specifically, the Wright Variable Map was inspected.

Quantitative Aspects for Model I

Unidimensionality. A fundamental assumption in the Rasch Model is the notion that the scale is *unidimensional.* In order words, there is only one latent variable that increases (or decreases) in a linear fashion that is being measured. Different criteria to assess *unidimensionality* are available (Embretson & Reise, 2000; Engelhard, 2013; Linacre; 1998; 2006; Reckase, 1979; Smith, 2002). Reckase recommended that the Rasch model should explain a minimum of 20% variance in order for a measure to be considered *unidimensional*. Linacre suggested more conservative heuristics, and stated that a ≥40% of the variance explained by the Rasch measure is indicative of a strong measurement dimension while ≥30% is considered a moderate measurement dimension. In addition, Embretson and Reise provided guidelines for *unidimensionality* evaluation from the principal component analysis of residuals. A minimum of 3:1 ratio resulting from the variance explained by a Rasch measure against the variance of the first principal component of residuals (largest secondary dimension) is recommended. Also, it has also been suggested that the first component of residuals (i.e., the variance explained by the first contrast of the residuals or second dimension) not be greater than 15%.

For the purposes of this project, *unidimensionality* was assessed by examining the variance explained by the Rasch measure, and by evaluating the existence of other potential substantial dimensions through principal component analysis of the residuals. For the measure to be considered *unidimensional,* three main components were requisite: (a) the variance explained by the Rasch measure (i.e., primary measurement dimension) was of at least 20%; (b) the variance explained by the first component of the residuals (second dimension) was less than 15%; (c) a minimum ratio of 3:1 for the variance in the measurement dimension compared to the variance of the first principal component of residuals.

For this model, the variance explained by the measure was 35.6%, which provided evidence of *unidimensionality* (i.e., FRT as the main and primary substantial dimension being measured) based on the aforementioned benchmarks; specifically, that there was evidence of a moderate measurement dimension of FRT. The largest secondary dimension (i.e., the first contrast in the residuals) explained 8.4% of the variance, which then met the criteria for *unidimensionality*. Finally, the ratio for the variance explained by measurement dimension compared to variance of the secondary largest dimension was 4:1, meeting the third requirement for *unidimensionality.*

Monotonicity of response categories, its usage, and coefficient order. The proper functioning of the scale was assessed by inspecting: (a) whether or not the average measures advanced monotonically (Linacre, 1999; 2002); (b) whether or not categories for the different items were actually used and whether they exhibited a regular distribution (e.g., uniform, normal, unimodal, bimodal [Engelhard & Wind, 2013]); and (c) whether or not the response category thresholds (coefficients) monotonically

progressed from low to high (Bond & Fox, 2007; Engelhard, 2013; Engelhard & Wind, 2013; Tennant & Conaghan, 2007). Table 4 provides the rating scale structure for all questions analyzed in model I. The average ability that each of the categories contained in the items, the category usage (presented in frequency and percentage), and the category coefficient location are provided.

As seen in Table 4, the average ability for the categories in all thirteen questions (third column) did exhibit the desired monotonic program from one category to the adjacent one. Thus, this was indicative that the directionality of the rating categories for the items in the scale appeared to be aligned with the latent variable. In other words, a respondent's endorsement or selection of high rating scale categories for items reflects high locations on the latent variable for the respondent.

In terms of category usage, the frequencies of category usage for each question is presented in Table 4, column 4. It was observed that there were more than 10 observations for each category for all questions. Linacre (2002) explained that categories with less than 10 observations limit the precision and stability of the model's estimates. Nonetheless, given the large sample utilized in this project, this method might not be the most optimal for the evaluation of category usage. Hence, the percentage and distribution of category usage was then examined. Column 4 of Table 4 contains the percentage of category usage for each question. Most of the items with three or more categories had a good spread that conformed to regular distributions. Nonetheless, there were some interesting observations. For example, several categories, such as category 4 for item 3, category 4 for item 6, category 4 for item 7 were lightly used (3% or less). Such categories might need revisions to ensure a more continuous advance of the latent

construct across the categories for such particular items. Additionally, column 6 of Table 4 provides the outfit MSNQ for the rating categories of each item. It is expected that the outfit MSNQ approached 1. Overall, most of the categories for each of the items were within the expected range of 0.8 and 1.2. However, note that the outfit statistics for the categories of question 3 (categories 1 and 3), 11 (category 4), 9 (both categories), and in particular 10 (both categories) were farther away from 1. Such response categories might require additional revision and functioning assessment.

Finally, as part of the category functioning assessment, the category coefficients or thresholds were evaluated. The category coefficient order should match the intended order of categories in terms of the linear progression along the latent construct (Andrich, 1978a; 1978b; 1988). If the thresholds do not progress in a linear manner, as expected, then problematic categories could be collapsed in order to improve category functioning (Tennant & Conaghan, 2007). As seen in Table 4, column 7, all category coefficients for each of the questions monotonically increased from one threshold to the next adjacent one. Thus, no category collapsing was suggested for any of the items.

Table 4.

Rating Scale Structure for Model I (11,905 respondents; 13 items)

Question	Response Category	Average Ability	S.E. Mean	Category Usage: Frequency (%)	Outfit MSNQ	Category Coefficient Location
1	1	-1.62	0.07	433 (4)	1.10	
	2	-0.74	0.01	3911 (33)	0.80	-2.76
	3	0.02	0.01	6925 (58)	0.80	-0.42
	4	0.83	0.07	636 (5)	1.10	3.18
2	1	-1.01	0.02	2541(21)	1.00	
	2	-0.3	0.01	5897 (50)	0.90	-1.57

	3	0.24	0.01	2174 (18)	0.80	0.77
	4	0.7	0.03	1293 (11)	1.20	0.80
3	1	-0.68	0.01	3970 (33)	1.30	
	2	-0.31	0.01	3488 (29)	1.10	-1.18
	3	0.09	0.01	4055 (34)	1.30	-0.97
	4	1.25	0.09	392 (3)	1.10	2.15
4	1	-1.13	0.02	2095 (18)	1.00	
	2	-0.53	0.01	3421 (29)	0.80	-0.21
	3	0.2	0.01	6389 (54)	0.90	0.21
5	1	-0.96	0.02	2717 (23)	1.00	
	2	-0.27	0.01	5656 (47)	0.90	-0.88
	3	0.35	0.02	3532 (30)	1.00	0.88
6	1	-1.08	0.03	1309 (11)	1.00	
	2	-0.46	0.01	6971 (59)	0.80	-2.71
	3	0.37	0.01	3240 (27)	0.80	0.32
	4	1.3	0.09	385 (3)	1.10	2.38
7	1	-0.67	0.01	4627 (39)	1.00	
	2	-0.16	0.01	5533 (46)	1.00	-1.71
	3	0.42	0.20	1465 (12)	1.10	0.31
	4	1.78	0.11	280 (2)	1.10	1.40
8	1	-1.5	0.04	882 (7)	0.90	
	2	-0.66	0.01	3377 (28)	0.90	-1.81
	3	-0.5	0.01	6384 (54)	0.80	-0.50
	4	0.77	0.04	1262 (11)	1.00	2.31
9	1	-0.58	0.10	7187 (60)	1.50	
	3	0.27	0.10	4718 (40)	1.40	0.00
10	1	-0.78	0.02	3177 (27)	2.30	
	3	-0.05	0.01	8728 (73)	1.80	0.00
11	1	-1.29	0.02	1695 (14)	0.90	
	2	-0.33	0.01	5951 (50)	1.00	-1.93
	3	0.27	0.01	3130 (26)	0.90	0.51
	4	0.32	0.04	1129 (9)	1.60	1.42
12	1	-1.02	0.01	3402 (29)	0.80	

	2	-0.14	0.01	6314 (53)	0.60	-1.13
	4	0.65	0.02	2190 (18)	0.90	1.13
13	1	-0.78	0.01	3971 (33)	1.00	
	2	-0.21	0.01	5592 (47)	1.00	-1.65
	3	0.42	0.02	1920 (16)	0.90	0.27
	4	1.35	0.08	422 (4)	1.10	1.38

Items fit. After evaluating the dimensionality of the scale and the category functioning, the item fit statistics for all thirteen questions were examined. Fit statistics allow testing fundamental measurement assumptions, such as directionality and dimensionality (Bond & Fox, 2007; Wright & Stone, 1979). The main idea of identifying misfit items is to further examine the causes of such. For example, poor fit could stem from unclear and ambiguous question wording or from the notion that more than one construct is being measured (*unidimensionality*) (Bond & Fox, 2007; Conrad, Iris, Ridings, Langley, & Anetzberger, 2010; Engelhard, 2013). The Rasch model presents MNSQ (i.e., the squared residuals based on the difference between observed and expected data by the Rasch model) in the form of two indicators of misfit: infit and outfit. The latter is particularly useful in identifying highly unexpected outliers responses (Engelhard, 2013; Engelhard & Wind, 2013; Tennant & Conaghan, 2007). Thus, for the purpose of this project, the outfit MNSQ was used and calculated using simple arithmetic means, to assess items fit[27].

Indices of model-data fit permit the comparison of the randomness observed in the data against the expected randomness by the model (Engelhard & Wind, 2013). When the analyzed data matches the Rasch model, a relatively uniform level of randomness should be seen. In fact, a good fit to the Rasch model is when the value of randomness

[27] Appendix D provides the formulas to calculate outfit MNSQ (Engelhard, 2013).

(outfit MSNQ, in this project) is around 1 (Bond & Fox, 2007). A low outfit statistic value (<1.00) suggests overfit to the Rasch model. In other words, the data is more predictable than the model would expect (Wright & Linacre, 1994). Nonetheless, those with seemingly excessive predictable values do not contradict what is known; yet, do not offer extra or unique information from that which is known. To an extent, overtfitting values, and in particular in the context of items, are inefficient. These items indicate redundancy, albeit harmless. Conversely, a high outfit statistic value (>1) indicates underfit to the Rasch model. This references that the data are less predictable than the model would expect (Wright & Linacre, 1994). As noted earlier, the difference between mean-square value and the expected value by the Rasch measure indicates randomness. Thus, a mean-square value of 1.30 shows that there is 30% more noise in the data than modeled (Wright & Linacre, 1994).

In terms of heuristics, a good model fit should have items with MNSQs ranging from 0.75 to 1.3 (Engelhard, 2013; Liu, 2010; Wilson, 2005). Wright and Linacre (1994) provided a guideline on reasonable item MNSQ ranges based on the type of test (e.g., multiple choice questions, rating scale, judged or rated based tests, etc.). Based on their guideline, the most conservative range for MNSQ value was used in this project; that is $0.8 \leq \text{MNSQ} \leq 1.2$.

Table 5 presents the calibration of the thirteen items in model I. Column 2 and 3 show the item parameters resulting from the calibration and their associated standard error, respectively. Column 4 provides the outfit MSNQ for the items (arranged from high MNSQ values to low MNSQ values), and column 5 presents the point-measure correlation (PTMEA), which indicates how each particular item contributes to the item

difficulty. As seen, items 10, 9, and 3 have MSNQ greater values than 1.2. Based on the adopted MNSQ range for this project, these items do not fit the Rasch model as their fit statistics are beyond the acceptable range. PTMEA shows that these items contribute the least to the measure.

Table 5.

Calibration of items in Model I (11,905 respondents; 13 items)

Question	Measure	SE	Outfit MSNQ	PTMEA
10	-0.98	0.01	2.20	0.34
9	0.07	0.01	1.45	0.44
3	0.78	0.01	1.23	0.42
11	-0.01	0.01	1.09	0.49
7	1.14	0.01	1.05	0.48
2	0.14	0.01	1.01	0.54
13	0.85	0.01	0.97	0.53
5	-0.42	0.01	0.96	0.49
1	-0.48	0.02	0.92	0.52
8	-0.47	0.01	0.91	0.56
6	0.35	0.02	0.90	0.55
4	-1.00	0.01	0.88	0.54
12	0.03	0.02	0.81	0.59
Mean	0.00	0.01	1.10	
S.D.	0.64	0.00	0.35	

Note. Items have been arranged in descending order based on outfit MSNQ.

Person reliability. The reliability of separation statistic for person is comparable to Cronbach's coefficient alpha reported in CTT (Engelhard, 2013; Engelhard & Wind, 2013). As such, values are interpreted in the same manner. It is important to note that unlike Cronbach's alpha, the reliability of separation in a Rasch analysis excludes extreme scores in its computation. Thus, perfect scores are not taken into account for such calculation. In the social sciences, generally, reliability coefficients within the range

of 0.70-0.79 are deemed acceptable; within 0.80-0.89, good; and above 0.90 are considered excellent (Cortina, 1993; DeVellis, 2012; George & Mallery, 2003; Kline, 2000). Consistent with the above heuristics, in the field of financial planning and consumer economics, acceptable reliability coefficients range from 0.70 to 0.79 (Saad, Carter, Rothenberg, & Israelson, 1999)

The results from model I showed the reliability of separation statistic for this model was 0.73 (N=11,849 respondents). The Cronbach's alpha "test" reliability was 0.75 (N=11,905 respondents). Thus, both statistics were deemed adequate following the acceptability guidelines used in the field.

Visual Aspects for Model I

Wright Variable Map. Figure 3 shows a graphical display of the spread of respondents' measures (i.e., financial risk tolerance exhibited by respondents), items measures (i.e., *endorsability* or location of items along the continuum of FRT), both mapped on a Rasch ruler (i.e., the dashed line on the map) that represent the latent construct of FRT. It is important to note that all measures are on the same interval logit scale. The first column *measure* (from left to right) is logits shown in descending order from top to bottom. Logits are the units of measurement—the common unit of measurement for all elements displayed on the map. Logits can be thought of as the feet, yards, or centimeter of the ruler. The second column, *persons,* shows the location of each of 11,905 respondents used in the analysis. The location of each respondent is based on the ability measures (or in other words, the FRT measures exhibited). Respondents with higher risk tolerance appear at the top of the column, and respondents with lower risk tolerance appear near the bottom. The *pound* or *number sign* "#," represents 88

respondents, while each *dot* "." can represent 1 to 87 respondents. The risk tolerance measures (ability measures) in this sample ranged from -5.44 to 5.66, measured in logits (M = -0.24, SD = 0.96, N = 11,905). The next column is the *map*, which shows a dashed line representing the ruler, or in other words, the continuum of FRT. Item *endorsability* or difficulty calibrations (in logit units) are shown in the next column. As seen, the location of each of the thirteen items has been mapped based on its respective *endorsability* measure. The item measures (*endorsabilities*) ranged from -1.00 (question 4) to 1.14 (question 7) (M = 0.00, SD = 0.64, N = 13). More hard to endorse (or less frequently endorsed) items are located near the top of the column, and easier to endorse (or more frequently endorsed) items are located closer to the bottom. Finally, note the letters "M," "S," and "T" are shown along the ruler. These represent the mean, standard deviation, and two standard deviations respectively. Thus, to the left side of the ruler (dashed line), the mean (-0.24), standard deviation to both directions of the continuum (-0.24 ± 0.96), and the two standard deviations (-0.24 ± 2*0.96) for the person measures are marked. Similarly, for items the mean "M" (0.00), the standard deviation "S" (-0.00 ± 0.64), and the two standard deviations marked on both directions of the continuum (-0.00 ± 2*0.64) are provided. Also, note that in order to map both facets or parameters (i.e., persons and items), one of them has to be anchored at 0.00 logits. In this case, items have been anchored at 0.00 logit (hence, the reason the mean item measure is equal to 0.00), and the other facet (person measures) float along accordingly.

From the variable map for Model I, several observations can be made. For persons, it is apparent that in this sample there were several outliers on the extreme ends of the continuums. Overall, the spread was pretty normally distributed. From the items

measures, it can be observed that there is a good spread of items across the section of-1.00 to 1.14. However, there were no questions in this scale that targeted respondents with FRT measures passed one standard deviation on the lower end of the continuum. On the other hand, there were questions that targeted respondents with FRT measures of up two standard deviations on the upper area of the construct (i.e., high FRT). There were gaps along the continuum that ideally should have questions to better target and measure respondents with FRT measures in these sections of the latent construct. Furthermore, extreme respondents (that is persons with extremely low FRT and persons with very extremely high FRT) were not targeted as respondents with more average FRT levels. Finally, the Wright map is a good tool to observe the item hierarchy (based on item *endosabilities*), which provides an indication of construct validity (Smith, 2002). Since this was the first time Rasch Measurement analysis has been applied to the GL-FRT scale, no priori hypothesis about the hierarchy of items was formulated. In fact, this was a good opportunity to examine the spread of items across the continuum. Overall, the scale measures targeted respondents with average abilities. This begs the validity study question of "are these items targeting the intended audience?" This idea warrants further research for the future. Finally, the category probability function, item and category information curves, and test information function are available in Appendix B of this dissertation.

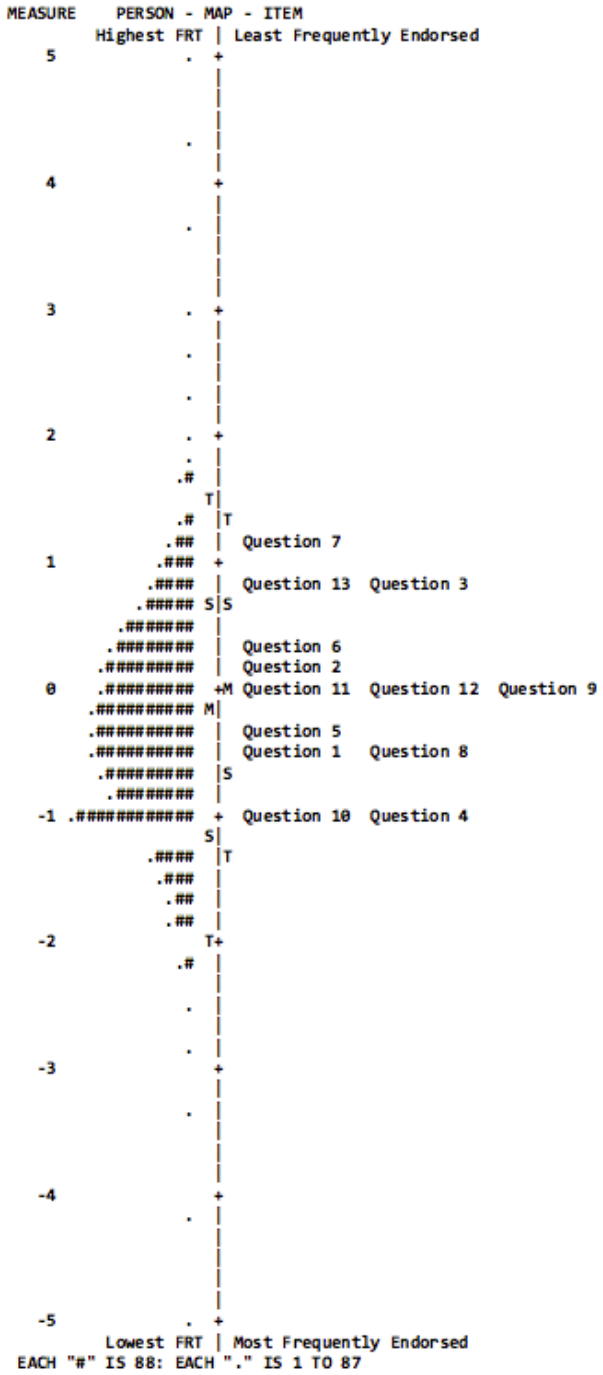


Figure 3. Wright map for the GL-FRT scale (Model I): Items' location (*endorsabilities*) and respondents' ability are both depicted on the continuum (in logit units).

