UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

STATISTICAL NUMERACY NORMS AND DECISION VULNERABILITY BENCHMARKS: A NORM-REFERENCED METHOD FOR ESTIMATING THE RISK LITERACY DIFFICULTY LEVEL OF CHOICES AND RISK COMMUNICATIONS

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

JINAN N. ALLAN
Norman, Oklahoma
2021

STATISTICAL NUMERACY NORMS AND DECISION VULNERABILITY BENCHMARKS: A NORM-REFERENCED METHOD FOR ESTIMATING THE RISK LITERACY DIFFICULTY LEVEL OF CHOICES AND RISK COMMUNICATIONS

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF PSYCHOLOGY

BY THE COMMITTEE CONSISTING OF

Dr. Edward T. Cokely, Chair

Dr. Rocio Garcia-Retamero, Co-Chair

Dr. Hank Jenkins-Smith

Dr. Adam Feltz

Dr. Scott Gronlund

Dr. Joseph T. Ripberger

Dr. Robert Terry

Acknowledgements

This dissertation would not have been possible without the guidance of my professors, friends, and colleagues at the University of Oklahoma. First, I would like to express my deepest gratitude and appreciation for my advisor Dr. Edward T. Cokely. His unwavering support and encouragement are unparalleled. In the last seven years, Ed has taught me so much about research, and always found new ways to challenge me, while still providing a safe place to explore. Thank you for always reminding me that you don’t care if I fail, only if I stop trying.

I would also like to thank my committee members for their constant dedication and inspiration. First, Drs. Joseph Ripberger and Hank Jenkins-Smith taught me how find and solve interdisciplinary problems and provided me with so many opportunities to grow while at OU. The CRCM/NIRR lunch table was the starting point of so many insightful conversations, and I am so grateful to know that I will always have a home in the office and at the lunch table. Drs. Rocio Garcia-Retamero and Adam Feltz are pillars of the Risk Literacy Program and the work they have done has set the foundation for so much of what I do. I am endlessly indebted to them, and only hope that one day I can pay it forward. Last, but certainly not least, Drs. Scott Gronlund and Robert Terry have been with me since my first days at OU. They have taught me so much about what it means to be a research psychologist as well as a teacher and an academic. I am grateful for their time, effort, and kind, encouraging feedback. I am forever grateful that so much of my PhD experience was positive and enjoyable. I owe so much of that to these seven thoughtful scholars who served on my committee. Thank you.

I have been fortunate throughout my life to have many strong, empowering, and encouraging women to look up too. In high school, Jennifer Gorsline, Kim Shawver, Diana

Hoehn, and Deena Alam-Parks were a force to be reckoned with. I owe so much of my educational success to those first lessons learned at DCMST. At Michigan Tech, I was always inspired by Kim Roberts, Sylvia Matthews, and Bonnie Gorman, for their unwavering dedication to their students. At OU, the guidance and friendship I found in Kuhika Gupta, Carol Silva, and Silke Feltz carried me through some of the hardest days. These women changed my life in ways that I am still discovering today. Thank you all for always having my back, and giving me a safe place to land.

They say that the right friends come at just the right time. The friends I made while at OU definitely came at just the right time. Kylie, Vinny, Wesley, Felicia, Megan, Madhuri, Jinhyo, Jon, Kenzie, Sean, Cory, Marc, Tom, Craig, Saima, Ashley, Grace, Jenna, Olivia, and Gwen: Thank you for the endless fun. From conferences, concerts, and half-marathons, to working for hours at coffeeshops, and late-night Taco Bell, in your own way, each of you helped me see the bigger picture, and reminded me why I came here in the first place. Norman became my home because of you. Special thanks go to Kylie, Vinny, and Wesley, who were my "first friends" at OU and who taught me what it means to be a grad student. Felicia, Megan, and Madhuri were the best roommates I could ask for. Thank you for spending entirely too much time watching trash TV, eating Thai food, and traveling around the world. Finally, to my undergraduate mentees: Jessica, Kaytlyn, Madeline, and Fizza, working with you has been such a joy. Thank you for being a constant source of inspiration. To the friends I made in the Psychology department, at the CRCM/NIRR, and in the Decision Analytics Lab: Thank you for the many important conversations about research, school, and life.

Elizabeth Laine, Jenna Brackett, Brittany Nelson, Audrey Hutton, Will Lehman, Kyle Yarusso, and Sophie Rizzo: It feels like you have been there since the beginning. You have seen me through so many stages and seasons of my life, and I cannot imagine having made it this far without you. Thank you for encouraging me to take walks, drink water, and for giving me the time and space to be my truest self. You are my people.