Model II: Partial Credit Model for the Revised Version of the GL-FRT

After careful evaluation of the results, particular items from Model I, (questions 3, 9 and 10), were removed due to poor fit. Subsequently, an additional PCM was used to conduct a Rasch analysis with the 10 remaining items (i.e., questions 1, 2, 4, 5, 6, 7, 8, 11, 12, and 13). A series of quantitative and visual aspects were examined to determine the psychometric quality of the reduced version of scale, as recommended by Rasch measurement evaluation guidelines (Engelhard, 2002; 2013; Engelhard & Wind, 2013; Linacre, 1999; 2000). Specifically, the *unidimensionality* of the new version, monotonicity of response categories and category usage, item fit, category coefficient order and their differences, and person reliability were inspected. Furthermore, the Wright variable map, conditional probabilities curves, item and category information curves, and test information fiction were examined.

Quantitative Aspects for Model II

Unidimensionality. The assumption of *unidimensionality* was evaluated using a principal component analysis of residuals. Specifically, the variances explained by the primary measurement dimension and secondary dimension were examined. The same criteria in Model I were used to assess the *unidimensionality* of data in Model II. These were that (a) the variance explained by the primary measurement dimension was greater or equal to 20%; (b) the variance explained by the secondary measurement dimension was less than 15%; (c) and that a minimum ratio of 3:1 for the variance in the measurement dimension compared to the variance of the first principal component of residuals was observed.

For Model II, the variance explained by the Rasch measure was 42.4%, which based on Linacre (2006) provided strong evidence of *unidimensionality.* The second criterion in assessing *unidimensionality* was the size of the variance explained by the largest secondary dimension, where a desirable size is >15%. For this model, the first contrast in the residuals explained 9.4% of the variance. Thus, the second requirement for data *unidimensionality* was met. The ratio for the variance explained the Rasch measure compared to the variance explained by the secondary largest dimension was assessed, a ratio of 3:1 is the minimum accepted value for this criterion. For Model II, the ratio was 4:1, meeting then the third requirement for *unidimensionality*. In summary, examination of dimensionality revealed one clear dimension that made sense statistically and theoretically. A discussion on the theoretical aspect of this dimension is offered in the next chapter.

Monotonicity of response categories, usage and category coefficient order

The adequacy of category functioning for the revised version of the GL-FRT scale was evaluated by examining the advancement of average measures across the categories for items, the usage of each category (e.g., percentage of respondents who selected a category); and the advancement of category threshold for each items, as well as the difference between thresholds (Bond & Fox, 2007; Engelhard, 2013; Engelhard & Wind, 2013; Linacre, 1999; 2000; Tennant & Conaghan, 2007).

Table 6 provides the rating scale structure for all questions analyzed in Model II. The average ability the categories for each item, category usage (presented in frequency and percentage basis), the category coefficient location for each item (i.e., thresholds), and the difference between thresholds are presented.

As seen in Table 6, column 3, the average ability for the categories in all ten questions increased monotonically, as expected, from one category to the next one. Such monotonic advance in categories for items in the revised version of GL-FRT scale suggested that the directionality of the rating categories for the item was aligned with the theoretical directionality of the latent construct.

Regarding the category usage for each question, the percentage and distribution of category usage was assessed (column 5). As observed, all items, save item 7, had a good spread of category usage. The selection of category 4 for item 7 was seldom (2%). Such seldom-selected categories might need revisions to ensure a more continuous advancement of the latent construct across the categories for this particular item. Moreover, column 6 of Table 4 provides the outfit MSNQ for the rating categories of each item. Ideally, outfit MSNQ values for the categories should approach closer to 1. As seen, category 4 for item 2 (*outfit MSNQ* = 1.40) and category 4 for item 7 (*outfit MSNQ* = 1.30) were out of the expected range for acceptability. These misfitting response categories indicate that additional revision and functioning assessment might be needed.

Next, the category coefficients or thresholds for each item were examined. As seen in Table 6, column 7, all category coefficients for each question included in the revised version of the scale monotonically increased, as desired, from one threshold to the adjacent one. These results confirmed that the order of categories for each item matched the linear progression of the latent construct, in this case FRT (Andrich, 1978a; 1978b; 1988).

Finally, the differences between thresholds (i.e., $\tau_{ix} - \tau_{ix-1}$) were assessed. Linacre (1999; 2002) suggested a minimum of 1.40 logits difference between categories

should be observed in order to conclude that categories are distinctive. Table 6, column 8 presents the differences for the category thresholds for each item. For item 2, the difference between threshold 3 ($\tau_{2\,3}$) and threshold 2 ($\tau_{2\,2}$) was 0.21, which was below the minimum recommended. It is interesting to note that the upper-FRT category for this item was also deemed misfitting when evaluating the outfit value for this category. This violation to the expected minimum threshold difference can be interpreted as a category distinction problem. In other words, there might not be sufficient difference between category 3 and 4 for item 2, such that collapsing the categories should be considered.

Items fit. For this project, a conservative range for MSNQ value interpretation, suggested by Wright and Linacre (1994), was used: $0.8 \leq \text{MNSQ} \leq 1.2$. Table 7 presents the calibration of the 10 items (i.e., questions 1, 2, 4, 5, 6, 7, 8, 11, 12, and 13) retained in Model II. Column 2 and 3 show the item parameters resulting from the calibration and their associated standard error, respectively. Colum 4 provides the outfit MSNQ for the items (arranged from high MNSQ values to low MNSQ values). Column 5 presents the point-measure correlation (PTMEA). As observed, all 10 items had outfit MSNQ values within the acceptability range, which confirmed the well functioning of items with respect to the directionality of the measure, and strengthened the evidence for *unidimensionality* for the item data. Item 11 had the highest outfit MSNQ value of 1.18, which indicated an underfit when using the Rasch Model. In order words, for this item, there was 18% more randomness (or noise) in the data than was modeled. On the other hand, item 12 exhibited the lowest outfit MSNQ value (i.e., 0.82), which suggested an overtfit to the Rasch model. This indicated 18% deficiency in Rasch model-predicted randomness, and an implied 22% more ambiguity in the inferred measure than modeled

(i.e., 100 * [1-0.82] / 0.82). As Wright and Linacre (1994) stated, items with low MNSQ values might indicate a degree of redundancy in the responses, yet this is not harmful to the model or the scale. All items showed a positive and moderate PTMEA correlation, which indicated that the items contributed to the measure and to the item difficulty as intended (e.g., items were not improperly scored or there was not a substantial sub dimension being measured) (Cavanagh & Waugh, 2011; Liu, 2010). Finally, it is important to note that the item parameters that resulted from this Rasch calibration were the ones to be used for the estimation (ability, standard error and fit statistic) tool presented in the following section.

Person reliability. For this model, the reliability of separation statistic for this model was 0.73 (*N* = 11,837 respondents), and the alpha reliability was 0.77 (*N* = 11,905 respondents). Based on the acceptability guidelines for reliability in the field (DeVellis, 2012; Engelhard, 2013; George & Mallery, 2003; Kline, 2000; Kuzniak et al., 2015; Saad, Carter, Rothenberg, & Israelson, 1999), both reliability estimates were considered acceptable.

Table 6.

Rating Scale Structure for Model II (11,905 respondents; 10 items)

Question	Response Category	Average Ability	S.E. Mean	Category Usage: frequency (%)	Outfit MSNQ	Category Coeff. Location	Diff. ($\tau_{ix} -$ $_{ix-1}$)
1	1	-2.11	0.08	433 (4)	1.20		
	2	-0.97	0.01	3911 (33)	0.80	-3.10	
	3	0.09	0.01	6925 (58)	0.90	-0.39	2.71
	4	1.01	0.07	636 (5)	1.10	3.49	3.88
2	1	-1.28	0.02	2541 (21)	1.10		
	2	-0.37	0.01	5897 (50)	1.00	-1.81	
	3	0.34	0.02	2174 (18)	1.00	0.80	2.61
	4	0.95	0.04	1293 (11)	1.40	1.01	0.21

4	1	-1.51	0.02	2095 (18)	1.10		
	2	-0.7	0.01	3421 (29)	0.80	-0.35	
	3	0.33	0.20	6389 (54)	0.90	0.35	0.70
5	1	-1.31	0.02	2717 (23)	1.00		
	2	-0.33	0.01	5656 (47)	0.90	-1.02	
	3	0.56	0.02	3532 (30)	0.90	1.02	2.04
6	1	-1.38	0.04	1309 (11)	1.10		
	2	-0.57	0.01	6971 (59)	0.90	-3.00	
	3	0.54	0.01	3240 (27)	0.80	0.36	3.36
	4	1.48	0.09	385 (3)	1.20	2.64	2.28
7	1	-0.87	0.02	4627 (39)	1.10		
	2	-0.17	0.01	5533 (46)	1.20	-1.94	
	3	0.64	0.03	1465 (12)	1.20	0.35	2.29
	4	2.04	0.11	280 (2)	1.30	1.59	1.24
8	1	-1.92	0.05	882 (7)	1.10		
	2	-0.84	0.02	3377 (28)	0.90	-2.11	
	3	-0.02	0.01	6384 (54)	0.90	-0.48	1.63
	4	0.99	0.04	1262 (11)	1.00	2.59	3.07
11	1	-1.72	0.03	1695 (14)	0.90		
	2	-0.4	0.01	5951 (50)	1.00	-2.19	
	3	0.42	0.02	3130 (26)	1.00	0.54	2.73
	4	0.43	0.04	1129 (9)	1.90	1.65	1.11
12	1	-1.33	0.02	3402 (29)	0.80		
	2	-0.14	0.01	6314 (53)	0.70	-1.27	
	4	0.89	0.02	2190 (18)	0.90	1.27	2.54
13	1	-0.99	0.02	3971 (33)	1.10		
	2	-0.24	0.01	5592 (47)	1.20	-1.88	
	3	0.6	0.02	1920 (16)	0.90	0.31	2.19
	4	1.66	0.08	422 (4)	1.20	1.57	1.26

Table 7.

Calibration of items in Model II (11,905 respondents; 10 items).

Question	Measure	SE	Outfit MSNQ	PTMEA
11	-0.01	0.01	1.18	0.53
7	1.26	0.02	1.16	0.50
2	0.17	0.01	1.11	0.56
13	0.95	0.01	1.07	0.54
5	-0.57	0.01	0.99	0.59
1	-0.59	0.02	0.96	0.56
8	-0.47	0.02	0.97	0.56
6	0.36	0.02	0.96	0.57
4	-1.14	0.01	0.92	0.60
12	0.03	0.02	0.82	0.63
Mean	0.00	0.02	1.01	
S.D.	0.70	0.00	0.11	

Note. Items have been arranged in descending order based outfit MSNQ.
Note. The items measures presented in this table were item parameters used for the construction of the spreadsheet.

Visual Aspects for Model II

Wright Variable Map. The variable map for the revised version of the GL-FRT scale is presented in Figure 4. Similar observations to the one for the variable map in Model I can be made for the variable map Model II. As shown, there was a good spread of persons along the ruler or the FRT continuum. The risk tolerance measures (ability measures exhibited by the respondents) in this sample ranged from -5.74 to 5.66, measured in logits (M = -0.29, SD = 1.20, N = 11,905). From the items measures, it can be observed that there was an overall good spread of items across the FRT continuum. The location of the items ranged from -1.14 to 1.25 in logits (M = 0.00, SD = 0.70, N = 10). As with Model I, in comparison, items were anchored at 0.00 logits, and person measures were allowed to float along the continuum accordingly. As mentioned

previously, the Wright map is a good visual aid to examine the item hierarchy of the scale (based on *endorsabilities*), which can be thought as of evidence for construct validity (Smith, 2002). It was expected that the item location in the revised scale would not change substantially, as this parameter (i.e., item *endorsabilities*) was stable. From the map, it can be observed that the locations of items were conserved. Though there was a good spread of items across the continuum, there were a couple of noticeable gaps that should be addressed to improve the measurement quality of the scale. For example, it can be observed that there was a gap between the location of item 6 (measure = 0.36 logits) and item 13 (0.95). Respondents with a risk tolerance between ranges of 0.36 and 0.95 are not well targeted with this scale. Similarly, there were gaps between item 5 (measure = -0.47) and item 11 (measure = -0.01), and between item question 4 (measure = -1.14) and item 1 (measure = -0.59). Additionally, extreme respondents (persons with extreme FRT values) were not well targeted with this scale.

Category probability functions and conditional probability curves.

Conditional probability curves for the items in the revised version of the GL-FRT scale are presented in Figure 5. Category probability functions are a good visual representation of the probability's relationship between category difficulty or *endorsability*, and the respondent's ability (FRT in this case). Each of the curves represents an individual response category, with the lowest response category being the farthest to the left, and the curve for the highest response the farthest to the right (Engelhard & Wind, 2013). PCM allows each category width to vary, and consequently the location of the category curve peak may differ across items. The category probabilities curves are useful in identifying categories that are never the most probable along any point on the continuum and are

useful for examining disordered categories. From Figure 5, it is apparent that for question 2, category 3, the probability curve reveals that there is only a slight probability that it will be selected when a respondent possesses an ability (or FRT in the context of this dissertation) of approximately 1.00 in logits. As shown, the peak for this curve is somewhat overshadowed by another probability curve--that of category 2 and 4. When inspected, this was an aspect that the quantitative indicators (e.g., category usage, and category coefficient) had previously revealed. Similarly, category 2 for question 4 exhibited the same problem. For the rest of the items, it can be seen that the category probability curves have very distinctive curves.

Conditional probabilities curves were also examined. Figure 6 presents the conditional probabilities curves for all 10 items in the revised scale. The conditional probability curves follow dichotomous logistics that are able to visually illustrate the relationship between probabilities for observed pairs of adjacent categories. In other words, each category function represents two categories (Engelhard, 2013; Engelhard & Wind, 2013). For item 2, for example, the conditional probability curve in red (farthest to the left) represents response category 1 and 2. From Figure 6, it can be seen that there was a notable lack of category distinctiveness for response categories 2 and 3, and 3 and 4 for question 2. This particular observation was also noted when evaluating the category coefficient order. This similar issue was observed for response categories 1 and 2, and 2 and 3 for item 4.

Item information, category information curves, and test information. Figures 7, 8, and 9 present the item information curves, category information curves, and test information functions, respectively. Item information curves represent the amount of

model-based Fisher statistical information provided by an item at a different locations on the latent construct (Engelhard & Wind, 2013; Fisher, 1958). In other words, item information can be thought as the precision of measurement yielded by each item. The visual aid can be used to identify locations along the latent construct at which the information is most useful for providing statistical information, where the *x-axis* represents the latent construct in logits and the y-axis is the information in logits yielded by the item. The peak of the curve (high values on the y-axis) means the item is measuring with more precision or higher information those respondents that match these locations on the latent construct (*x-axis*). Take for example, the item information curve for question 11 in Figure 7. This curve shows that item 11 more precisely measures respondents with a FRT measure of approximately 1.0 logits. Further, category information can be interpreted likewise. But to do so, it is in terms of categories for each item, instead of items themselves. Thus, categories information curves are helpful in identifying the level of precision for which each category has when measuring respondents with a particular ability measure on the continuum. Figure 8 presents the information category curves for all the items used in the revised version of the GL-FRT scale. Finally, a test information curve is analogous to item and categories information curves, but in this case, the information is in terms of the instrument or scale. Figure 9 presents the test information curve for the revised form of the GL-FRT scale. As seen, the test is best measuring (or assessing with more precision) those respondents with a FRT measure in the range of -2.61 to 2.70, approximately. The peak of the information curve is approximately 1.00 (in logits). It might help to think of information as the reliability for items. Also, note that the inverse of the information function (whether it is the item,

category, or test function) is the standard error function (Baker, 2001; Baker & Kim, 2004). Thus, where information is lower, the standard errors are higher. In the context of item information function and of the scale evaluated in this section, it can be noted that the revised version of the scale measures with less precision those individuals with very extreme FRT measures; for example, smaller than -6.6 logits or greater than 6.6 logits.

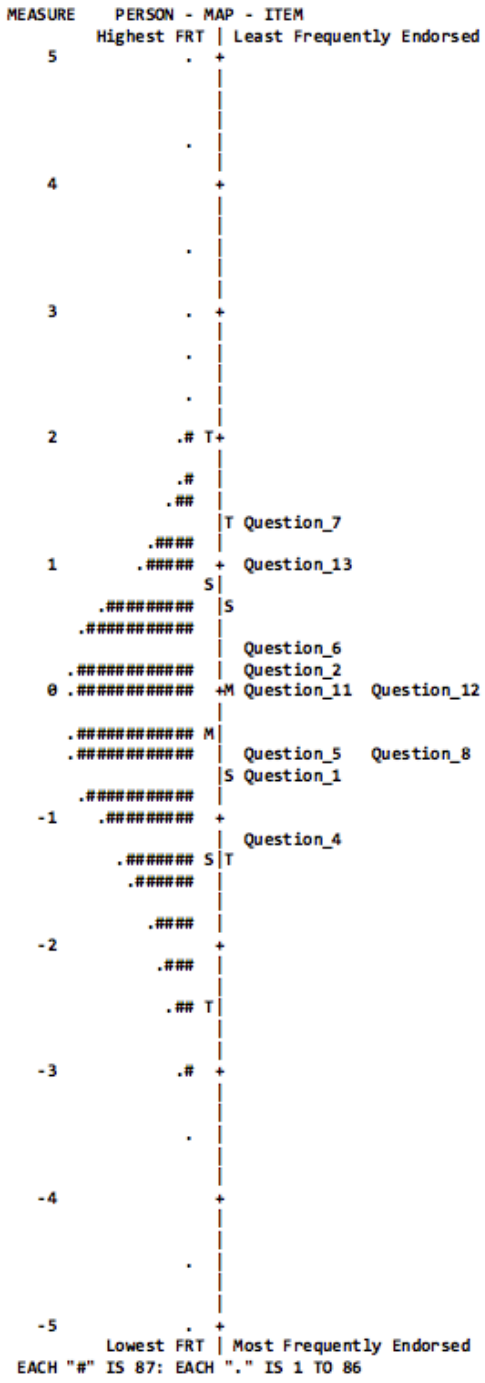


Figure 4. Wright map for the GL-FRT scale (Model II): Items' location (*endorsabilities*) and respondents' ability are both depicted on the continuum (in logit units).

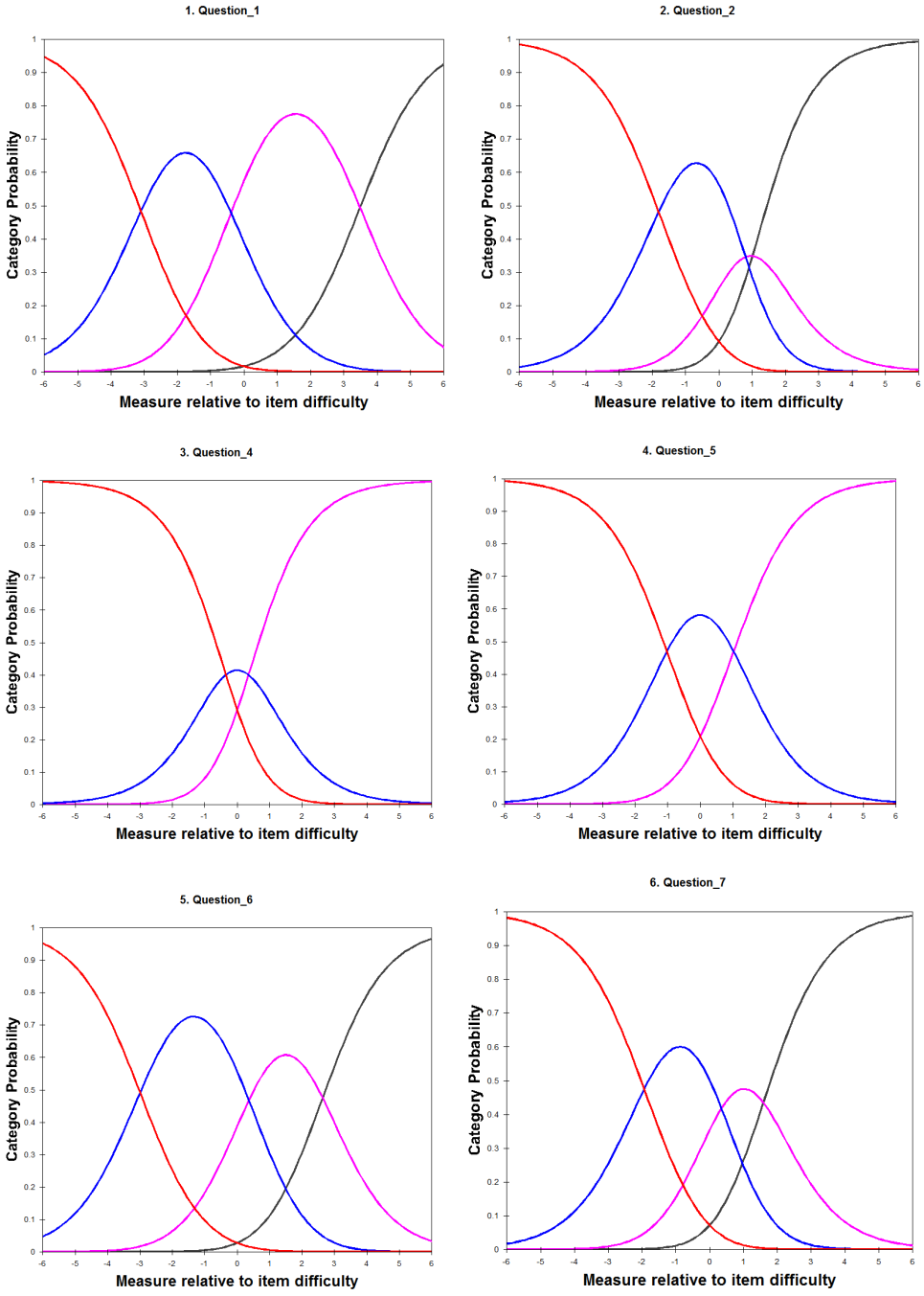


Figure 5. Category probability functions for items in Model II.

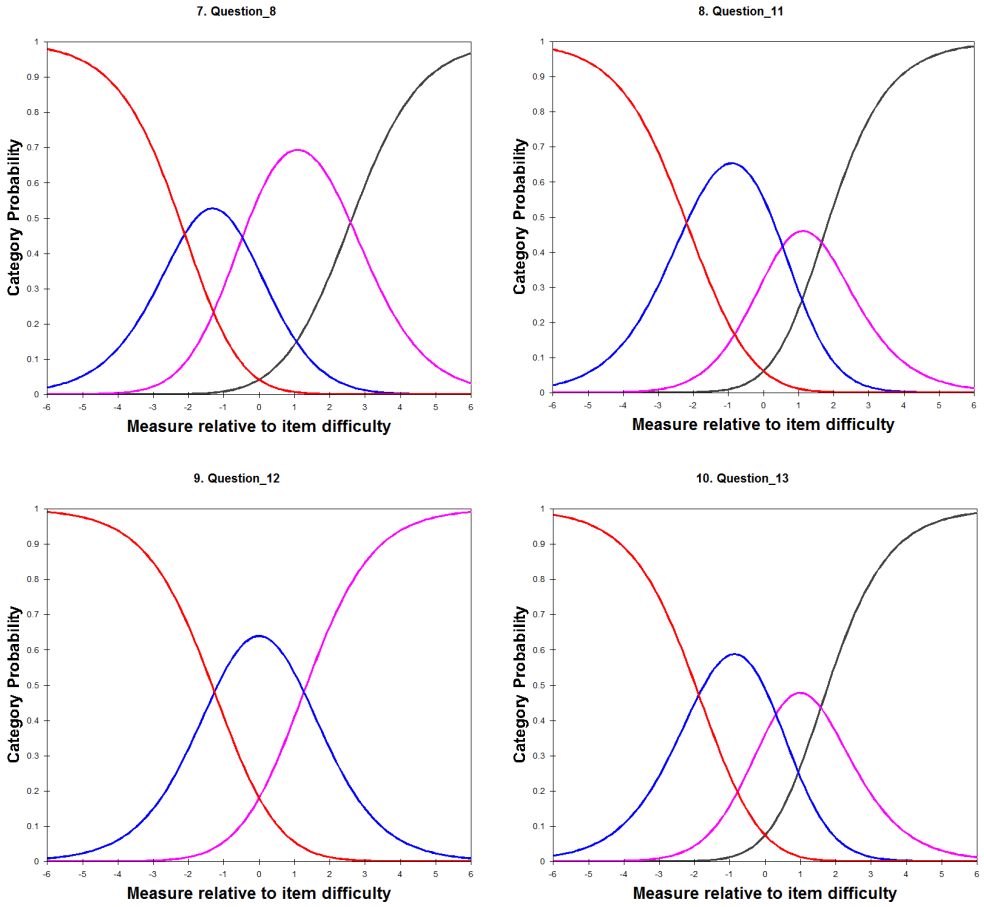


Figure 5. Category probability functions for items in Model II (continued).

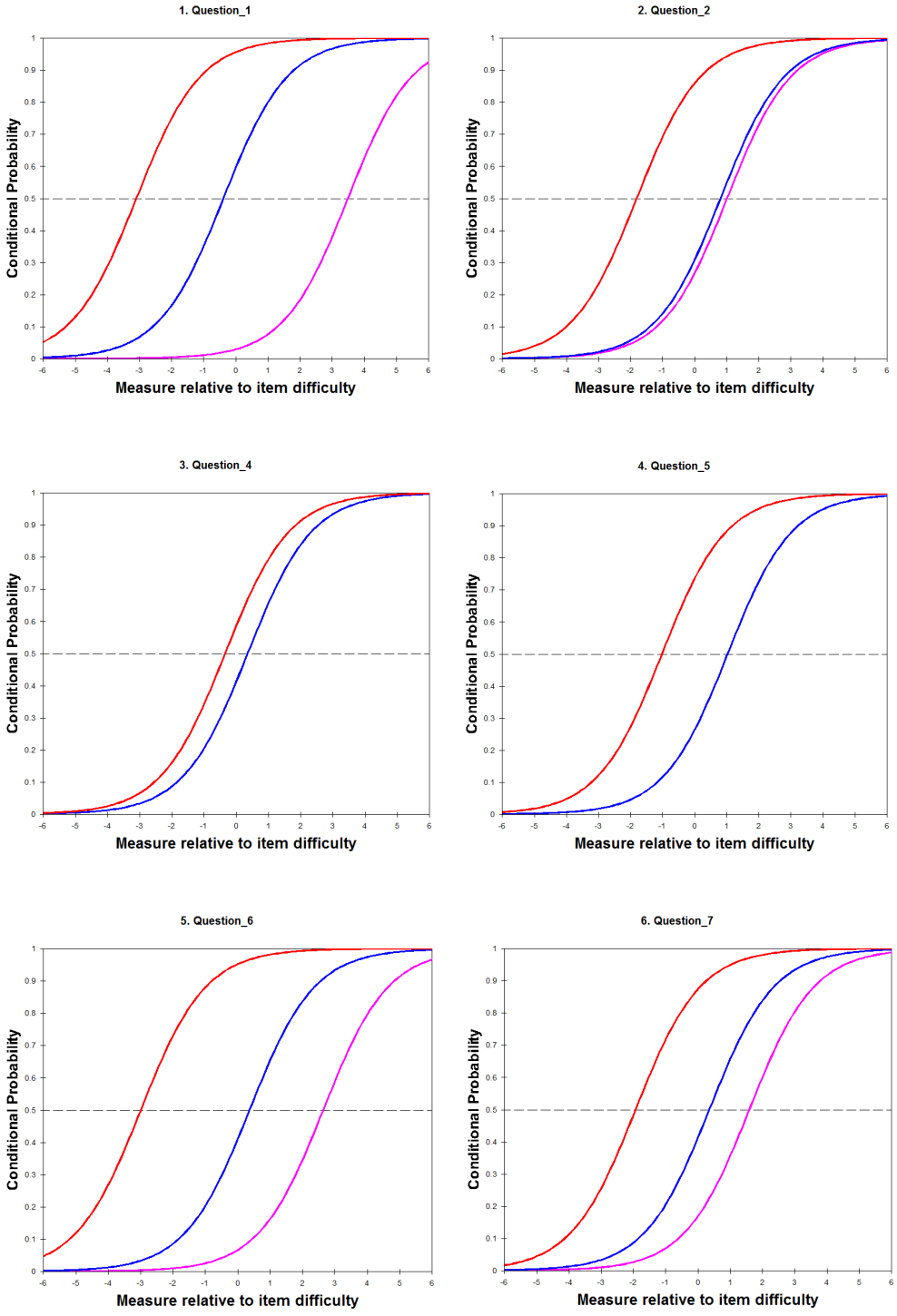


Figure 6. Conditional probability curves for items in Model II.

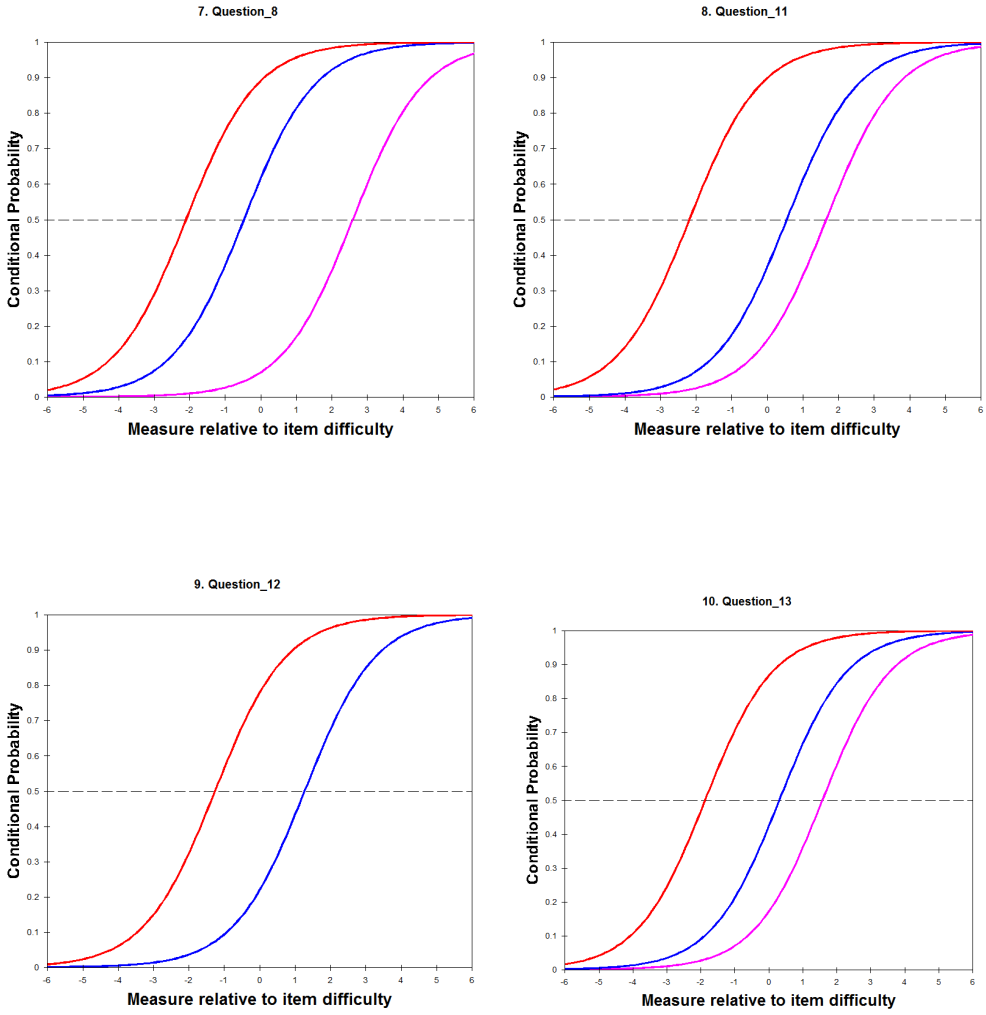


Figure 6. Conditional probability curves for items in Model II (Continued).

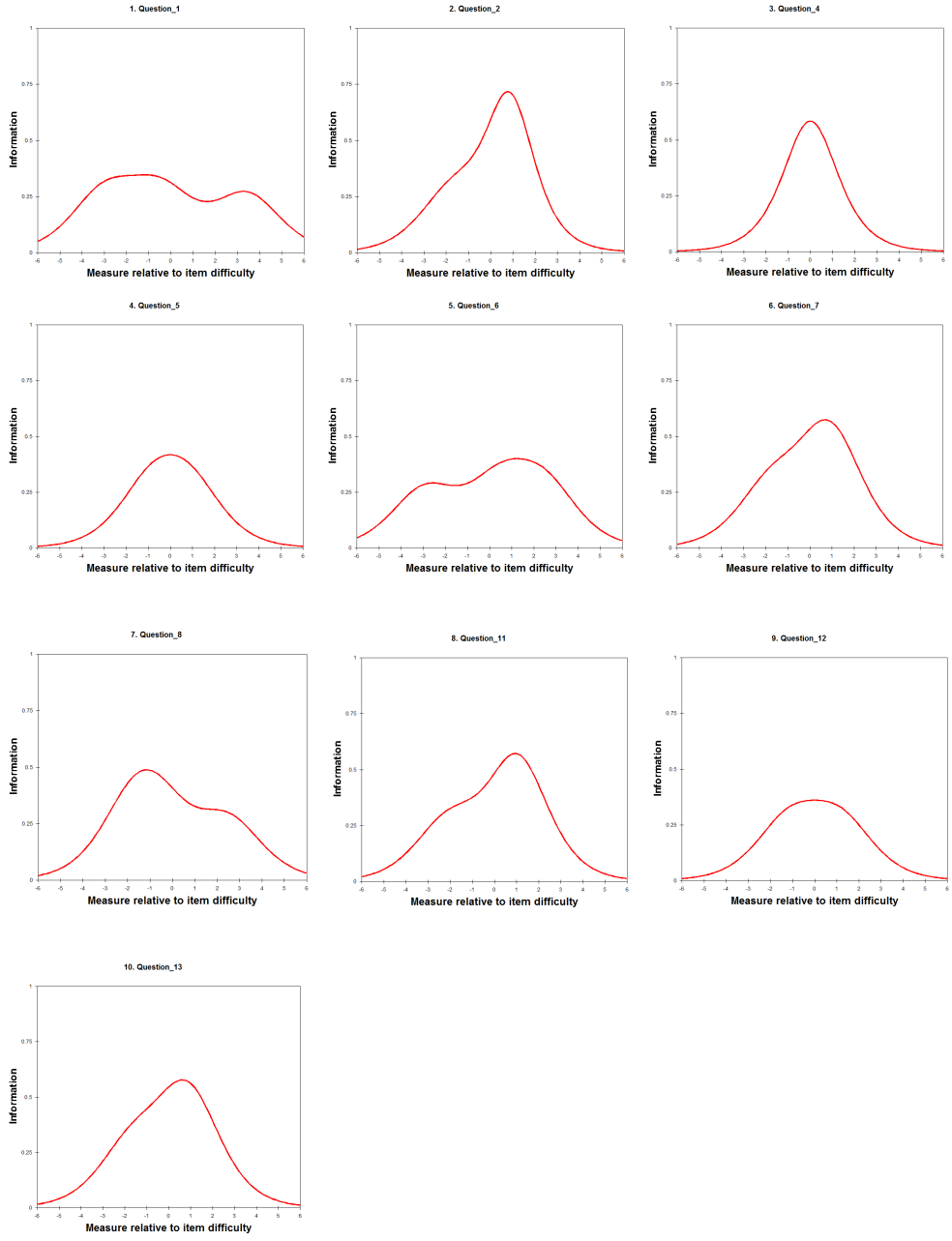


Figure 7. Item information curves for items in Model II.

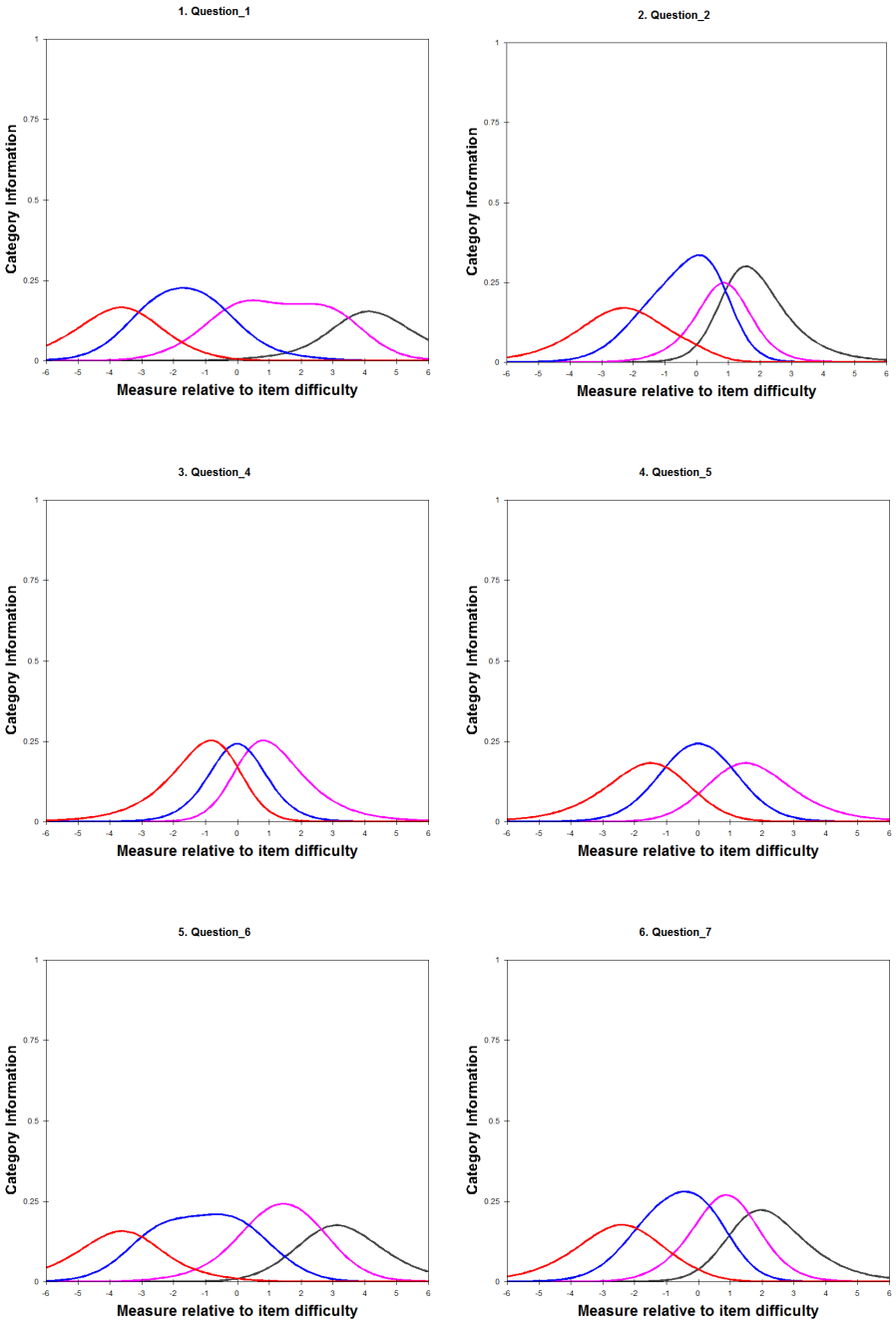


Figure 8. Category Information Curves for items in Model II.