This dissertation would not have been possible without the support of my family. I am lucky to have grown up in a home that valued education and lifelong learning. I am endlessly grateful that my passion for school was always supported, and that from a young age I was surrounded by and encouraged to participate in intellectually stimulating activities. I thank God every day for my life and for the countless opportunities I have been afforded. Importantly, my grandparents, from whom I received my name, and so much more, are the reason I am here today. Their sacrifices made it possible for me to dream this dream. Before I knew how far I could soar, they helped to build an environment where I could spread my wings. Finally, this dissertation is for the man who has been by my side since the very beginning. “My father didn’t tell me how to live; he lived and let me watch him do it.” My dad sacrificed his life to watch me grow and succeed. He has been the most important constant in my life, and I am still learning so much from him. Thank you for the life lessons, travel adventures, home cooked meals, and most importantly for teaching me to “Speak the truth and fear no one.”

Table of Contents

List of Tables

List of Figures

Abstract

People who misunderstand risk are more likely to experience costly decision biases. In addition to serious personal and economic implications, risk misunderstanding can further undermine high-stakes risk mitigation efforts. What makes someone vulnerable to misunderstanding risk? Recent research suggests individual differences in statistical numeracy (i.e., ones' practical probabilistic reasoning) tend to be the strongest predictor of general decision making skills and risk literacy—i.e., the ability to evaluate and understand risk (Cokely et al., 2012, 2018). The current study aims to develop some relatively straightforward and practical tools that provide insights and make meaningful inferences about who, when, how much, and why people may misunderstand risks (i.e., Decision Vulnerability Analyses). First, the development of a unidimensional measure of Risk Literacy lends itself to the development of numeracy norms and comparative risk literacy levels. Cumulative percentile rank norms are provided for the general public, as well as stratified by education, age, gender, and race. Next, a major potential threat is addressed: differential item functioning and measurement invariance. Results suggest that the Berlin Numeracy Test-Schwartz and the Risk Literacy Test pass strict measurement invariance, with good model fit (RMSEA < .06). Finally, a template for decision vulnerability analysis is developed and validated using five example artifacts (e.g., risk communications). Initial results suggest that over 90% of predicted risk literacy difficulty levels are within ten percentile points of the observed value. An additional out-of-sample application with hurricane risk communications is explored, and discussion focuses on theoretical implications, future research for methodological improvements, and further implications for high-stakes decision making.

Keywords: Numeracy, risk literacy, decision vulnerability, norms, invariance

Chapter 1

Introduction: Difficult Decisions & Decision Vulnerability

"*If you cannot measure it, you cannot improve it.*" – Lord Kelvin

Linda and Douglas deSilvey, their daughter Donna, and Linda's parents Nadine and Ted Gifford were all aware of the warnings. Their family had sheltered-in-place through severe weather before at the Gifford's home, which had been designed to meet the updated NIST standards developed after Hurricane Camille (1969). Together, they decided not to evacuate. It was not an easy decision, but they gathered at the family home in Gulf Hills and prepared to ride out the storm. On Monday, August 29th, 2005, shortly after breakfast the family realized that Hurricane Katrina was different. In a matter of hours, storm surge was flooding the streets, and last-minute evacuation was probably no longer an option. By the end of the day, their roof collapsed under the growing force of Hurricane Katrina, leaving Douglas deSilvey the sole survivor. One day earlier, about 100 miles west of Gulf Hills, Daniel Aldrich decided to take his family and evacuate New Orleans. He, his wife, and their two young children had recently moved, and Daniel was scheduled to start his new job on Monday. However, thanks in part to a discussion with his neighbor, the family left for Houston that Sunday. The next day, August 29th, 2005, Daniel and his family learned their home had been destroyed and all their belongings were "nothing more than a big, black smear on the ground."

Hurricane Katrina was one of the worst natural disasters in the History of U.S., leading to at least 1,833 fatalities, destroying 352,930 homes, and displacing roughly half a million people. Over the last 15 years, many investigations have considered and documented the various factors that contributed to the evacuation decisions people made during Hurricane Katrina (Boin et al., 2019). These reports overwhelmingly suggested that there is no single factor that explains people's decisions. While some factors stand out as more influential, for better or worse, there appear to be complex and often justifiable reasons for the difficult choices people made. For example, many people who decided to shelter-in-place were older adults who had retired, like the Giffords. Many retirees also had medical conditions that would make travel difficult even under normal circumstances. Other people reported staying primarily because of personal, moral, or professional responsibilities, including people in medical and emergency response positions. And there were those who otherwise preferred to leave, but who stayed to be with family and friends who would not or could not evacuate (e.g., perhaps as the deSilveys did). Far too many people also decided to stay because they felt it was truly their only option, since they lacked crucial resources to do otherwise (e.g., no savings, no car, no social support).