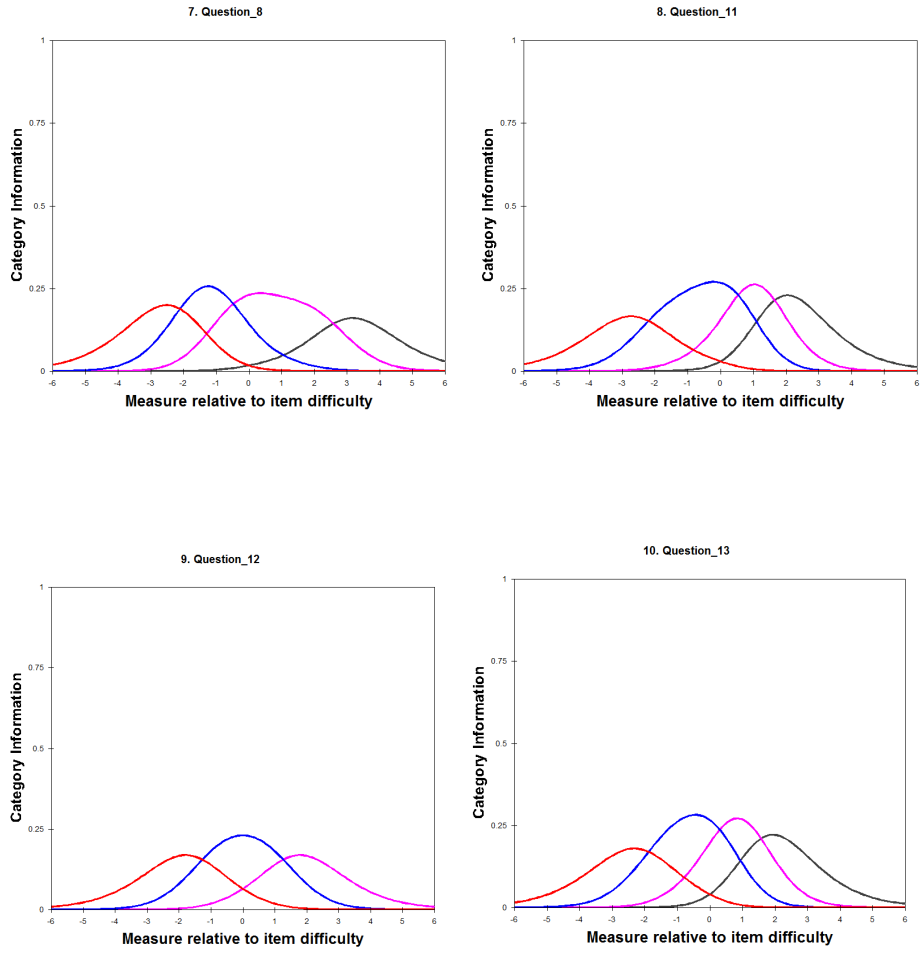


Figure 8. Category Information Curves for items in Model II (Continued).

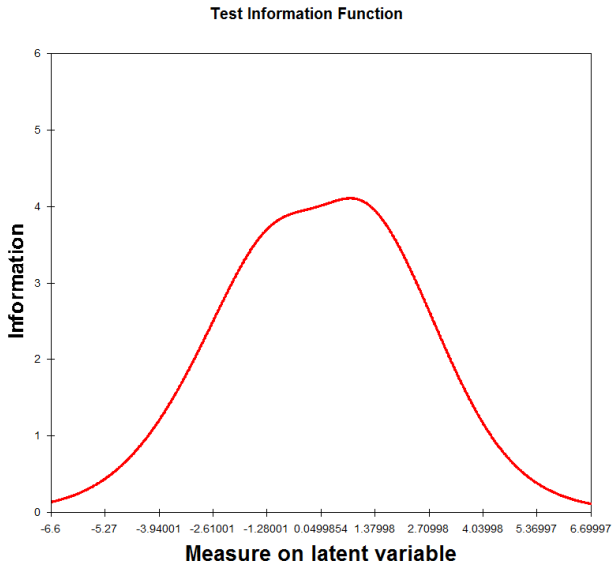


Figure 9. Test information function for the condensed version of the GL-FRT scale.

Presentation of a Person's Ability Estimation and Person's Fit Statistic Tool

In this section, an automatic and interactive tool for the estimation of a person's ability, standard error, and person's misfit is presented. The estimations are based on the revised version of the GL-FRT scale analyzed in Model II. Appendix C presents the 10 questions included in this version of the scale, as well as the scoring system and the proportionally adjusted cut-off values. The calibrated item parameters from Model II were used for the computations in this tool (Table 7 presents these parameters).

Specifically, Excel template allows for the calculation of the FRT measure (in logits) by entering the raw score (or answers to each question in the scale)[28]. Using the

[28] The ability parameters shown in the spreadsheet correspond to the Rasch transformed scores presented in Table 8. Such parameters were estimated via JMLE using Winsteps. In Rasch measurement, two respondents with the same raw score receive the same ability parameter (Baker, 2001). This is a main difference between Rasch model and other Scaling Tradition models (e.g., IRT models: 2 parameter model and 3 parameter model). A brief description of the JMLE estimation is provided in Chapter 2 of this dissertation.

scoring system and cut-off values pre-determined by Grable and Lytton (1999), the automatic tool generates the categorical level of FRT associated with that score (i.e., low risk tolerance, below-average risk tolerance, average/moderate risk tolerance, above-average, and high risk tolerance). In addition to the FRT measure estimation, the tool automatically computes a standard error and fit statistic for each person. A misfit value is indicative of an outlier response pattern. Therefore, qualitative assessment is recommended for that particular person.

Figure 10 shows a screenshot of the template created with Excel. As shown, the template requires the user to manually enter responses to the scale. An example is shown in Figure 10, where person A has a raw score of 25. Based on the raw score, the associated Rasch transformed score in logits is provided, as well as, the corresponding standard error. Based on cut-off values[29], a nominal category is assigned to the score. In this case, a raw score of 25 is equivalent to a Rasch transformed score of 0.48, which can be interpreted as above-average FRT. Further, the Rasch transformed score can be used to locate the person's ability on the continuum using the variable map provided in Figure 4. Finally, an outfit statistic is provided. Appendix D presents the formulas used to compute the outfit MSNQ value (Engelhard, 2013). In this example, the response pattern is inconsistent across items, and indicative of an outlier. This scale might not be the most optimal measure for assessing the FRT of this individual; therefore, the comment section

---

For those interested in reading further about how to calculate these parameters using different methods (e.g., CMLE, JMLE), see Baker and Kim (2004). Additionally, those interested in specifically understanding how Winsteps calculates the ability parameter via JMLE, see Cohen (1979) and Linacre (2004).

[29] Cut-off values for the revised from of the GL-FRT scale were adjusted proportionally to match the ones originally suggested in Grable and Lytton (1999). The original cut-off values of the scale are available in Appendix. And, the proportionally adjusted cut-off for the revised scale are presented in Appendix C.

on the spreadsheet advises the user (e.g., financial planner, financial counselor, etc.) to proceed with a more in depth, personalized qualitative assessment of this construct.

An additional version of this tool was created; and a visualization of this version is presented in Figure 11. This template was created with the notion of the assessment of a single individual's FRT (John Doe in this example) at different points of time (6/1/15 and 1/1/16 for this example). For instance, a financial planner might be interested in creating a file where a log sheet with the record of the FRT assessment at different points of times for a client is maintained. As illustrated in Figure 11, the template allows the planner to record the date the evaluation of FRT was conducted. It provides similar information (i.e., FRT measure in raw score, Rasch transformed score, fit statistic) as in the previous version of the template shown in Figure 10.

Additionally, for the revised version of the scale, the Rasch transformed score can be obtained without using this automatic tool. Table 8 provides a Raw score to Rasch Transformed score conversion chart. The standard error and the corresponding FRT category (determined based on the cut-off values recommended by the authors of the scale) are provided.

Finally, for those who prefer to utilize the original version of the scale (13 questions), a raw score-to-Rasch transformed score chart is provided in Table 9. Note that results from Model I showed that a set of items (i.e., questions 3, 9, and 10) were problematic due to poor fit. However, this was determined using conservative heuristics provided by Wright and Linacre (1994). If a more liberal benchmark (such as the one presented in Liu [2010] and Bond and Fox [2007]) is utilized to evaluate the quality of

the original form of the GL-FRT scale, then overall, the psychometric quality of the scale could be deemed as acceptable.

Summary of Chapter Four

This chapter has presented the results from a two-stage analysis procedure. First, the results of the application of the PCM to the original 13-item scale were reported. Later, the results from the Rasch analysis to the revised version of the scale were provided. Lastly, using the items parameters obtained from the Rasch analysis in the second stage, a tool to estimate persons' FRT measures, individual standard errors, and fit statistics was presented. The remainder of this dissertation presents the implications, limitations, and future research directions derived from this project.

Intstruction: Enter the responses to the scale

Person ID/name	Question 1	Question 2	Question 4	Question 5	Question 6	Question 7	Question 8	Question 11	Question 12	Question 13
A	3	3	2	1	2	3	2	3	2	3
B	1	2	3	3	2	3	2	1	2	3
C										
D										
E										
F										
G										
H										
I										
Calibrated Item Parameters	-0.59	0.17	-1.14	-0.47	0.36	1.26	-0.57	-0.01	0.03	0.95

Person ID/Name	Raw Score	Rasch-Transformed Score	Standard Error	FRT Category*	Outfit Statistics	Comment
A	24	0.23	0.50	Above average FRT	0.88	Proceed
B	22	-0.27	0.50	Average/Moderate FRT	1.77	more assessment neeed
C	0					
D	0					
E	0					
F	0					
G	0					
H	0					
I	0					

*Note. FRT categories were adjusted proportionally from the cut-off values suggested in the original version of the GL-FRT scale.

Figure 10. Screenshot of Excel template for Multiple Persons' Abilities Estimation and Fit Statistic Tool

Name of client: John Doe

Intstruction: Enter the responses to the scale indicated by the respondent(s)

Round	Items										Date administed
	Question 1	Question 2	Question 4	Question 5	Question 6	Question 7	Question 8	Question 11	Question 12	Question 13	
1	3	3	2	1	2	3	2	3	2	3	6/1/15
2	1	2	3	3	2	3	2	1	2	3	1/1/16
3											enter date
4											enter date
5											enter date
6											enter date
7											enter date
8											enter date
9											enter date
Calibrated Item Parameters	-0.59	0.17	-1.14	-0.47	0.36	1.26	-0.57	-0.01	0.03	0.95	

Round	Date administed	Raw Score	Rasch-Transformed Score	Standard Error	FRT Category	Outfit Statistics	Comment
1	6/1/15	24.00	0.23	0.50	Above average	0.88	Proceed
2	1/1/16	22.00	-0.27	0.50	Above Average	1.77	Need more assessment
3	enter date						
4	enter date						
5	enter date						
6	enter date						
7	enter date						
8	enter date						
9	enter date						

Figure 11. Screenshot of Excel template for a Person's Ability Estimation and Person's Fit Statistic Tool Different Points of Time.

Table 8.

Raw Score, Rasch-Transformed Score, and standard Error for the revised version of the GL-FRT scale

Raw Score	Rasch-Transformed Score	SE	FRT categories
10	-5.75	1.87	Low Financial Risk Tolerance
11	-4.42	1.08	
12	-3.56	0.81	
13	-3.00	0.70	
14	-2.57	0.63	
15	-2.20	0.58	Below-Average Financial Risk Tolerance
16	-1.88	0.55	
17	-1.58	0.53	
18	-1.31	0.52	Average/Moderate Financial Risk Tolerance
19	-1.04	0.51	
20	-0.78	0.51	
21	-0.52	0.50	
22	-0.27	0.50	
23	-0.02	0.50	Above-Average Financial Risk Tolerance
24	0.23	0.50	
25	0.48	0.50	
26	0.72	0.49	High Financial Risk tolerance
27	0.97	0.49	
28	1.21	0.50	
29	1.46	0.51	
30	1.73	0.52	
31	2.01	0.54	
32	2.31	0.57	
33	2.65	0.61	
34	3.06	0.67	
35	3.58	0.78	
36	4.39	1.06	
37	5.67	1.86	

Note. FRT categories were adjusted proportionally from the cut-off values suggested in the original version of the GL-FRT scale.

Table 9.
Raw Score, Rasch-Transformed Score, and standard Error for original GL-FRT.

Raw Score	Rasch-Transformed Score	SE	FRT categories
13	-5.45	1.86	Low Financial Risk Tolerance
14	-4.15	1.06	
15	-3.34	0.78	
16	-2.83	0.66	
17	-2.45	0.58	
18	-2.14	0.53	
19	-1.88	0.49	Below-Average Financial Risk Tolerance
20	-1.65	0.46	
21	-1.45	0.44	
22	-1.26	0.42	
23	-1.08	0.41	Average/Moderate Financial Risk Tolerance
24	-0.92	0.40	
25	-0.76	0.40	
26	-0.60	0.39	
27	-0.45	0.39	
28	-0.30	0.39	
29	-0.15	0.39	Above-Average Financial Risk Tolerance
30	0.01	0.39	
31	0.16	0.39	
32	0.31	0.40	
33	0.47	0.40	High Financial Risk Tolerance
34	0.64	0.41	
35	0.81	0.42	
36	0.99	0.43	
37	1.18	0.44	
38	1.38	0.46	
39	1.60	0.47	
40	1.83	0.49	
41	2.08	0.52	
42	2.37	0.55	
43	2.69	0.59	
44	3.08	0.66	
45	3.59	0.78	
46	4.39	1.05	
47	5.66	1.85	

Note. The original version of the GL-FRT is based on 13 financial risk tolerance (Grable & Lytton, 1999). The FRT categories are based on the cut-off values suggested by the authors of the GL-FRT scale.

CHAPTER 5

DISCUSSION

Summary

The development, evaluation, and improvement of financial risk tolerance (FRT) assessment continues to be an important are of research among scholars and professionals. The implementation of accurate and consistent FRT measures is crucial, as it is a gauge through which both financial planning researchers and counseling professionals implement their advice and recommendations for consumers. Refining the accuracy of FRT assessment requires dedicated researchers who are committed to understanding and predicting individuals' financial decisions and behaviors under uncertainty. Research findings can then be applied to individualize increasingly precise recommendations.

In recent years, regulatory and licensing board entities in the financial planning profession have issued rules of best practice as guidelines to foster a more careful and diligent assessment of each client's FRT—especially when advisors and financial planners are presented with the task of creating investment plans for clients (CFP, 2015; FINRA, 2011; 2012a; 2012b). Within the profession, as the measurement of FRT becomes more prevalent, critiquing the quality of instruments employed to assess FRT will continue to garner greater attention. Ultimately, a measurement tool should consistently and accurately measure FRT so that financial advice and strategies are

appropriate, realistic, and abide to the standards of the profession, which have been specified by the profession's regulators.

Similarly, in academia, researchers and scholars have increasingly been recognizing the prerequisite of valid and reliable measures of FRT during the process of modeling and predicting financial and economics behaviors under risk. Over the years, important efforts have been devoted to refining and improving FRT measures for accuracy and in consistency in the measurement of the elusive FRT construct (e.g., Carr, 2015; Grable, 2010; Grable & Lytton, 2001; Hanna, Gutter, & Fan, 2001; Hanna & Lindamood, 2004; Kuzniak et al., 2015; Nobre & Grable, 2015). Scholars have developed instruments using theoretical approaches (i.e., Classical Test Theory) and also empirical perspectives. The composition of FRT has been examined through the evaluation of its properties (e.g., item selection, word choice, questionnaire length) and the psychometric quality (i.e., reliability and validity) of existing instruments. Additionally, these investigations have been to document the association of relevant reference financial behaviors, such as actual portfolio allocation, with FRT. Ultimately, researchers have sought to advance FRT assessment research by developing effective, improved, and efficient metrics to capture the FRT latent construct. Doing so is important as these metrics of latent constructs can possess good predictive power for modeling the complexities of financial human behavior—from everyday economic decisions to more formal investment choices, such as retirement portfolio allocation.

A widely known, but little researched, are of concern among FRT researchers is that many existing FRT instruments lack a formal theoretical psychometric background to guide test construction, evaluation, test interpretation, and subsequent maintenance

(Grable & Schumm, 2010; Ruiz-Menjivar et al., 2014). The incorporation of measurement theory is crucial to fill this gap in the literature. Measurement theory can be used, as illustrated in this dissertation, to provide insights into the evaluation practices, guidelines, and methodologies to examine and improve the precision and accuracy with which instruments measure FRT (Croker & Algina, 2006; Engelhard, 2013; Messick, 1993; 1995; Roszkowski, Davey & Grable, 2005).

It is important to note, however, that some of the existing FRT measures have incorporated the use of formal psychometric test procedures during the development and evaluation process (e.g., GL-FRT scale and the SOFRT scale)—though these are exceptions and not the rule (Grable & Lytton, 1999; Grable, 2000; Roszkowski, Davey & Grable, 2005). From psychometric theory, CTT analysis has typically been the chosen framework to develop and evaluate FRT measures. CTT's acceptance and popularity lies in its historical acceptance among researchers and it conceptual and theoretical simplicity (Hambleton & Swaminathan, 1984; Liu, 2010). Researches who use CTT must accept certain tradeoffs within the scale development process. For example, CTT tends to be based on weaker theoretical assumptions. Some of these assumptions (e.g., that a concept can be measured effectively using an ordinal scale) have been identified among quantitative test developers as potential limitations when scale scores are used primarily for research purposes (e.g., Rasch Measurement theory and Item Response Theory).

In an effort to promote and support the continuing discussion on ways to measure FRT with more reliability and validity, this dissertation presented an alternative probabilistic methodology with a strong mathematical foundation (i.e., Rasch Measurement Theory) to evaluate the psychometric quality of FRT scales. Rasch

Measurement Theory has been extensively used in different academic and professional disciplines, ranging from educational psychology and applied psychology, to health sciences and clinical and behavioral sciences, such as medicine and nursing (Engelhard, 2013). The work here demonstrated the application of this theory to a widely used FRT instrument (i.e., GL-FRT) in the field of financial planning and consumer economics. A revised version of the GL-FRT scale, with improved psychometric properties, was the outcome of this project. Scoring and interpretation guidelines were made available for this new version.