Taken together, considering the many complications and unique circumstances people were faced with, it is possible that most choices genuinely reflected people's best effort to make "the right" decision in an extremely difficult and risky situation. Despite the complexity and rapidly evolving nature of the risks, it is noteworthy that the weight of the evidence suggests that the vast majority of people had access to remarkably accurate forecasts, warnings, and risk communications well-before the storm made landfall. While some of the risk communications may have caused confusion for some people, the forecasts of the National Weather Service and the National Hurricane Center were generally quite timely, precise, and accurate. To illustrate, consider these examples reported in subsequent media coverage: "Storm-track projections released to the public more than two days (56 hours) before Katrina came ashore were off by only about 15 miles... [and] Two days before the storm hit, the hurricane center predicted Katrina's strength at landfall; the agency was off the mark by only about 10 mph." (NBCUniversal News Group, 2005). Given the speed and accuracy of these forecasts, there should be little doubt that these risk communications protected the lives of many people, including Daniel Aldrich, his family, and his neighbors. Nevertheless, even the most accurate and useful information does not necessarily make risks easy to interpret, nor does this kind of information guarantee that decisions will be easier or less biased.

Decision Vulnerability. It seems reasonable to expect higher-quality decisions are likely to result when people have both access to, and awareness of, more accurate risk communications, which was generally the case with Hurricane Katrina. Nevertheless, research shows that there are many examples of accurate yet difficult, technical risk communications, that cause risk misunderstanding and biased decisions among members of the public and highly trained professionals alike (Garcia-Retamero & Cokely, 2017; Ghazal et al., 2014; Petrova et al., 2018). Research suggests two major sources of decision vulnerability often result from (a) differences in cognitive skills (e.g., risk literacy skills) and (b) the difficulty vs. transparency of data depicted in risk communications (e.g., displaying part-to-whole data relations). Here, I will roughly define the notion of decision vulnerability as the relative probability of biases or errors during judgment and decision making, by drawing comparisons to a specific reference class (e.g., norm referenced test score distributions). This is roughly consistent with notions of social vulnerability as indexed for use in the assessment of behavioral risks, developed by the Centers for Disease Control (CDC; Tate, 2013).

Theoretically, a high level of decision vulnerability does not imply that a bad outcome will definitely result, just as a low degree of vulnerability does not necessarily mean that a decision maker will avoid bias. Instead, as it is used here, it is useful to assume that decision vulnerability generally functions the same way a good decision making *process* does (e.g., a good decision is a good bet such that it pays off *on average*, though sometimes a good decision does not work out). Accordingly, decision vulnerability may be associated with (i) differences in the relative difficulty of a risk communication (e.g., *item-level* difficulty or *what* is hard) as well as (ii) skill-related individual differences (e.g., *person-level* difficulty or *for whom* it is hard). To the extent these decision vulnerability factors are at play, an interesting theoretical question is *why*?

Cognitive Abilities & Decision Biases. It is well-established that decision biases are robustly predicted by individual differences in general cognitive abilities, including intelligence and statistical numeracy tests (Allan, 2018; Cokely et al., 2012, 2018; Del Missier et al., 2012; Peters, 2012; Stanovich & West, 2000). In some sense this is not surprising: people who are more intellectually prepared to engage in sophisticated reasoning should also be better prepared to avoid misunderstandings and other cognitive errors and biases. As such, people with higher levels of general cognitive abilities are likely to be less vulnerable to downstream decision biases and costly decision outcomes—a finding that has been demonstrated many times over the past 20 years in decision psychology and across a vast array of judgment and decision tasks and domains (e.g., health, wealth, relationships, and happiness). While it is notable that general cognitive skills and abilities are robust predictors of decision biases, these relations in and of themselves do not necessarily answer the question of *why*? While such individual differences are almost certainly determined by multiple factors, what follows are brief descriptions of two well-integrated theoretical accounts of the cognitive processes and causal mechanisms that primarily give rise to, and explain, differences in generally superior decision making (i.e., biases and error avoidance):

(a) **Dual Process / Dual Systems Theory** (Evans & Stanovich, 2013; Kahneman, 2003, 2011). This theory emphasizes the central role of *logical, cold calculation* in unbiased decision making (e.g., System 2 overrides emotions and biased intuitions generated by System 1, so that System 2 can then logically compute or analyze decisions). Theoretically, a primary mechanism that gives rise to differences in the cognitive process associated with *cold, logical calculation* is a fundamental difference in the *heritable and abiding* inhibitory and storage capacity of System 2 (e.g., larger short-term working memory capacity, greater inhibitory and attentional control). Theoretically, because more intelligent people have larger System 2 capacities, they can more easily and reliably (a) inhibit biased and emotional System 1 intuitions, and (b) hold more abstract and complex logical equations in their (short-term) working memory (e.g., computing a formal decision analysis or solving a difficult math equation).