Discussion of Results

Rasch Measurement Theory was selected as the theoretical framework to complete this dissertation. The primary reason for this selection was that Rasch Theory proposes a set of probabilistic models developed for the purpose of describing response patterns of respondents to individual items. This framework provides some advantages over other measurement approaches, such as CTT, in the sense that measures obtained with a Rasch analysis are expressed on an interval scale. Additionally, items' measures are not sample dependent nor vice versa (i.e., persons' measures are not dependent on the items included in an instrument). Also, individual errors for each subject are provided that can be used in practice to further examine a test taker's score.

Specifically, the Wright and Masters (1982) Partial Credit Model was employed to evaluate the psychometric quality of the GL-FRT scale. Unlike other available versions of Rasch models (e.g., dichotomous Rasch model or rating scale mode), PCM allows each item to have its own structure. Given that the GL-FRT scale has diverse response category structure across the set of items, PMC was deemed suitable for the

analysis. As mentioned earlier, the sample of 11,905 respondents (delimited to individuals age 25 or older) and 154,765 responses were analyzed (Table 2 present the sample profile). The data for the analysis came from a repeated cross-sectional data collection project hosted by Rutgers New Jersey Agricultural Experiment Station.

The PCM application to the original version of the GL-FRT (i.e., Model I) provided mixed results in terms of the psychometric quality of the scale—as per conservative evaluation benchmarks suggested in the literature (Linacre, 2006; Wright & Linacre, 1994). In fact, after a thorough assessment of a series of quantitative and visual evaluation criteria, areas for improvement were identified. Guidance on how to improve the quality was obtained from different sources in the literature (Engelhard; 2002; 2013; Engelhard & Wind, 2013; Linacre, 2002; 2006; Wright & Linacre, 2006). Thus, a second analysis was performed using PCM to evaluate the psychometric quality of a reduced version of the GL-FRT scale (Model II). A sample of 11,905 and 119,050 responses were evaluated during this phase of the study. For organizational purposes, the discussion of these results is divided in terms of Model I and Model II.

Model I. The primary concern that emerged with the analysis of the first model involved the notion of *unidimensionaltity* of the scale. Results from the Rasch analysis showed that the original version of the scale might be measuring more than one construct concurrently. This was not a surprise because Grable and Lytton (1999) noted that the original scale was thought to be comprised of three dimensions. Within the Rasch framework, however, this is a problem. The Rasch model requires a scale to consist of one primary construct in order to develop a stable "yardstick" composed of a set of psychometrically sound quality items. The presence of multiple latent constructs in a

single scale introduces noise, and ultimately hinders the precision and accuracy with which a single construct can otherwise be measured. Thus, *unidimensionality* is a desired property for constant measures that are designed for the purpose of creating a research measure. This scale requirement, among researchers, comes from an idea borrowed from the physical sciences (Andrich, 1978a; 1978b; Liu, 2010). For instance, if one wants to determine the weight and height of an object, then two different calibrated and valid instruments (e.g., a valid and well calibrated scale and a well-marked tape measure) are needed—one to measure weight and one to measure height. Each instrument should be one dimensional. The procedure itself involves the measurement of two properties, but each is individually obtained. Similarly, within a Rasch model framework, for example, a researcher can administer a questionnaire to measure two distinctive latent constructs. Though the procedure or the administration of the survey is done in one sitting, the items used to measure each construct should be comprised of separate and stable sets of questions in order to yield reliable and valid scores or measures.

From Model I, it was observed that item 3, 9, and 10 were problematic in terms of *unidimensionality*. The fit statistic used (i.e., outfit MNSQ) showed that these items poorly fitted the Rasch model, and consequently introduced substantial noise or randomness into the estimation of FRT measures. Upon further assessment of items 9 and 10, it was noted that these particular questions were based on Prospect Theory—a behavioral economic framework that describes how individuals make decisions based on the potential value of losses and gains, rather than the ultimate outcome itself (Kahneman & Tversky, 1979). In particular, questions 9 and 10 of the GL-FRT were adapted from the seminal economic paper by Amos Tversky and Daniel Kahneman (1994) entitled

"Judgment Under Uncertainty: Heuristics and Biases" to describe a complex behavioral bias in which individuals with aversion to loss react to financial losses more profoundly than the satisfaction that might stem from gains of equal size. Though loss aversion is related to risk aversion (i.e., inverse of FRT), subtle differences do exit. The former describes a cognitive bias (i.e., the tendency to prefer options that avoid losses rather than acquiring gains), while the latter refers to a preference for certainty over uncertainty (or risk) (Erev, Ert, &Yechiam, 2008; Ert & Erev, 2008; Gal, 2006). Thus, it was not surprising that the Rasch analysis suggested the existence of secondary dimensions measured by the original form of the scale. The other problematic item, question 3, related to risk-taking behavior associated with life style risk rather than financial risk.

Model II. With the revised version of the scale, the issue of *unidimensionality* was addressed, as evidenced by all 10 remaining items fitting reasonably well with the use of the Rasch model. However, other complications were identified. For example, the category functioning for some questions was not optimal. More specifically, this was seen in some of the categories for question 2 in the revised version of the GL-FRT scale (See appendix C for the complete list of questions and scoring system). Here, the results from the Rasch analysis suggested that response category 3 did not function well, as it was overshadowed by response categories 2 and 4. In practice, this lack of category distinctiveness may indicate that category 3 is almost never the most probable along any point on the continuum. Collapsing categories is often a solution recommended to address this issue (Linacre, 2006). One possible solution involves revising the item. At a minimum, future research should be conducted to determine if the finding noted here is consistent across samples or whether the result was an artifact of the sample.

Additionally, the Wright variable map (shown in Figure 4) provided information in terms of item hierarchy and construct representatives. The map can be thought of as evidence for construct validity (Smith, 2002). As illustrated, the map showed that although there was a good spread of items across the FRT continuum with the revised scale, there were noticeable gaps among items. For example, a gap was observed between the location of item 6 (measure = 0.36 logits) and item 13 (0 = 0.95). This is suggestive that individuals with FRT measures between the ranges of 0.36 and 0.95 may not be well targeted with this scale. One solution involves creating a new question or item that can be used to fill the gap. This would increase the number of items in the revised scale, but the result could provide a more robust scale scoring system for those who need an interval scale. Similar gaps were identified between item 5 (measure = -0.47) and item 11 (measure = -0.01) and between item 4 (measure = -1.14) and item 1 (measure = -0.59). Adding items with locations along these gaps could lead to better measurement precision of individuals with FRT measures in these sections. Finally, from the variable map and the test information function, extreme respondents (persons with extreme FRT values) were not well targeted with this scale. This was observed when inspecting the test information function (Figure 9). Essentially, respondents with a FRT measure below -4.00 logits and above 4.00 logits were assessed with less precision. This is only a problem, however, for scale users who are working with very risk-averse or extremely risk-tolerant populations. Based on a reading of Grable and Lytton (1999), it does not appear that the scale was designed for either of these groups. The original and revised scales are more appropriate for use with "average" respondents or those who fall within the following range: below-average, average, and above-average FRT.

Implications

This study was undertaken to advance FRT research assessment through the introduction of the Rasch measurement paradigm with the expectation that it will potentially lead others to use the framework in the development, evaluation, and refinement of other FRT measures. Several inferences and implications can be drawn from the results of this dissertation.

First, the evaluation of a widely used FRT scale (i.e., GL-FRT), employing Rasch Measurement model, resulted in the presentation of a refined form of the GL-FRT—a new version that possesses rigorous and strong psychometrical properties. The culmination of these properties led to a new version that can be conceived of as the end-product of this dissertation. This scale is available in Appendix C. Also shown are the scoring system and recommended cut-off values.

Second, for interpretability purposes, a raw score-to-Rasch transformed score chart was provided in Table 8. With the help of this figure, a scale user (e.g., a researcher, financial planner, consumer, policy maker, etc.) can convert raw scores to Rasch measurement logits, which allows for a test taker to be located according to the person's score on the variable map for the scale (presented in Figure 4). For the convenience of scale users, the cut-off values were proportionally adapted from ones suggested by Grable and Lytton (1999). The FRT categories from the original version (i.e., low FRT, below-average FRT, average/moderate, above-average FRT, and high FRT) were conserved. This allows a straightforward comparison between the original version and the new version. Doing so makes for an easier transition for those utilizing the original form versus the refined version.

Another important outcome from this study was the development of a novel tool designed specifically to be used for this scale that allows for the automatic estimation of respondents' FRT measures in raw score and Rasch scale logits. Such a tool not only provides FRT measures scores but also fit statistics for respondents. This can be used to determine whether or not a person's score is reliable and valid. An indication of misfit for a person's FRT measure can be interpreted to mean that there is a need of following up the FRT assessment with qualitative techniques. In the profession, and in the context of financial planning, this tool can provide advisors and financial planners a convenient, easy-to-use, and efficient mechanism to determine which clients may need a more in-depth examination where FRT assessment is concerned. This, undoubtedly, is a clear implication of this study, and will be a useful contribution for the profession.

Furthermore, the newly presented scale allows for a more systematic and consistent measurement of FRT across samples, studies, and programs, as it possesses invariant measurement properties embedded in Rasch Measurement Theory (Englehard, 2013). The psychometric evaluation of the scale, using a framework with strong theoretical and mathematical properties, yielded important conclusions about the scale. For example, this study was able to provide strong evidence of the suitability or "goodness" of the items included in the revised version of GL-FRT. An important property of scaling tradition models (e.g., Rasch Measurement model) over test-score tradition analysis (e.g., CTT) is that each question contributes differently to the estimation of the FRT measures for individuals—this is based on the location of each item along the continuum. Including items with relatively different locations along the continuum enables users to obtain more precise information about FRT across a wide

range of ability or *endorsability* levels. As mentioned previously, the revised version of the scale has good spread, overall, over the continuum, especially when measuring individuals from below average through high FRT.

Lastly, two characteristics associated with a well-fitted scale, within a Rasch measurement model, are person-free test calibration and item-free person calibration. In practice, the utility of such invariant measurement properties allows researchers flexibility when using the scale according to their needs. For example, adding or removing items from the refined version of the scale will not change or make obsolete the FRT measures compared to obtaining scores from a non-modified form of the revised version. This is true because the stable parameters were obtained through Rasch calibration. Measures are stable even when different forms of the revised version of the GL-FRT are used. Some may ask, in terms of comparability purposes, if different questions are being used in a test administration, can these be compared to the FRT measures of two individuals who responded to different items of the new revised version of the scale? The answer to this question is yes. This is, essentially, the strength of the Rasch model. The ability to remove questions without significantly altering conclusions arises from the technique called equating, which allows the comparison of scores from a Rasch calibrated scale. This has great potential when considering opportunities to engage in cross-cultural studies or research with substantially heterogeneous groups. For example, a researcher studying cross-cultural similarities and differences in terms of willingness to take risk from two different populations, say Americans and Asians, could modify the revised version of the scale accordingly (e.g., some items asking about stock markets in the scale might be content specific for an American audience), and yet, be able

to have reliable, accurate, and fair measures for comparison purposes. Moreover, a researcher can use an equating technique to tie together scores from both groups via a common set of items or what are termed anchor items. This provides a powerful tool for obtaining and evaluating FRT scores from different groups on a common metric without compromising the psychometric quality of the scale or penalizing individuals for being a less targeted audience for the scale.

Yet another implication from this dissertation is the estimation of FRT scores using an interval level of measurement. This is important, as a major assumption for inferential statistical methods, such as *t*-test or *F*-test, is that data be measured on an interval scale. Through usage of the Rasch transformed scores in logits, interval data can be used for the aforementioned methods and other advanced inferential statistical methods without violating the assumption of interval data. For instance, if a researcher wants to measure changes in FRT, interval data is preferred over ordinal data, especially when using parametric methods to assess the size of such changes over time. In the field of medicine, where clinical studies involving experimental trials are prevalent, Rasch measurement models have been widely used for this purpose. It is important to note that some research studies have documented that the use of ordinal data with inferential statistics does not substantially increase the likelihood of a Type I error. From a pure statistical perspective, however, the use of ordinal data does violate the assumption imbedded in many widely used statistical tests. Thus, even if the chance of Type II error is slight, if the Rasch Measurement model provides a more robust scale, the approach should be favored over scaling methods that provide only an ordinal outcome. In other

words, if two tests are available, researchers would be better served by choosing the one that leads to an interval measure of FRT.

The contribution of this project is not exclusive to the FRT literature, as the utilized measurement framework and model are applicable to many types of studies within financial planning and consumer economics. This research can be conceived of as a stepping stone for researchers in these fields to apply the Rasch methodology to other instruments that assess latent constructs of interest (e.g., financial well-being, financial satisfaction, and financial stress). In addition, the use of modern psychometrics in these fields should encourage a reconsideration of how measurement issues might be explored. The Rasch approach permits finer granularity and potentially greater specificity. The ideas of validity and reliability can be evaluated and confirmed using the analysis and tools provided by Rasch Measurement theory. For example, evidence for construct validity can be established using one of the main Rasch analysis tools; namely, the variable map. Construct validity, as described by Messick (1996), deals with the trustworthiness of score meaning and its interpretation. Messick (1989) suggested that a major threat to the meaningfulness of construct validity that could lead to a potentially limited score, resulting in false interpretation, is construct underrepresentation. When used appropriately, the variable map obtained from a Rasch analysis allows for inspection of the spread of items across the continuum. In order words, it allows for the observation of how well represented the latent construct is based on the location of questions measuring different levels of the construct. Another threat to construct validity is construct irrelevance (Messick, 1996). This refers to the presence of unrelated, sub-dimensions that could contaminate the accurate measurement of the main latent construct

of interest. The *unidimensionality* assessment that performed with a Rasch analysis (e.g., fit statistics and principal component analysis of the residuals) aids in assuring that only items that are relevant to the construct of interest are used in the scale to measure a respondent's ability. In short, Rasch analysis provides a powerful tool for evaluating "Messickian" (Baghaei, 2008) construct-validity issues.

In terms of reliability, Rasch and other scaling tradition models, such as Item Response Theory models, allow for the estimation of not only a reliability indicator for persons but also for items. This feature is a powerful tool in the sense that it allows for evaluating both person and items, individually, and with finer granularity. This is a major advance for scale development and refinement, and an advantage over test focused reliability indicators (e.g., alpha or KR_{20}) used in CTT (Boone, Staver, & Yale, 2013). In addition, the visual aid produced by a Rasch analysis, such as test information functions, item information curves, and category information curves allows for the visual inspection of the precision with which test (scale) items and categories are measuring ability at different points on the construct continuum.

As noted in this dissertation, two important notions from Test Theory could also be incorporated into the measurement of latent constructs and scale development and maintenance practices in the field of financial planning and consumer economics. These are consequential validity and test fairness. The former refers to the after effects, and possible social implications, from the score interpretation from a particular measure. For example, consequential validity helps in identifying a test that is not truly measuring what it claims to be measuring. Stated another way, consequential validity can be used to identify a scale that is falsely measuring respondents who have taken the measure. Test

fairness refers to the constant and routine bias analysis of questions to ascertain that items in a scale do not unfairly contribute to group differences (AERA, APA, & NCME, 1999; Engelhard, 1990).

To illustrate this notion, take the following example in the context of FRT. Two investors, A and B, both have the same very high FRT score but substantially different socioeconomic and cultural backgrounds. They also both wish to create an investment plan using U.S. investments. As such, they hire a financial planner. Investor A has lived in a country where there are organized (exchange) markets. She holds financial securities via her retirement plan. Investor B, on the other hand, has lived in a country that lacks a formal stock market, and for Investor B, examples of risky investments are translated into venture capital investments or real-estate investment. The financial planner wants to make portfolio recommendations for both investors. So, following the best practice standards of the profession, the financial planner measures each investor's FRT with a 5-item FRT instrument that contains a few items that are context specific to aspects of organized and well-structured stock markets (for simplicity assume this scale is measuring FRT *unidimensionally* and is reliable). Problematically, the results from the scale administration suggest that investor A has a very high FRT (i.e., correct estimation), while investor B has moderate FRT (i.e., incorrect estimation). If all other variables for these two investors are equal, and FRT is the main aspect taken into consideration for the portfolio allocation, the financial planner correctly suggests a portfolio allocation to investor A that reflects a higher risk and potentially higher returns. But, for investor B the recommendation is a portfolio with moderate risk with potentially for only moderate returns. Thus, for investor B, the recommendations are not optimal given B's level of

FRT. This demonstrates the importance of obtaining accurate scores to optimize interpretations, and how these scores can have an effect and create consequences for portfolio allocation. The differences in scores might be an artifact of biased items. Thus, fairness (whether the items are biased) and consequential validity are pivotal elements to assess. Rasch modeling helps address these issues.

A Rasch model analysis allows for the evaluation on how items are functioning for two different populations. This is called differential item functioning. Following the example from above, it would be possible to use a Rasch model to evaluate whether certain items in the scale are functioning differently or inaccurately when measuring investor B, and thus possibly allow for correction. Later, differences in suggested portfolio allocation based on FRT could be evaluated to determine any consequential validity issue with the scale. These two measurement notions—test fairness and consequential validity—can be informative resources for determining whether fair and objective measures are being obtained across distinctive groups. Cross-cultural comparison studies in financial planning and consumer economics could benefit greatly from the incorporation of these ideas.

Study Limitations

This dissertation, and the revised version of the GL-FRT scale, can be conceived of as a starting point in the ongoing cycle of re-calibration, optimization, re-evaluation, and maintenance of this and other FRT scales (Hattie, Jaeger, & Bond, 1999). Although the benefits and implications of this project, and the psychometrically sound scale, are noteworthy, this study is not exempt from limitations. These must be considered when drawing conclusions. For example, the sample used in this study was based on

respondents who answered survey questions online in 2013. It is possible that had a different sample frame, or a different time period, been chosen the results might have changed. Related to this limitation is the concept of generalizability. While steps were taken to delimit the sample to be representative of a broad range of investors, the actual generalizability of findings to the broader U.S. population is unknown. As such, readers should use caution when extrapolating results to groups that do not match the sample characteristics.

Another limitation is that differential item functioning was not performed in this study. Thus, little is known if different items are functioning optimally across different groups. Unlike CTT's perspective where all items contribute to the scale and function across groups equally, a Rasch analysis allows for item bias tests across groups via differential item functioning. This undoubtedly should be a future research study as a way to determine the objectivity and fairness with which items in the scale measure substantially different groups.

A further limitation is that there was no attempt to qualitatively pre-test each item via focus groups, nor was cognitive interviewing done. This study relied on the fact that the scale has been widely used for over 15 years (Kuzniak et al., 2015), and that previous psychometric work supports the validity and reliability of the scale. Nonetheless, a future endeavor should include an in depth qualitative assessment of the items retained in this scale that could include further review of the already revised categories with functioning problems in the scale (such as question 2, where category "c" seems to have marginal distinctiveness). Similarly, the evidence of validity provided in the study was limited to construct validity via the evaluation of construct irrelevance and construct

underrepresentation in the scale. The notion of consequential validity, an integral part of the "Messickian" conceptualization of validity, was not examined in this project.