(b) **Skilled Decision Theory** (Cokely et al., 2018). This theory emphasizes the role of *vivid, representative understanding* in skilled and adaptive decision making (e.g., System 1 and System 2 collaborate and iteratively evaluate and encode risks into a personally meaningful understanding in long-term memory). Theoretically, the primary mechanisms that enable the cognitive process differences associated with vivid, representative understanding are differences in *acquired* skills, knowledge, and the organization of information in long-term memory. As such, because people are more skilled at probabilistic reasoning they may (a) use System 2 to more accurately evaluate and interpret risks so they can then (b) use elaborative encoding to develop a personally meaningful understanding in long-term working-memory, thereby circumventing the capacity constraints of (short-term) working memory (System 2), by creating a well-informed and emotionally-calibrated intuitive understanding in long-term memory (e.g., developing a realistic, vivid narrative story detailing the probabilities and imagining possible decision outcomes and how they might feel).

According to the popular theoretical account provided by *Dual Systems Theory*, the differences in decision making biases result from differences in (largely) stable and abiding cognitive capacities (Frederick, 2005; Kahneman, 2003). As such, only *a small proportion of* individuals might ever be able to engage in superior and unbiased decision making. That is, highly intelligent people avoid biases because they are endowed with a larger mental capacity and greater attentional control (i.e., inhibit, shift, update) allowing them to disregard intuitive feelings and override automatic cognitive processes (e.g., heuristics), while also more effectively holding logical (cold calculating) reasoning processes in their (short-term) working memory (e.g., the calculation of some subjective expected utilities, as described in *The Foundations of Statistics*; Savage, 1954). In contrast, *Skilled Decision Theory* suggests that individual differences in decision making quality are not (primarily) constrained by differences in intelligence or other working memory and attentional control, but rather are primarily constrained by the kinds of skills and knowledge one has acquired and can bring to bear on the task.

Clearly, both theories posit that general cognitive abilities of some type play a defining role in supporting specific decision making processes (e.g., cold rational optimization vs. meaningful, representative understanding). However, because the Dual Systems perspective emphasizes the importance of heritable cognitive capacities, this theory suggests that in general, the quality of decision making should not be directly or substantially improved by training or skill acquisition (e.g., heritable intellectual capacities largely define functional capacity limits).

In contrast, Skilled Decision Theory suggests the opposite: trainable, acquired skills and knowledge may causally give rise to differences in general decision quality. Specifically, Skilled Decision Theory suggests that training, decision support, and other interventions can and should be beneficial, in many ways. Ultimately, Skilled Decision Theory explains differences in general superior decision making via the same mechanisms that promote the profoundly superior judgment and decision making processes of verifiable experts (i.e., specific differences in knowledge and skills that are acquired through extended deliberate practice). To further contextualize theoretical, technical, and historical foundations with respect to cognitive ability testing, I next provide a brief review of (i) measurement in psychology (ii) measurement of intelligence, and (iii) the assessment of numeracy, its components and its relation to risk literacy, and superior decision making.

Psychology, Measurement, & Science

In a specific sense, scientific measurement is not possible in Psychology or related Behavioral and Social Sciences. This was the conclusion and official determination of the Ferguson Committee (1940), which was formed in 1932 by the British Association for the Advancement of Science. Composed of 19 scientists, including well-respected Physicists and Psychologists, their charge was to investigate the scientific validity of psychological measurement practices. Although all members did not agree on the formal conclusion, based on common standards and assumptions in philosophy of science and in Newtonian Physics, they determined that psychological variables were not measurable or scientific in the same way as other scientific variables because even the most rigorously measured psychological variables did not measure actual physical "quantities."

Broadly, the argument of the Ferguson Committee was that science generally requires objective "direct" measurement of physical quantities that are a function of some *invariant* (e.g., constant, universal) physical substance, such as length, weight, or time, and so naturally affords precise quantification and direct discovery. That is, if all such substances can be precisely and reliably measured with respect to their real (invariant) properties, natural scientists can simply use direct measurement to reveal, discover, and catalog invariants. For example, with the right tools, any competent and qualified observer could measure the "true" quantity of 8 seconds, which would necessarily be exactly four times longer than 2 seconds, regardless of who measured the quantity. By contrast, because psychological variables are primarily theoretical, psychologists would not likely aim to actually measure some real "mental quantity." For example, in measuring the contents of working memory psychologists do not purport to be measuring something like an actual physical "space" that is filled up with 7+/- 2 "cognitions," which take up 7 times more space than any 1 "cognition" or idea.

While much has changed, in some ways the ideals of realism in scientific measurement reflected by the Ferguson Committee continues to be represented in many modern standards of measurement in science and engineering. For example, the International System of Units (SI) defines seven fundamental measurement units (i.e., ampere, candela, kelvin, kilogram, meter, mole, and second), all of which are precisely defined in physical terms with reference to a specific, directly measurable, *invariant* (universal) physical constant (e.g., temperature is measured in Kelvin units based on the Boltzmann Constant; weight is measured in kilograms based on Planck's Constant). Although there are clearly some relatively direct physical relationships psychologists might reasonably aim to measure (e.g., the extent to which changes in pain are directly linked to changes in the firing rate of nociceptors neurons), most variables of psychological interest should, by definition, be considered conceptual variables or theoretical constructs (e.g., attention, memory, anger, intelligence). Thus, in psychology, the measurement of any psychological variable is by its very nature a theoretical enterprise.

Representational Theory of Measurement. In some way, the Ferguson Committee's finding (1940) that psychologists do not measure anything universally invariant or otherwise "real," was probably poorly timed and certainly largely ignored. By that time, measurement and quantitative methods were already fairly well-established and were also quite well-funded, thanks in part to the need for testing and assessment in educational, clinical, occupational, and military contexts. For example, nearly 20 years before the Ferguson report was released, many U.S. psychologists who were members of the American Psychological Association were actively involved with some aspect of measurement and test development in applied psychology (Terman, 1921). Gustav Fechner's (1860) work on psychophysics from nearly a century before the report was released had served (with other works) to set a foundation for the professional sub-field of quantitative psychology. Moreover, "measurement" and quantitative practices in psychology had produced fundamental and practical advances in statistics that were of interest well-beyond the field. For example, Francis Galton's work is recognized for making essential contributions to the conceptualization of the standard deviation and correlation (e.g., regression to the mean). Charles Spearman's work led to the establishment of the rank-order correlation and is often more generally credited as developing factor analysis.

Given the many quantitative achievements, and the vast amount of research taking place in applied psychology and testing in the 1940s, perhaps it is not surprising that soon after publication of the report a formal response was published by Stevens (1946). That paper advanced and rather firmly established what remains today to be the modern, standard approach to measurement in psychology—i.e., the operationalization of psychology and quantitative measurement. Roughly, this view holds that measurement of psychological and other theoretical constructs may be accurately described using the assignment of some operationally defined variable, in accord with a specified measurement scale (i.e., nominal, ordinal, interval, ratio). As such, although the construct may only be theoretical in nature, and thus does not necessarily have a truly *invariant* form, it nevertheless may be usefully, reliably, and accurately characterized by its functional measurement properties, provided an adequate theoretical conceptualization and operationalization of the measured variable (e.g., cognitive abilities as measured by X according to Y scale). While many have presented concerns about the logic and rigor of this initial formulation, more recent axiomatic treatments of related representational theories of measurement have been successful, including Luce and Tukey's (1964) conjoint theory of measurement, further refined in Luce and Suppes (2002), which today stands as one of the leading theories of scientific measurement in any scientific field of study. As such, considerable care is merited when making claims about how to interpret some measurement scale (e.g., ordinal vs. interval), including the need for careful investigation of aspects and assumptions of test properties. These notions and requirements are today commonly embodied in the standards for development of tests, norms, and other methodological elements, as described by various frameworks for construct validity (Messick, 1995), and codified in professional standards such as the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999).

Measuring Intelligence. At the turn of the twentieth century, French psychologist Alfred Binet was hired by the French Ministry of Education to determine which students were not learning effectively in standard classrooms, so they could be given additional remedial work (i.e., a type of special education). Binet (1903) devised a written test and in the next decade this work spread to the U.S. In 1916, Lewis Terman standardized the Stanford-Binet Intelligence Scale, and then worked on a committee to create the Army Alpha, another test used to determine a soldier's capability to serve (1917). While Binet did not believe that a single, permanent, inborn level of intelligence could be determined using these psychometric instruments, Terman believed there was a hereditary component to intelligence and was particularly interested in studying individuals who demonstrated extreme talent (i.e., geniuses and gifted children).