Future Research and Recommendations

Several research opportunities, in terms of FRT assessment, emerged from this dissertation. Of primary relevance is the notion that Rasch Measurement Theory can open avenues of exploration for subsequent analysis that can be conducted in the pursuit of a better measurement of FRT. Further analysis of FRT should be done utilizing the Rasch model in financial planning and consumer economics research. Additionally, more research is needed to extend this dissertation into new areas. For example, differential item functioning analysis needs to be employed to determine if items presented in this dissertation are behaving differently when answered by different groups (e.g., men versus women and younger versus older investors). In the context of FRT, groups formed by gender, income level, financial education, cultural background, country of origin, and language need to be considered in future studies. Results from such research could shed insight on how to improve the revised scale further as an approach to better administer the scale to certain groups of respondents.

Numerous applications for those interested in cross-cultural research related to FRT came to light from this study. For example, a comparison of FRT measures with two or more culturally different groups via equating is an intriguing research direction to explore. The concept of fairness and consequential validity are notions to be further examined in the context of cross-cultural comparisons.

More work is needed in terms of the longitudinal assessment of FRT. The revised scale presented in this dissertation provides an opportunity to track individuals over time

with a valid and parsimonious instrument. The ability to measure FRT using an interval scale (logits units) can be explored in comparison to more traditional ordinal scaling traditions. Whether an interval scale actually provides a better level of FRT discrimination is something that was not addressed in this study; however, this question needs to be addressed in a future research project. More specifically, does "intervalness" in data allow for more confidence and fewer statistical violations when estimating the size of changes via parametric and inferential statistical methods? While the answer may seem obvious, little evidence exists in the FRT literature, at this time, to answer the question conclusively. More research is needed. Finally, the continued use of the scaling tradition presented in this dissertation, and in particular Rasch Measurement model, can eventually lead to a computer-adaptive testing (CAT) system to assess FRT more effectively and parsimoniously, without compromising the quality of the scale. This research endeavor is ambitious in nature, but the creation of such a tool will constitute a major breakthrough in the optimal assessment of FRT.

REFERENCES

Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory.* Long Grove, IL: Waveland Press.

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Andersen, E. B (1972a). The numerical solution of a set of conditional estimation equations. *The Journal of the Royal Statistical Society, 34*, 42-54.

Andersen, E. B. (1973). *Conditional inference and models for measuring*. Copenhagen, Denmark: Mentalhygiejnisk Forlag.

Andrich, D. A. (1978a). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2*, 581–594.

Andrich, D. A. (1978b). A rating formulation for ordered response categories. *Psychometrika, 43,* 561–573.

Andrich, D. (1988). *Rasch models for measurement*. Beverly Hills: Sage Publications.

Baghaei, P. (2008). The Rasch model as a construct validation tool. *Rasch Measurement Transaction, 22*, 1145-1146. Available: http://www.rasch.org/rmt/rmt221a.htm.

Baker, F. (2001). *The Basics of item response theory*. College Park, MD: Clearinghouse on Assessment and Evaluation-University of Maryland.

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

Barsky, R. B., Juster, F. T., Kimball, M. S., & Shapiro, M. D. (1997). Preference parameters and behavioral heterogeneity: An experimental approach in the Health and Retirement Study. *The Quarterly Journal of Economics, 112*, 537-579.

Bassett, W. F., Fleming, M. J., & Rodrigues, A. P. (1998). How workers use 401 (k) plans: The participation, contribution, and withdrawal decisions. *National Tax Journal*, 263-289.

Bentler, P. M., & Kano, Y. (1990). On the equivalence of factors and components. *Multivariate Behavioral Research, 25*, 67-74.

Binder, A. (1984). Restrictions on statistics imposed by method of measurement: Some reality, some myth. *Journal of Criminal Justice, 12,* 467-481.

Birnbaum, A. (1968). Some latent train models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*, 395-479.

Brennan, R. L. (1998). Raw-score conditional standard errors of measurement in generalizability theory. *Applied Psychological Measurement, 22*, 307–331.

Bond, T. G., & Fox, C. M. (2007). *Fundamental measurement in the human sciences*. Mahwah, NJ: Erlbaum Associates.

Bonoma, T. V., & Schlenker, B. R. (1978). The SEU calculus: Effects of response mode, sex and sex role on uncertain decisions. *Decision Sciences, 9*, 206–227.

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

Bouchey, P. (2004). Questionnaire quest: New research shows that standard

questionnaires designed to reveal investors' risk tolerance levels are often flawed or misleading. *Journal of Financial Planning, 1*, 97-99.

Boyle, G. J. (1991). Does item homogeneity indicate internal consistency or item redundancy in psychometric scales? *Personality and Individual Differences*, 12, 291-294.

Brayman, S., Finke, M., Bessner, E., Bessner, B., Grable, J. E., & Griffin, P. (2015). *Current Practices for Risk Profiling in Canada & Review of Global Best Practices*. Ontario Securities Commission. Retrieved from http://www.osc.gov.on.ca/documents/en/Investors/iap_20151112_risk-profiling-report.pdf

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Campbell, J. (2006). Household Finance. *Journal of Finance, 61*, 1553-1605.

Carr, N. (2014). *Re-assessing the assessment: Exploring the factors that contribute to comprehensive financial risk evaluation* (Unpublished doctoral dissertation). Manhattan, Kansas: Kansas State University.

Certified Financial Planning Board of Standards, Inc. (CFP) (2015). *Student-centered learning objectives based upon CFP Board Principal Topics.* Retrieved from http://cfp.net/docs/for-education---resources-for-registered programs/cfpboard_learning_objectives_resource_document.pdf.

Certified Financial Planning Board of Standards, Inc. (CFP) (2016). *Practice Standards 300*. Retrieved from https://www.cfp.net/for-cfp-professionals/professional-standards-enforcement/standards-of-professional-conduct/financial-planning-

practice-standards/practice-standards-300.

Chen, P., & Finke, M. S. (1996). Negative net worth and the life cycle hypothesis. *Financial Counseling and Planning*, 7, 87-96.

Cohen, L. (1979). Approximate expressions for parameter estimates in the Rasch model. *The British Journal of Mathematical and Statistical Psychology, 32*, 113-120.

Iris, M., Ridings, J., & Conrad, K. J. (2010). The development of a conceptual frame-work for understanding elder self-neglect. *The Gerontologist, 50*, 303–315.

Constantinides, G. M., Donaldson, J. B., & Mehra, R. (2002). Junior can't borrow: A new perspective on the equity premium puzzle. *Quarterly Journal of Economics, 117*, 269-296.

Cordell, D. M. (2001). RiskPACK: How to evaluate risk tolerance. *Journal of Financial Planning, 14*, 36-40.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78,* 98-104.

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory*, Mason, OH: Cengage Learning.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

DeVellis, R. F. (2012). Scale development: Theory and applications. Thousand Oaks, CA: Sage Publications.

Embretson S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Maheah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Englehard, Jr., G. (1990). Gender differences in performance on mathematics items: Evidence for the United States and Thailand. *Contemporary Educational Psychology, 15*, 13-26.

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal and T. Haladyna (Eds.), *Large-Scale Assessment Programs for All Students: Development, Implementation, and Analysis* (pp. 261–287). Mahwah, NJ: Lawrence Erlbaum Associates.

Engelhard Jr, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Abingdon, UK: Routledge.

Engelhard Jr, G., & Wind, S. A. (2013). Rating quality studies using Rasch measurement theory. *College Board Research Report 2013-3*.

Erev, I., Ert, E., & Yechiam, E. (2008). Loss aversion, diminishing sensitivity, and the effect of experience on repeated decisions". *Journal of Behavioral Decision Making, 21*, 575–597.

Ert, E., & Erev, I. (2008). The rejection of attractive gambles, loss aversion, and the lemon avoidance heuristic. *Journal of Economic Psychology 29*, 715–723.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*, 272-299.

Fisher, R. A. (1958). *Statistical methods for research workers.* New York, NY: Hafner Publishing Co.

FinaMetrica (2015). *The FinaMetrica risk profiling system*. Retrieved from https://www.myrisktolerance.com/riskprofilingsystem.

Financial Industry Regulatory Authority (FINRA). (2011). *Regulatory Notice 11-25: Know your customer and suitability*. Retrieved from https://www.finra.org/sites/default/files/NoticeDocument/p123701.pdf.

Financial Industry Regulatory Authority (FINRA). (2012a). *Regulatory Notice 12-55: Guidance on FINRA's Suitability Rule*. Retrieved from https://www.finra.org/sites/default/files/NoticeDocument/p197435.pdf

Financial Industry Regulatory Authority (FINRA). (2012b). *Regulatory Notice 12-25: Additional Guidance on FINRA's New Suitability Rule*. Retrieved from https://www.finra.org/sites/default/files/NoticeDocument/p126431.pdf

Financial Industry Regulatory Authority (FINRA). (2014). *2111: Suitability*. Retrieved fromhttp://finra.complinet.com/en/display/display_main.html?rbid=2403&element_id=9859.

Financial Industry Regulatory Authority (FINRA) (2016). *Newsroom: Statistics.* Retrieved from http://www.finra.org/newsroom/statistics.

Ford, J. K., MacCallum, R. C., & Tait, M. (1986). The application of exploratory factor-analysis in applied psychology: A critical review and analysis. *Personnel Psychology, 39*, 291-314.

Gal, D. (2006). A psychological law of inertia and the illusion of loss aversion. *Judgment and Decision Making 1*, 23–32.

George, D., & Mallery, M. (2003). *Using SPSS for Windows step by step: a simple guide and reference*. Boston, MA: Allyn & Bacon.

Gilliam, J., Chatterjee, S., & Grable, J. (2010). Measuring the perception of financial risk tolerance: A tale of two measures. *Journal of Financial Counseling and Planning,*

21, 30-43.

Google. (2014). John E Grable. Retrieved from http://scholar.google.com/citations?user=6DdmMukAAAAJ&hl=en&oi=ao

Gorsuch, R. L. (1990). Common factor-analysis versus component analysis: Some well and little known facts. *Multivariate Behavioral Research, 25*, 33-39.

Grable, J. E. (2000). Financial risk tolerance and additional factors which affect risk taking in everyday money matters. *Journal of Business and Psychology, 14*, 625-630.

Grable, J. E., & Lytton, R. H. (1998). Investor risk tolerance: Testing the efficacy of demographics as differentiating and classifying factors. *Financial Counseling and Planning, 9*, 61- 73.

Grable, J. E., & Lytton, R. H. (1999). Financial risk tolerance revisited: The development of a risk assessment instrument. *Financial Services Review, 8*, 163-181.

Grable, J. E., & Lytton, R. H. (2001). Assessing the concurrent validity of the SCF risk tolerance question. *Financial Counseling and Planning, 12*, 43-53.

Grable, J. E., & Lytton, R. H. (2003). The development of a risk assessment instrument: A follow-up study. *Financial Services Review, 12*, 257-274.

Grable, J. E., & Joo, S-H. (2004). Environmental and biopsychosocial factors associated with financial risk tolerance. *Financial Counseling and Planning, 15*, 73-88.

Grable J. E. (2008). Risk tolerance, in Xiao, J. J., (ed.), *Handbook of Consumer Finance Research.* New York, NY: Springer.

Grable, J. E., Archuleta, K., & Nazarinia, R. (2010). *Financial planning and counseling scales*. New York: Springer.

Grable, J. E., & Schumm, W. (2010). An estimate of the reliability of the Survey of Consumer Finances risk-tolerance question. *Journal of Personal Finance, 9*, 117-131.

Guilford, J. P. (1965). *Fundamental statistics in psychology and education.* NewYork: McGraw-Hill, Inc.

Guiso, L., Haliassos, M., & Jappelli, T. (2002). *Houesehold Portfolios*. MIT Press, Cambridge.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Hambleton, R. K., & Swaminathan, H. (1984). *Item response theory: Principles and applications*. Hingham, MA: Kluwer, Nijhoff.

Hanna, S.D., Finke, M.S., & Waller, W. (2008). The concept of risk tolerance in personal financial planning. *Journal of Personal Finance, 7*, 96-108.

Hanna, S. D., Gutter, M. S., & Fan, J. X. (2001). A measure of risk tolerance based on economic theory. *Financial Counseling and Planning, 12*, 53-60.

Hanna, S. D., & Lindamood, S. (2004). An improved measure of risk aversion. *Journal of Financial Counseling and Planning*, *15*(2), 27-45.

Hanna, S. D., & Lindamood, S. (2005). Risk tolerance of married couples. *Paper presented at the 2005 Academy of Financial Services.* Retrieved October 26, 2007 from a CD of the Proceedings of the Academy of Financial Services, G2, 1-28.

Harvey, R. J., & Hammer, A. L. (1999). Item response theory. *The Counseling Psychologist*, *27*, 353-383.

Hattie, J., Jaeger, R. M., & Bond, L. (1999). Persistent methodological questions in educational testing, *Review of Educational Research, 24*, 393-446.

International Organization for Standardization. (2005). *ISO 22222:2005 Personal financial planning –Requirements for personal financial planners.*

Institute for Objective Measurement (IOM) (2000). *Definition of objective measurement*. Retrieved from http://www.rasch.org/define.htm

Jaccard, J., & Wan, C. K. (1996). *LISREL approaches to interaction effects in multiple regression*. Thousand Oaks, CA : Sage Publications.

Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis.* Upper Saddle River, NJ: Pearson Prentice Hall.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An analysis of decision under risk. *Econometrica 47*, 263-291.

Kimball, M., Sahm, C., & Shapiro, M. (2008). Imputing risk tolerance from survey responses. *Journal of the American Statistical Association 103*, 1028–1038.

Kline, P. (2000). *The handbook of psychological testing.* London, UK: Routledge.

Kuzniak, S., Rabbani, A., Heo, W., Ruiz-Menjivar, J., & Grable, J. E. (2015). The Grable & Lytton risk tolerance scale: A 15-year retrospective. *Financial Services Review, 24*, 177-192.

Linacre, J. M. (1997). *Guidelines for rating scales.* Retrieved from http://mesa.spc.uchicago.edu/rn2.htm

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement*, *3*, 103–122.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*, 85–106.

Linacre, J. M. (2004). Rasch model estimation: Further topics. *Journal of applied*

measurement, 5, 95-110.

Linacre, J. M. (2006). Rasch analysis of rank-ordered data. *Journal of Applied Measurement*, *7*, 129-139.

Liu, X. (2010) *Using and developing measurement instruments in science education: A Rasch modeling approach*. Charlotte, NC: Information Age Publishing.

Labovitz, S. (1967). Some observations on measurement and statistics. *Social Forces, 46*, 151-160.

Labovitz, S. (1970). The assignment of numbers to rank order categories. *American Sociological Review, 35*, 515-524.

Loehlin, J. C. (1990). Component analysis versus common factor-analysis: A case of disputed authorship. *Multivariate Behavioral Research, 25*, 29-31.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Welsley Publishing Company.

Lord, F. M., & Novick, M. R. (1980). *Applications of item response theory to practical testing problem.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

MacCallum, R. C., & Tucker, L. R. (1991). Representing sources of error in the common-factor model: Implications for theory and practice. *Psychological Bulletin, 109*, 502-511.

MacCrimmon, K. R., & Wehrung, D. A. (1986). *Risk Management.* New York: The Free Press.

Madrian, B. C., & Shea, D. F. (2000). *The power of suggestion: Inertia in 401 (k) participation and savings behavior* (No. w7682). National bureau of economic research.

Magnusson, D. (1967). *Test theory*. Reading, MA: Addison-Wesley.

Messick. S (1981). Evidence and ethics in the evaluation of tests. *ETS Research Report Series*, *1981*, i-41.

Messick, S. (1993). Validity, in Linn, R.L. *Educational Measurement*. Phoenix, AZ: American Council on Education and The Oryx Press.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*, 5-8.

Messick, S. (1996), *Validity and washback in language testing*. Princeton, NJ: Educational Testing Service.

Mittra, S. (1995). *Practicing financial planning: A complete guide for professionals*. Michigan: Mittra & Associates.

Mitchell, M., & Jolley, J. (2012). *Research design explained*. Boston, MA: Cengage Learning.

Mulaik, S. A. (1990). Blurring the distinctions between component analysis and common factor analysis. *Multivariate Behavioral Research, 25*, 53-59.

Mullainathan, S., & Thaler, R. H. (2000). *Behavioral economics* (No. w7948). National Bureau of Economic Research.

Nobre, L., & Grable, J.E. (2015). The role of risk profiles and risk tolerance in shaping client investment decisions. *Journal of Financial Service Professionals, 69*, 18-21.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3,* 1-18.

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

Peterson, R. A. (1994). A meta-analysis of Cronbach's coefficient alpha. *Journal of Consumer Research, 21*, 381-391.

Rasch, G. (1960 [1980]). *Probabilistic models for some intelligence and attainment tests.* [Reprint, with a Foreword and Afterword by Benjamin D. Wright. Chicago: University of Chicago Press, 1980.]

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 321-333). Berkeley: University of California Press.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational and Behavioral Statistics*, *4*, 207-230.

Roszkowski, M. J. (1992). *How to assess a client's financial risk tolerance: The basics. Personal Financial Risk Tolerance*. Bryn Mawr, PA: The American College.

Roszkowski, M. J., Snelbecker, G. E., & Leimberg, S. R. (1993) Risk-tolerance and risk aversion. In Leimberg, S. R., Satinsky, M. J. LeClair, R. T. & Doyle Jr, R. J. (eds). *The Tools and Techniques of Financial Planning* (4th ed., pp.213-225). Cincinnati, OH: National Underwriter.

Roszkowski, M. J., & Grable, J. (2005). Estimating risk tolerance: The degree of accuracy and the paramorphic representations of the estimate. *Financial Counseling and Planning, 16*, 29-47.

Roszkowski, M.J., Davey, G., & Grable, J.E. (2005). Questioning the questionnaire method: Insights on measuring risk tolerance from psychology and psychometrics. *Journal of Financial Planning, 18*, 68-76.

Ruiz-Menjivar, J., Blanco, A., Çopur, Z., Gutter, M. S., & Gillen, M. (2014). A cross-

cultural comparison of three risk tolerance measures: Turkey and The United States case. *International Journal of Research in Business and Social Science*, *3*, 1-14.

Saad, S., Carter, G. W., Rothenberg, M., & Israelson, E. (1999). *Testing and assessment: an employer's guide to good practices.* Washington, DC: U.S. Department of Labor Employment and Training Administration.

Schooley, D. K., & Worden, D. (1996). Risk aversion measures: Comparing attitudes and asset Allocation. *Faculty Publications School of Business. Paper 24.*

Smith E. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement,* 3, 205-231.

Snelbecker, G.E., Roszkowski, M J., & Cutler, N.E. (1990). Investors' risk tolerance and return aspirations: A conceptual model and exploratory data. *Journal of Behavioral Economics, 19,* 377-393.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15,* 72-101.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology, 18,* 161-169.

Spearman, C. (1913). Correlations of sums and differences. *British Journal of Psychology, 5,* 417-426.

Sung, J., & Hanna, S. (1996). Factors related to risk tolerance, *Financial Counseling and Planning, 7*, 11-20.

Suhr, D., (2003). Principal component analysis vs. exploratory factor analysis. *SUGI 30,*

Statistics and Data Analysis, 203–230.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 677-680.

Tennant, A., & Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care & Research*, *57*, 1358-1362.

Thorndike, R., & Hagen, E. (1961). *Measurement and evaluation in psychology and education*. New York: John Wiley and Sons.