Over the next few decades, research on intelligence and individual differences of abilities took many turns. In 1921, Terman started his 'Study of the Gifted,' and enrolled 1,444 "gifted" students in one of the longest running longitudinal studies to date (Terman & Oden, 1947). As notions of the heritability of intelligence (and feeblemindedness) grew, Francis Galton promoted the Eugenics movement (also supported by Edward Thorndike and Terman), which led to critical cases in American history, including the case of the Kallikak family (studied by Goddard in 1912), and the 1927 U.S. Supreme Court Case of Buck vs. Bell, which ultimately upheld the sterilization of Carrie Buck (who supposedly was part of a long line of "feebleminded" women). Additionally, Edward Thorndike, one of the early proponents of the eugenics movement, was particularly interested in individual differences in intelligence, and believed that these variations in intelligence were (primarily) due to innate, hereditary capacities, and *not* due to environmental influences. As such, he believed that genetic endowment was one of the most critical variables for social and economic progress. As his research progressed, his son, Robert Thorndike became recognized as one of the "fathers of personnel psychology," having created tests to weed out unfit employees and locate those who would perform well on the job (Thorndike, 1949). He worked to determine the lowest IQ necessary to carry out any particular job, and advocated for educational practices that prepared students for careers that matched their innate abilities.

Despite these early traditions that continue to exert influence in many ways (e.g., Murray, 2021) there are also many other modern trends in intelligence research that acknowledge a more nuanced and sophisticated view of the knowns and unknowns in intelligence research. For example, as the individual differences movement in intelligence research continued in the early part of the century, Spearman (1927) developed Factor Analysis (i.e., a method of separating a construct into a number of hypothetical factors or abilities) and posited that there was likely one primary factor in determining intelligence (i.e., *g*; or the general factor). However, in 1941 Raymond Cattell proposed two primary factors of intelligence – including fluid (i.e., the capacity to reason and solve novel problems, independent of any knowledge from the past), and crystallized (i.e., the ability to use acquired skills, knowledge, and experience, which relies on accessing information from long-term memory; Cattell, 1963; Cattell et al., 1941).

Modern Intelligence Research & Theory. Today, research on intelligence has grown from these initial roots. Hiring organizations continue to use tests of intelligence for selection purposes (Schmidt & Hunter, 1998), schools use clinical measures to assess student aptitude for learning, and people affected by the eugenics movement remain alive. However, while there have been many new models and assessments of intelligence (e.g., Hunt, 2010; Raven, 1938; Sternberg, 1977; 2003; Thurstone, 1938; Weschler, 1955), the work of Spearman (1927) and Cattell remain pervasive in psychometric models of cognitive abilities (Carroll, 1993; Cattell, 1971, 1987; McGrew, 2009)

Notably, in his seminal work, Carroll (1993) proposed a now widely endorsed *Three Strata Theory* of human intelligence by analyzing over 460 data sets from 1927 and 1987. This model suggested the cognitive abilities of humans are typically well-characterized by three levels. Stratum 1 represents the rote tasks that can be used as indicators. Stratum 2 provides eight broad factors of cognitive abilities, namely: Fluid Intelligence (Gf), Crystallized Intelligence (Gc), Broad Visual Perception (Gv), Broad Auditory Perception (Ga), General Memory and Learning (Gy), Broad Retrieval Ability (Gr), Broad Cognitive Speediness (Gs), and Reaction Time & Decision Speed (Gt). Finally, Stratum 3 follows from Spearman's factor analysis hypothesis: there is one overarching general ability factor, *g*. Carroll (1993) further suggested that three essential reasoning factors best explained fluid intelligence, namely (i) sequential reasoning, (ii) inductive factors, and (iii) quantitative reasoning.

Around the same time, *The Bell Curve: Intelligence and Class Structure in American Life* (Herrnstein & Murray, 1994) was published, making a number of controversial claims, including: (i) intelligence (fixed innate capacities) is the primary predictor of life outcomes, (ii) the reason for this connection is that intelligence is a strong determinant of decision making quality, (iii) only a small number of people have the aptitude required to make consistently good decisions, and (iv) there are a few appropriate policy prescriptions, which would place the cognitive elite in ruling positions (Herrnstein & Murray, 1994; but see also Cokely et al., 2018; Heckman, 1995).

Following its publication, the mainstream media as well as many acclaimed researchers responded with several criticisms and defenses. First, in 1994 Linda Gottfredson published a statement in the Wall Street Journal that included 25 statements (largely in support of *The Bell Curve*), which was signed by 52 psychologists (Gottfredson, 1997). Then, the American

Psychological Association (APA) compiled a task force to publish a report on the current state of intelligence research (Neisser et al., 1996). Stephen Jay Gould also issued a reprint of his book, *The Mismeasure of Man* (1981; 1996) with a new forward explicitly critical of Herrnstein and Murray's (1994) claims. Additionally, Heckman (1995) presented a reanalysis of the data presented in *The Bell Curve* and demonstrated (amongst other things) that a brief test of mathematical operations was excluded from the original analyses, but that this short test predicted wages (i.e., life outcomes) as well as the Armed Forces Qualifications Test which Herrnstein and Murray (1994) reported.

Of note for the current project, this last finding suggests that "numerical abilities must explain important life outcomes that are otherwise missed by more standard general intelligence metrics" (Cokely et al., 2018, p. 485). Though in recent years there has been increasing research on the relationship between cognitive abilities and decision making (Bruine de Bruin et al., 2007; Frederick, 2005; Stanovich & West, 2000), relatively less research has considered the relationship between (i) standard cognitive abilities, (ii) decision making skill, and (iii) numeracy. Said differently, though Carroll (1993) analyzed over 460 datasets, measures of decision making skill and statistical numeracy were relatively underrepresented (Cokely et al., 2018). As such, while previous analyses of cognitive abilities have determined that fluid and crystallized intelligence are the two primary factors of intelligence (and primarily predict all following life outcomes; Carroll, 1993; Cattell, 1971; McGrew, 2009), recent research suggests that when properly represented, another factor, *General Decision Making Skill* is also likely an important and potentially defining factor in a more comprehensive model of human cognitive abilities (Allan, 2018; Cokely et al., 2018).

The reason for this shortcoming in the analysis of intelligence tests is likely two-fold. First, while psychometric research on cognitive abilities has been ongoing for over a century, the research on decision making competence has only emerged within the last 50 years (Dhami, et al., 2012). Second, while logic can be divided into two major categories (deductive and inductive), fluid intelligence tests are primarily about careful and thorough *deductive* reasoning under conditions of certainty – ultimately neglecting the importance of *inductive* logic in accounting for decision making ability (i.e., under conditions of uncertainty). More specifically, the standard tasks used to measure fluid intelligence tend to rely on working memory and attentional control capacities (Kane et al., 2004; McCabe et al., 2010). Given the importance of long-term representations in memory (i.e., representative understanding), these fluid intelligence tasks are not likely representative of typical human decision making.

This finding on the neglected and under-represented role of statistical numeracy in intelligence assessments suggests a few important implications. First, previous models overemphasized the role of fluid intelligence and working memory in models of cognitive abilities (i.e., relatively innate capacities). Despite recent attempts, the effectiveness of working memory and fluid intelligence training remains unknown (Jaeggi et al., 2011). In contrast, recent attempts to train and improve decision making skill and statistical numeracy are abound (Jenny et al., 2018; Nisbett, 2009; Peters, 2017; Peters et al., 2010; Ybarra et al., 2017). Taken together with the other previously reviewed work on numeracy, this research suggests that perhaps the most influential variables that link cognitive abilities, decision making, and life outcomes are not inherent, but are rather *acquired* skills (e.g., crystalized intelligence, statistical numeracy, risk literacy).

Measuring Numeracy & Decision Making

Measuring Numeracy. The conceptual origins of the construct of numeracy can be traced in large part to an interest in mathematics for use in everyday life, beyond the classroom (e.g., “practical numerical literacy,” “quantitative literacy,” or “mathematical literacy”). One of the earliest documented uses of the term numeracy was in England, where it was presented as part of the Crowther Report on education in 1959 (but see also Huff, 1954; Paulos, 1988). Crowther used “numeracy” to specifically distinguish applied and practical mathematics, from the abstract mathematics performed in classrooms and in many technical professions. While mathematical skills and achievement have been of keen interest throughout the history of formal education, the specific scholarly interest of theoretical and practical *numeracy* is only 60 years old. Nonetheless, it has been a busy 60 years.

Modern definitions of numeracy are largely consistent with early conceptual foundations. For example, the Organization for Cooperation and Economic Development has defined numeracy as the “ability to access, use, interpret and communicate mathematical information and ideas in order to engage in and manage the mathematical demands of a range of situations in adult life,” (OECD, 2013; see also Ginsburg et al., 2006; Krenzke et al., 2020). Programs such as the National Assessment of Adult Literacy and the Programme for International Student Assessment have also focused on related assessments to evaluate literacy levels more comprehensively (i.e., reading and numerical) across time and groups (e.g., ability levels in different countries; see Breakspear, 2012; Kutner et al., 2006). Results from studies such as these suggest that although numeracy may be a fundamental and relatively essential skill, more than 62 million adults (e.g., nearly 30% of U.S. adults) have such low levels of numeracy skills that they struggle with even very basic and common calculations (NCES, 2020).