Traub, R. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice*, 16, 8-14.

Torgerson, W. S. (1958). *Theory and Methods of Scaling.* New York: John Wiley.

Van de Venter, G., & Michayluk, D. (2007). Subjectivity in judgments: Further evidence from the financial planning industry. *The Journal of Wealth Management. 10*, 17-26.

Van Rooij, M., Lusardi, A., & Alessie, R. (2011). Financial literacy and stock market participation. *Journal of Financial Economics*, *101*, 449-472.

Wainer, H., & Thissen, D. (2001). *Test Scoring*. New York: Lawrence Erlbaum

Wright, B. D., & Panchapakesan, N. (1969) A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29*, 23-48.

Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis*. Chicago: MESA Press.

Wright B. D., & Stone M. H. (1979). *Best test design*. Chicago, IL: MESA Press.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement: Transactions of the Rasch Measurement*, 8(3), 370.

Yang, Y. (2004). Characteristics of risk preferences: Revelations from Grable and

Lytton's 13-item questionnaire. *Journal of Personal Finance, 3*(3), 20-40.

Yao, R., Hanna, S. D., & Lindamood, S. (2004). Changes in financial risk tolerance, 1983-2001. *Financial Services Review, 13*, 249-266.

Zumbo, B. D., & Zimmerman D. W. (1993). Is the selection of statistical methods governed by level of measurement? *Canadian Psychology, 34*, 390-399.

APPENDIX A

Grable and Lytton's (1999) Scale

Part I. Grable and Lytton's (1999) 13-Item Risk Tolerance Measure

1. In general, how would your best friend describe you as a risk taker?
a. A real gambler
b. Willing to take risks after completing adequate research
c. Cautious
d. A real risk avoider

2. You are on a TV game show and can choose one of the following. Which would you take?
a. $1,000 in cash
b. A 50% chance at winning $5,000
c. A 25% chance at winning $10,000
d. A 5% chance at winning $100,000

3. You have just finished saving for a "once-in-a-lifetime" vacation. Three weeks before you plan to leave, you lose your job. You would:
a. Cancel the vacation
b. Take a much more modest vacation
c. Go as scheduled, reasoning that you need the time to prepare for a job search
d. Extend your vacation, because this might be your last chance to go first-class

4. If you unexpectedly received $20,000 to *invest*, what would you do?
a. Deposit it in a bank account, money market account, or an insured CD
b. Invest it in safe high quality bonds or bond mutual funds
c. Invest it in stocks or stock mutual funds

5. In terms of experience, how comfortable are you investing in stocks or stock mutual funds?
a. Not at all comfortable
b. Somewhat comfortable
c. Very comfortable

6. When you think of the word "risk" which of the following words comes to mind first?
a. Loss
b. Uncertainty
c. Opportunity
d. Thrill

7. Some experts are predicting prices of assets such as gold, jewels, collectibles, and real estate (hard assets) to increase in value; bond prices may fall, however, experts tend to agree that government bonds are relatively safe. Most of your investment assets are now in high interest government bonds. What would you do?
a. Hold the bonds
b. Sell the bonds, put half the proceeds into money market accounts, and the other half into hard assets
c. Sell the bonds and put the total proceeds into hard assets
d. Sell the bonds, put all the money into hard assets, and borrow additional money to buy more

8. Given the best and worst case returns of the four investment choices below, which would you prefer?
a. $200 gain best case; $0 gain/loss worst case
b. $800 gain best case; $200 loss worst case
c. $2,600 gain best case; $800 loss worst case
d. $4,800 gain best case; $2,400 loss worst case

9. In addition to whatever you own, you have been given $1,000. You are now asked to choose between:
a. A sure gain of $500
b. A 50% chance to gain $1,000 and a 50% chance to gain nothing

10. In addition to whatever you own, you have been given $2,000. You are now asked to choose between:
a. A sure loss of $500
b. A 50% chance to lose $1,000 and a 50% chance to lose nothing

11. Suppose a relative left you an inheritance of $100,000, stipulating in the will that you invest ALL the money in ONE of the following choices. Which one would you select?
a. A savings account or money market mutual fund
b. A mutual fund that owns stocks and bonds
c. A portfolio of 15 common stocks
d. Commodities like gold, silver, and oil

12. If you had to invest $20,000, which of the following investment choices would you find most appealing?
a. 60% in low-risk investments 30% in medium-risk investments 10% in high-risk investments
b. 30% in low-risk investments 40% in medium-risk investments 30% in high-risk investments
c. 10% in low-risk investments 40% in medium-risk investments 50% in high-risk investments

13. Your trusted friend and neighbor, an experienced geologist, is putting together a

group of investors to fund an exploratory gold mining venture. The venture could pay back 50 to 100 times the investment if successful. If the mine is a bust, the entire investment is worthless. Your friend estimates the chance of success is only 20%. If you had the money, how much would you invest?
a. Nothing
b. One month's salary
c. Three month's salary
d. Six month's salary

Part II. Scoring System (presented for each question)

1. a=4; b=3; c=2; d=1

2. a=1; b=2; c=3; d=4

3. a=1; b=2; c=3; d=4

4. a=1; b=2; c=3

5. a=1; b=2; c=3

6. a=1; b=2; c=3; d=4

7. a=1; b=2; c=3; d=4

8. a=1; b=2; c=3; d=4

9. a=1; b=3

10. a=1; b=3

11. a=1; b=2; c=3; d=4

12. a=1; b=2; c=3

Part III. Cut-off values (in raw scores)

a) 18 or below: low financial risk tolerance (i.e., conservative investor),

b) 19-22: below-average financial risk tolerance,

c) 23-28: average/moderate financial risk tolerance,

d) 29-32: above-average financial risk tolerance,

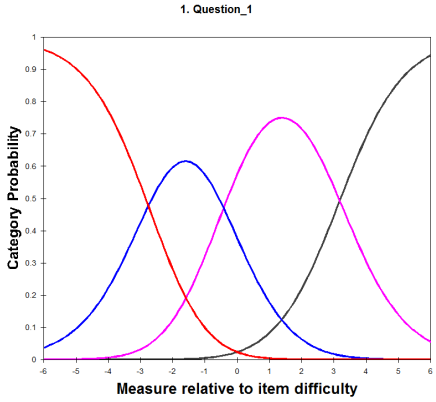
e) 33 and above: high financial risk tolerance (i.e., aggressive investor).

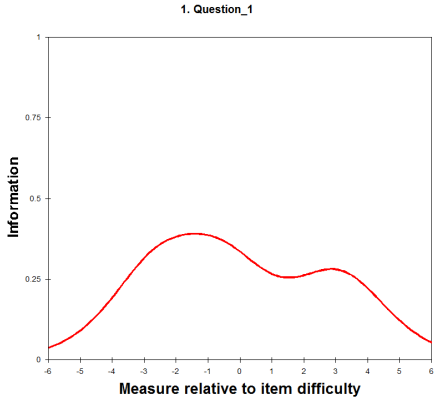
Appendix B

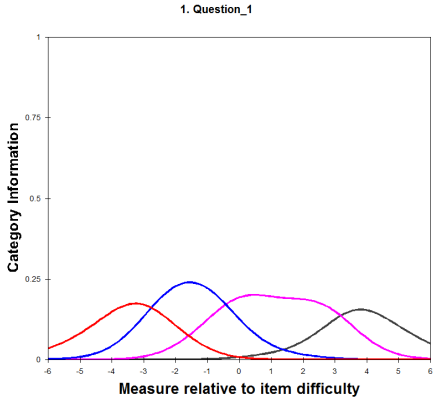
Visual Aspects for Model II

Part I: item information, and category information curves for items in Model I.

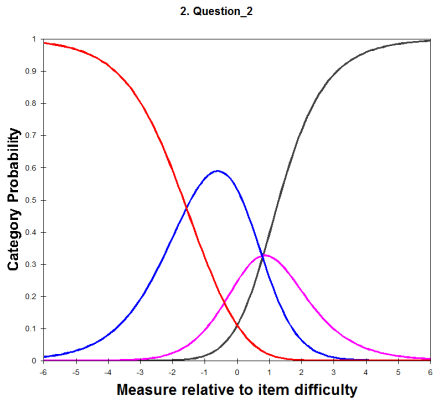
Question 1

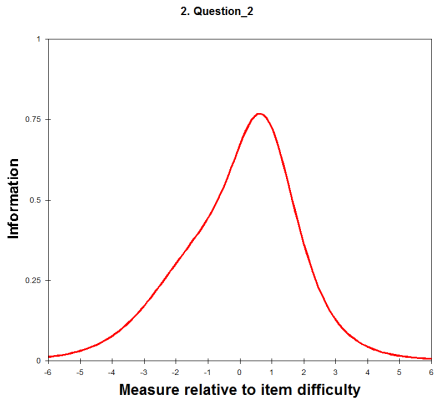


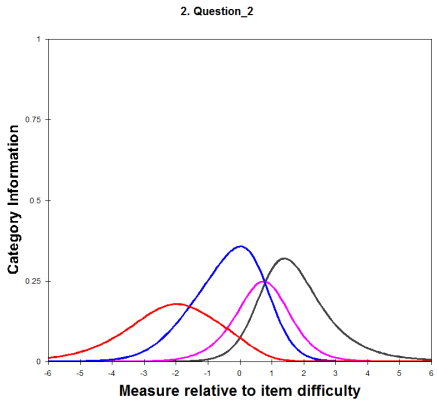




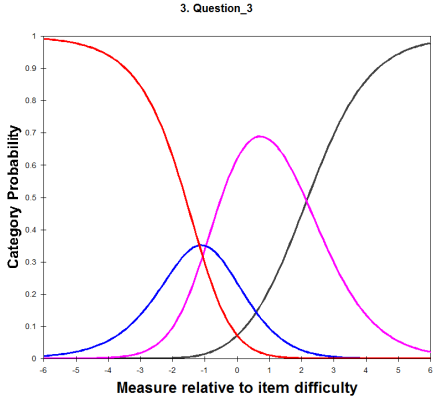
Question 2

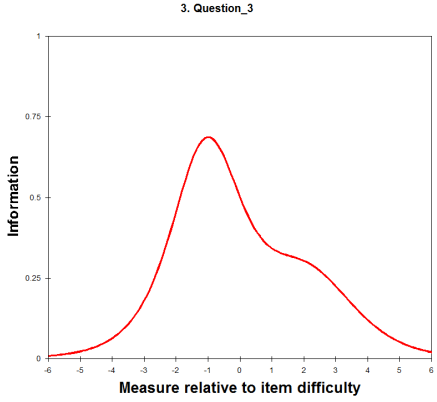


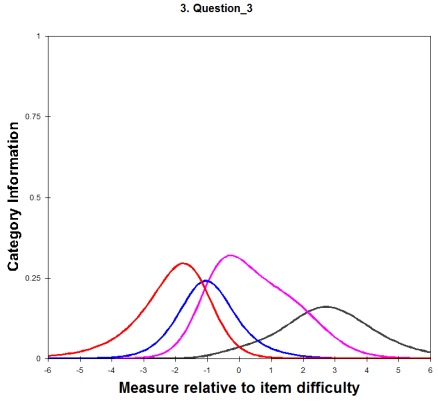




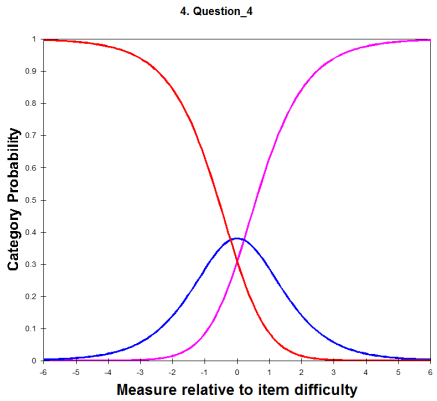
Question 3

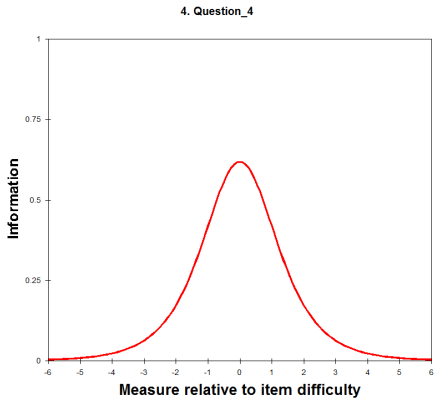


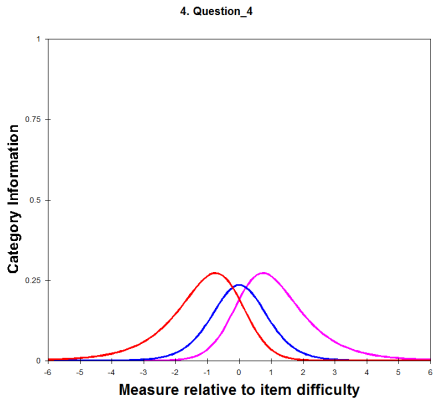




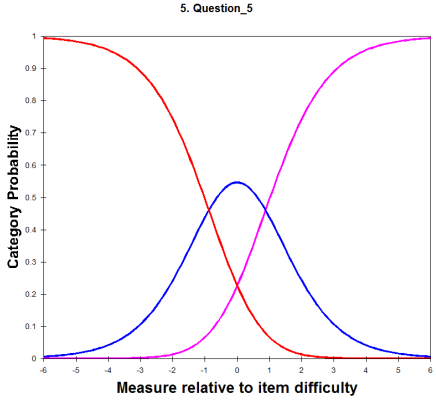
Question 4

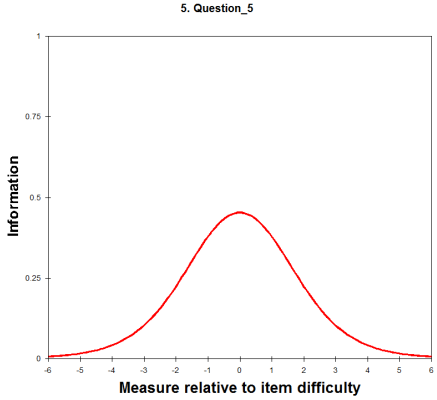


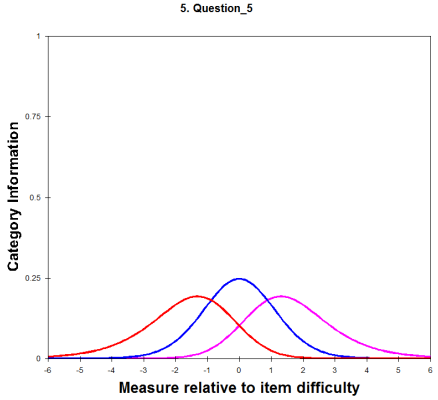




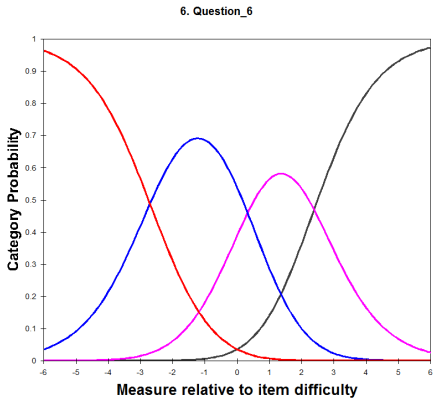
Question 5

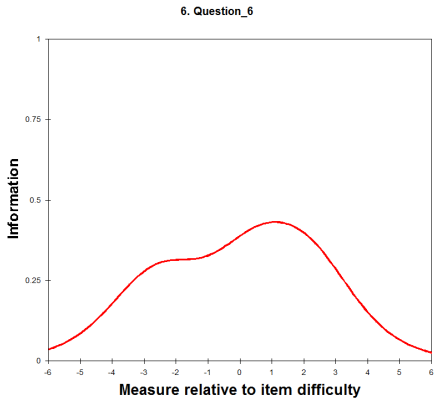


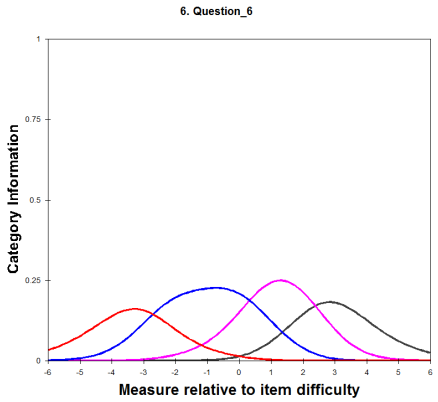




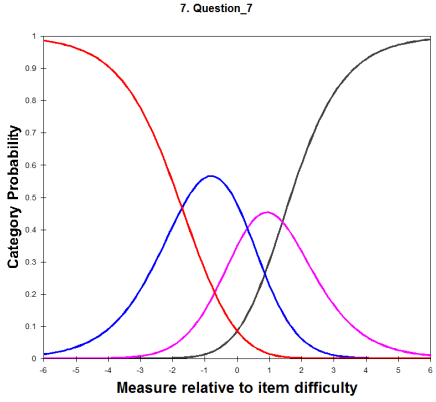
Question 6

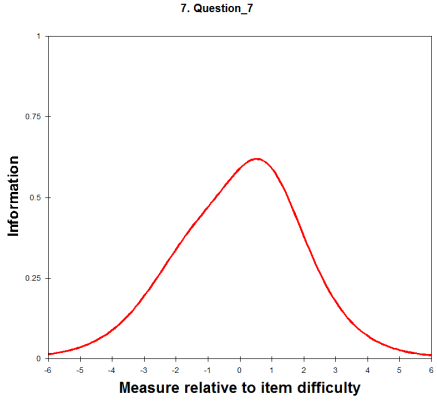


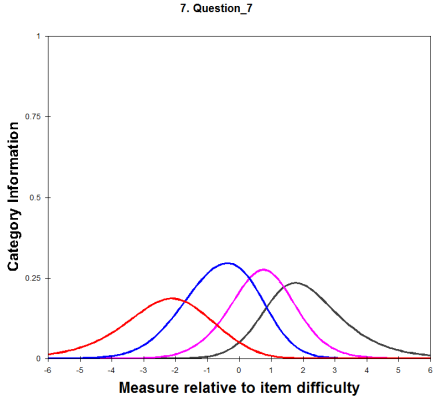




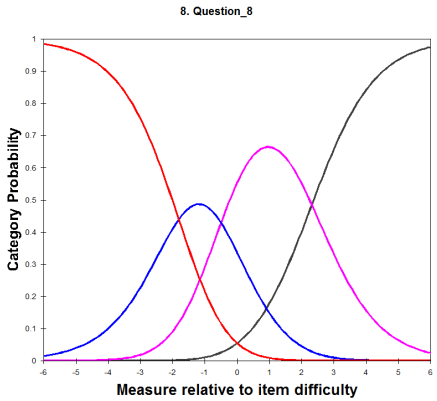
Question 7

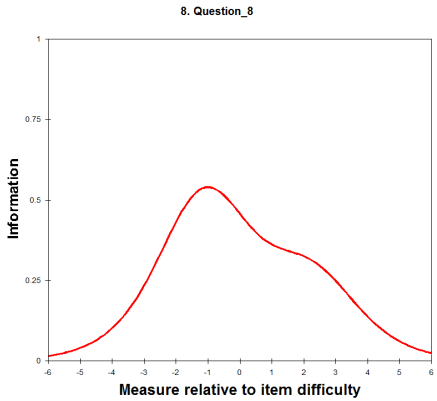


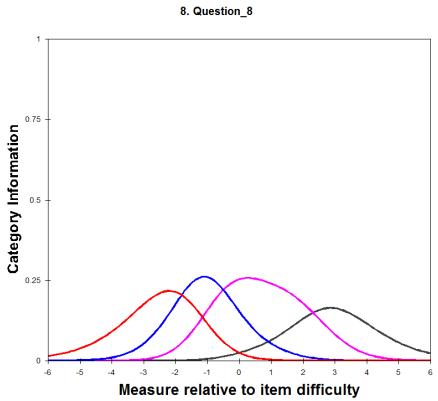




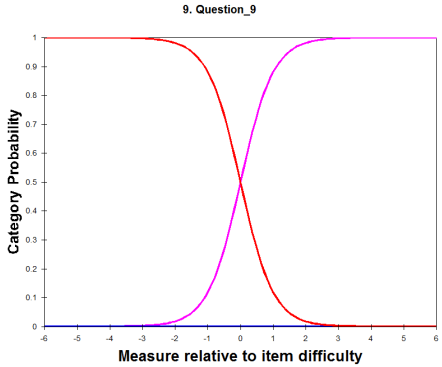
Question 8



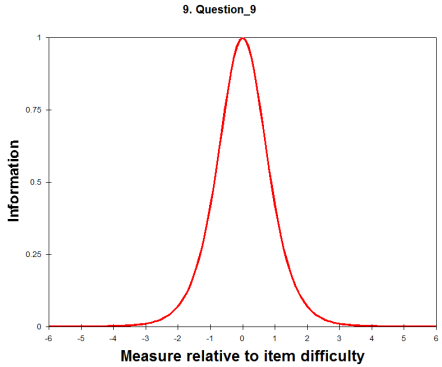




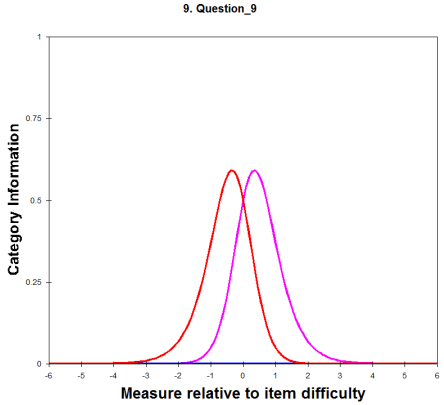
Question 9



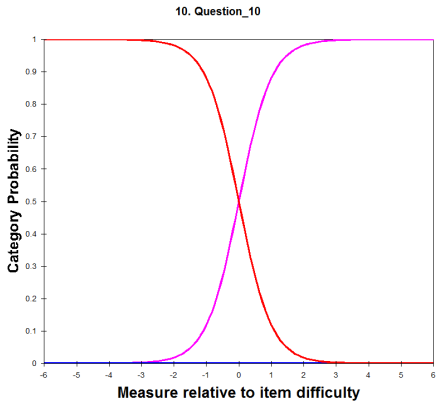
Category probability: 1

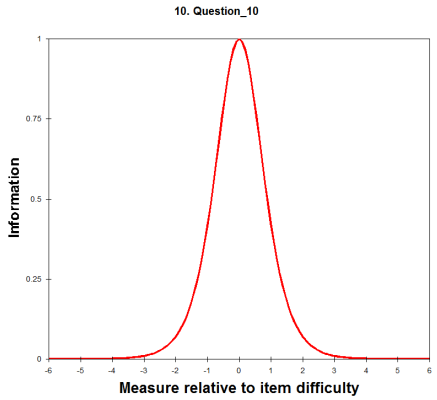


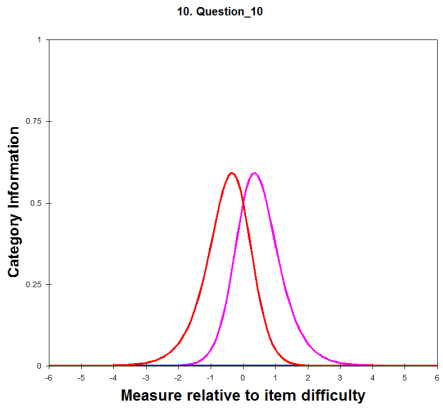
Item information function



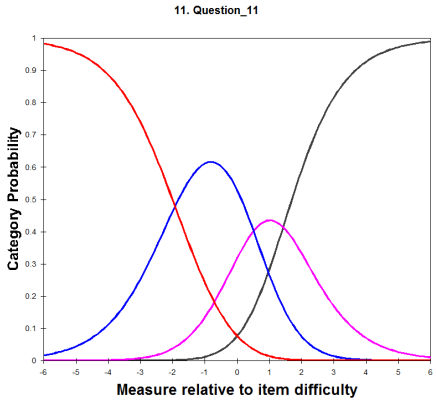
Question 10

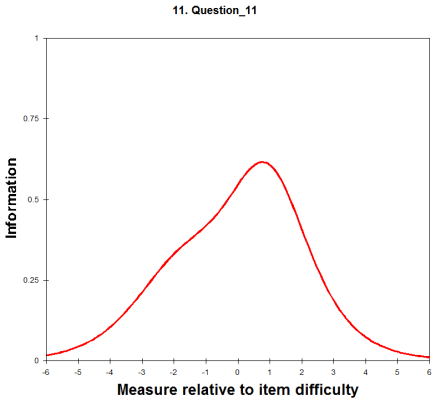


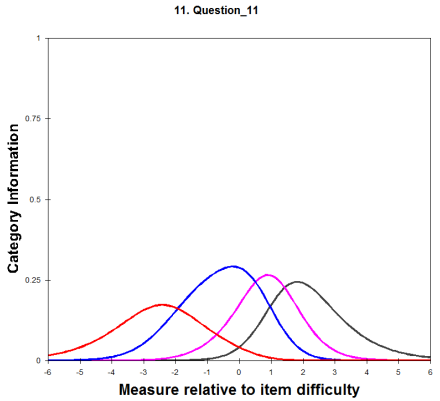




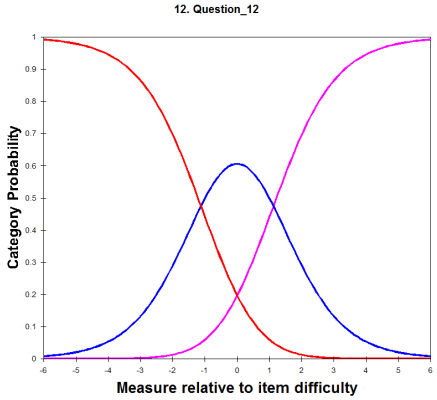
Question 11

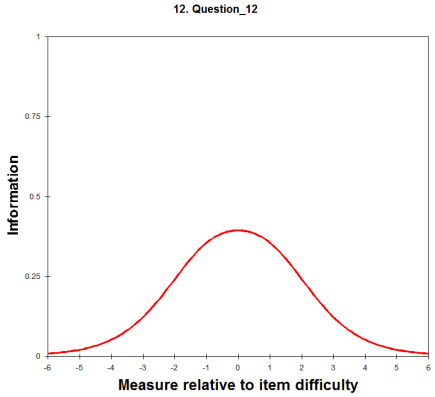


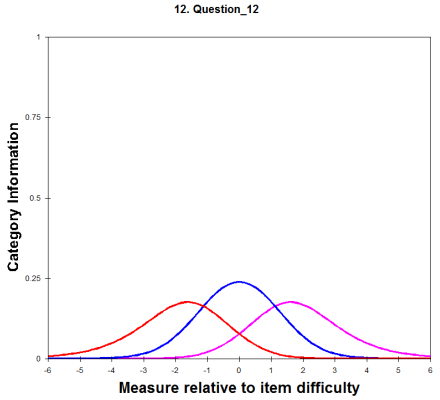




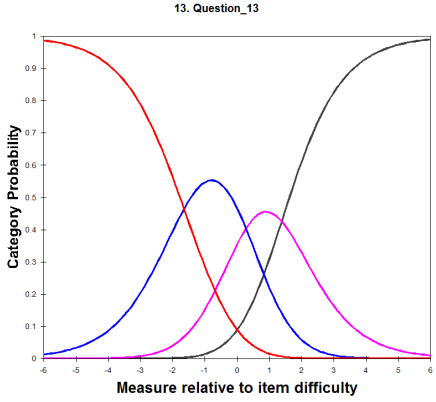
Question 12

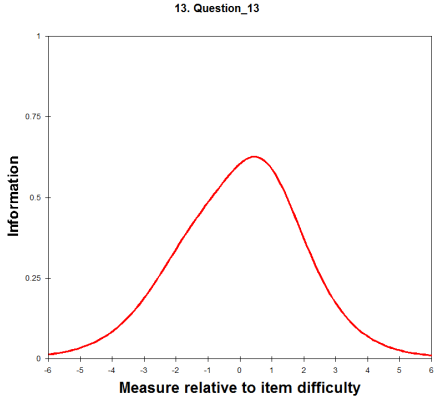


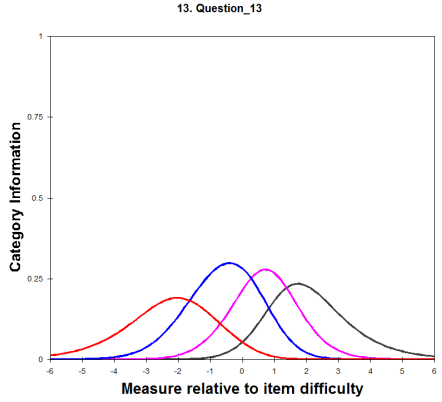




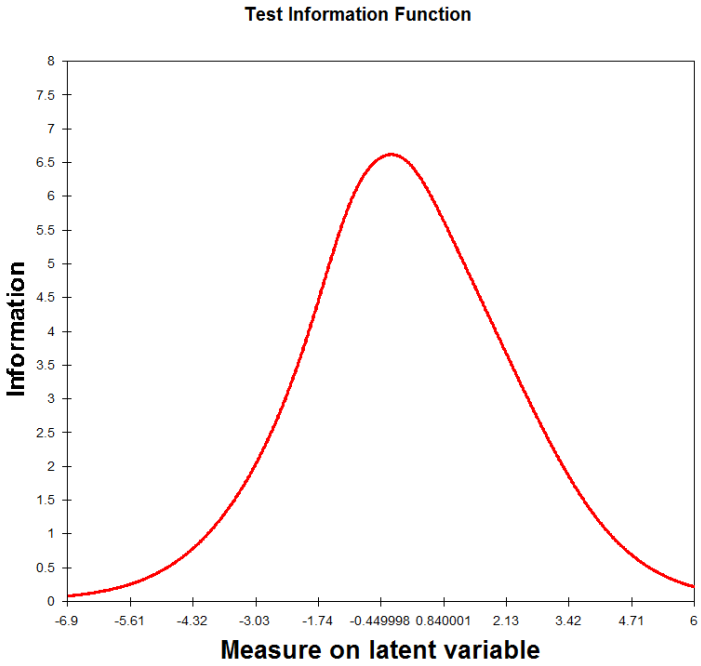
Question 13







Part II: Test Information Function for Model I.



Appendix C

Revised Version of the GL-FRT Scale

Part I. Revised version of the GL-FRT scale (10 items)

1. In general, how would your best friend describe you as a risk taker?
a. A real gambler
b. Willing to take risks after completing adequate research
c. Cautious
d. A real risk avoider

2. You are on a TV game show and can choose one of the following. Which would you take?
a. $1,000 in cash
b. A 50% chance at winning $5,000
c. A 25% chance at winning $10,000
d. A 5% chance at winning $100,000

4. If you unexpectedly received $20,000 to *invest*, what would you do?
a. Deposit it in a bank account, money market account, or an insured CD
b. Invest it in safe high quality bonds or bond mutual funds
c. Invest it in stocks or stock mutual funds

5. In terms of experience, how comfortable are you investing in stocks or stock mutual funds?
a. Not at all comfortable
b. Somewhat comfortable
c. Very comfortable

6. When you think of the word "risk" which of the following words comes to mind first?
a. Loss
b. Uncertainty
c. Opportunity
d. Thrill

7. Some experts are predicting prices of assets such as gold, jewels, collectibles, and real estate (hard assets) to increase in value; bond prices may fall, however, experts tend to agree that government bonds are relatively safe. Most of your investment assets are now in high interest government bonds. What would you do?
a. Hold the bonds
b. Sell the bonds, put half the proceeds into money market accounts, and the other half

into hard assets
c. Sell the bonds and put the total proceeds into hard assets
d. Sell the bonds, put all the money into hard assets, and borrow additional money to buy more

8. Given the best and worst case returns of the four investment choices below, which would you prefer?
a. $200 gain best case; $0 gain/loss worst case
b. $800 gain best case; $200 loss worst case
c. $2,600 gain best case; $800 loss worst case
d. $4,800 gain best case; $2,400 loss worst case

11. Suppose a relative left you an inheritance of $100,000, stipulating in the will that you invest ALL the money in ONE of the following choices. Which one would you select?
a. A savings account or money market mutual fund
b. A mutual fund that owns stocks and bonds
c. A portfolio of 15 common stocks
d. Commodities like gold, silver, and oil

12. If you had to invest $20,000, which of the following investment choices would you find most appealing?
a. 60% in low-risk investments 30% in medium-risk investments 10% in high-risk investments
b. 30% in low-risk investments 40% in medium-risk investments 30% in high-risk investments
c. 10% in low-risk investments 40% in medium-risk investments 50% in high-risk investments

13. Your trusted friend and neighbor, an experienced geologist, is putting together a group of investors to fund an exploratory gold mining venture. The venture could pay back 50 to 100 times the investment if successful. If the mine is a bust, the entire investment is worthless. Your friend estimates the chance of success is only 20%. If you had the money, how much would you invest?
a. Nothing
b. One month's salary
c. Three month's salary
d. Six month's salary

Part II. Scoring System (presented for each question)

1. a=4; b=3; c=2; d=1

2. a=1; b=2; c=3; d=4

4. a=1; b=2; c=3

5. a=1; b=2; c=3

6. a=1; b=2; c=3; d=4

7. a=1; b=2; c=3; d=4

8. a=1; b=2; c=3; d=4

11. a=1; b=2; c=3; d=4

12. a=1; b=2; c=3

Part III. Cut-off values (in raw scores)

a) 14 or below: low financial risk tolerance (i.e., conservative investor),

b) 15-17: below-average financial risk tolerance,

c) 18-22: average/moderate financial risk tolerance,

d) 23-25: above-average financial risk tolerance,

e) 26 and above: high financial risk tolerance (i.e., aggressive investor).

Appendix D

Outfit MNSQ Calculation

Part I. Outfit MNSQ statistic Formula (Retrieved from Engelhard, 2013).

$$U_i = \frac{\sum_n^N Z_{ni}^2}{N}$$

Where,

U_i is the outfit MNSQ for a person *i,*

Z_{in}^2 is the squared standardized residual for person *i* in question *n*,

N is the number of questions included in the instrument.

Part II. Additional statistics and respective formulas.

The standardized residuals (used to compute the outfit MNSQ) are derived using the following statistics and formulas.

a) Observed responses: $X_{ni} = 0, 1$

b) Expected responses: $P_{ni} = \frac{\exp(\theta_n - \delta_i)}{1+\exp(\theta_n - \delta_{i)}}$

c) Response variance: $Q_{ni} = P_{ni} * (1 - P_{ni})$

d) Score residuals: $Y_{ni} = X_{ni} - P_{ni}$

e) Standardized residuals: $Z_{ni} = \frac{Y_{ni}}{Q_{ni}^{1/2}}$

Note. Item parameters and ability parameters are presented in Table 7 and Table 8, respectively. Also, outfit statistics estimated with these formulas may vary with the outfit statistics provided in Rasch software packages (e.g., Winsteps).

Appendix E

Expected Responses by Ability Level

Table 10.

Expected Response by ability level

Raw Score	Measure	Item 1	Item 2	Item 4	Item 5	Item 6	Item 7	Item 8	Item 11	Item 12	Item 13
10	-5.75	1.11	1.02	1.01	1.01	1.04	1.01	1.04	1.03	1.01	1.01
11	-4.42	1.34	1.06	1.05	1.05	1.14	1.02	1.16	1.1	1.04	1.03
12	-3.56	1.58	1.13	1.12	1.11	1.28	1.05	1.32	1.2	1.09	1.07
13	-3.00	1.77	1.21	1.22	1.19	1.43	1.09	1.51	1.33	1.15	1.11
14	-2.57	1.94	1.31	1.34	1.28	1.56	1.14	1.7	1.45	1.23	1.18
15	-2.20	2.06	1.4	1.47	1.38	1.66	1.19	1.87	1.56	1.3	1.24
16	-1.88	2.14	1.47	1.57	1.45	1.73	1.24	1.98	1.64	1.36	1.29
17	-1.58	2.27	1.59	1.76	1.57	1.84	1.31	2.16	1.76	1.46	1.38
18	-1.31	2.35	1.67	1.89	1.66	1.9	1.37	2.27	1.85	1.53	1.45

19	-1.04	2.43	1.76	2.03	1.75	1.97	1.44	2.38	1.94	1.61	1.53
20	-0.78	2.55	1.9	2.24	1.9	2.07	1.56	2.54	2.07	1.73	1.66
21	-0.52	2.63	2.01	2.36	2	2.14	1.65	2.64	2.17	1.82	1.76
22	-0.27	2.7	2.12	2.48	2.1	2.22	1.74	2.74	2.28	1.9	1.86
23	-0.02	2.76	2.25	2.57	2.19	2.29	1.83	2.82	2.39	1.99	1.96
24	0.23	2.83	2.39	2.66	2.29	2.37	1.93	2.91	2.51	2.08	2.07
25	0.48	2.89	2.54	2.73	2.38	2.46	2.04	2.98	2.64	2.16	2.19
26	0.72	2.94	2.71	2.78	2.47	2.55	2.15	3.06	2.77	2.25	2.31
27	0.97	3	2.88	2.83	2.55	2.64	2.27	3.14	2.91	2.33	2.44
28	1.21	3.05	3.05	2.86	2.62	2.73	2.39	3.21	3.05	2.41	2.57
29	1.46	3.11	3.21	2.89	2.68	2.83	2.52	3.29	3.18	2.49	2.71
30	1.73	3.16	3.36	2.92	2.74	2.92	2.66	3.36	3.3	2.56	2.85
31	2.01	3.25	3.54	2.94	2.81	3.07	2.86	3.47	3.47	2.66	3.05
32	2.31	3.32	3.64	2.95	2.84	3.16	3	3.54	3.56	2.71	3.18
33	2.65	3.41	3.75	2.97	2.89	3.3	3.2	3.64	3.68	2.79	3.36

34	3.06	3.51	3.83	2.98	2.92	3.43	3.37	3.72	3.77	2.84	3.51
35	3.58	3.66	3.91	2.99	2.95	3.61	3.61	3.83	3.87	2.91	3.7
36	4.39	3.82	3.96	2.99	2.98	3.8	3.81	3.92	3.94	2.96	3.86
37	5.67	3.94	3.99	3	2.99	3.93	3.94	3.97	3.98	2.99	3.96

Note. Measures above represent the ability parameters that correspond to each of the possible total raw scores for the revised version of the GL-FRT scale.