To illustrate common limits of statistical numeracy (i.e., practical probabilistic reasoning), consider a probabilistic national sample comparing statistical numeracy scores collected in U.S. and Germany. The study found that only about 30% of adults could correctly identify the “larger probability” of getting sick (e.g., 1 out of 100 vs. 1 out of 10; Galesic & Garcia-Retamero, 2010). These low numeracy levels have real-world implications, because low numeracy is often tied to low health literacy and worse health outcomes (e.g., CDC; 2021). Extensive evidence also demonstrates that quantitative skills more generally are among the most influential educational variables associated with economic prosperity in industrialized countries (Hanushek & Woessmann, 2010; Hunt & Wittmann, 2008). However, in the last 25 years empirical research on the relationship between numeracy and decision making proliferated, in part thanks to the publication of a simple three-item numeracy test (Schwartz, et al., 1997) that focused specifically on probabilistic reasoning or what has become known as *statistical numeracy*. This test was created to assess the relationship between numeracy and one’s understanding of information about risks associated with breast cancer screening. However, it has since developed a much wider interest in the relationship between statistical numeracy, risk, and decision making in many risk communication domains (Nelson et al., 2008; Schapira et al., 2009).

Broadly speaking, numeracy research can be conceptually divided into four number-related competences including:

(i) **Objective numeracy and numeracy subcomponents**: the ability to solve practical math problems accurately, including various subdivisions of component numeracy skills such as (Allan, 2018; Cokely et al., 2012, 2018; Ghazal, 2014, see also Peters, 2012; Reyna et al., 2009):

a. Statistical numeracy (e.g., operations and probabilities)

b. Conventional numeracy (e.g., geometry and algebra)

c. General numeracy (e.g., geometry, algebra, operations, probabilities)

(ii) **Subjective numeracy**: one's subjective self-assessment of one's own numeracy competencies, as compared to others, or with reference to different tasks.

(iii) **Symbolic-number mapping abilities**: fundamental spatial mappings of magnitude, space, ratio, etc., defined as "internal representations of numeric magnitude" (Peters & Bjalkebring, 2015, p. 1).

(iv) **Domain-specific numeracy:** numeracy for use in specific contexts such as medical or health numeracy or numeracy for financial literacy (see for example the NUMi; Schapira et al., 2014),

Of note, when considering these divisions, research has started to identify interdependencies as well as dissociations (and distortions) that can emerge when comparing across different measurements. For example, while some research suggests that subjective numeracy can serve as a valuable and easier proxy for objective numeracy, evidence has also revealed limits. For example, though 70% of people in one sample self-report they are (subjectively) numerate, only 2% answered all objective numeracy items correctly (Miron-Shatz et al., 2014). Related, in educational and developmental psychology, some evidence suggests that symbolic number mapping may emerge earlier than numeracy, acting as a "precursor" to further development of objective numeracy. Theoretically, this connection could potentially provide another indirect path toward the acquisition of more robust objective numeracy skills (e.g., developing skills across individuals with low objective numeracy in more basic symbolic number mapping tasks; Peters & Bjalkebring, 2015).

Statistical Numeracy & Risk Literacy. Many studies of numeracy and decision making have shown that the various measures of numerical competencies can and often do predict differences in decision making and decision vulnerability. Nevertheless, a large body of research suggests that generally superior decision making may be most tightly linked to individual differences in statistical numeracy, as measured by tests such as the Berlin Numeracy Test (i.e., probabilistic reasoning and problem solving). Since 2012, more than 100,000 diverse adults from 150+ countries have participated in research on the relations between statistical numeracy, skilled decision making and cognitive biases, including the investigation of thousands of different judgments and decisions sampled from naturalistic and field-based studies (e.g., delay in seeking medical care during a heart attack, misunderstanding the risk of Ebola and other diseases, mitigating financial credit risk, reducing risk of disease transmission), and paradigmatic laboratory-based investigations (e.g., risky prospect evaluation, risk perceptions and interpretations, intertemporal choices, self-regulation and self-assessment, theory-of-mind; see Skilled Decision Theory for a review, Cokely et al., 2018; see also Cokely & Kelley, 2009, Garcia-Retamero & Cokely, 2017; Ghazal et al., 2014; Gigerenzer, 2015; Peters, 2020; Petrova et al., 2016; Reyna et al., 2009). Moreover, statistical numeracy has recently been identified as one of the strongest predictors of COVID-19 misunderstanding in samples from diverse countries on three continents (Pennycook et al., 2020; Roozenbeek et al., 2020). Statistical numeracy skills have also been found to reduce the influence of motivated cognition in several controversial, high-stakes domains (e.g., climate change beliefs, risk perceptions; Cho, 2020; Johnson, 2008; Ramasubramanian, 2020). Statistical numeracy has also been linked to a growing number of other demonstrations concerning high-stakes natural hazards and weather-related decisions (e.g., interpreting forecasts, avoiding weather myths, recognizing flood risks; food and water quality standards; see Allan et al., 2017, 2020; Cokely et al., 2012; Feltz & Feltz, 2019; Mahmoud-Elhaj et al., 2020; Ramasubramanian et al., 2019).

Although early judgment and decision making research emphasized the role of abstract, emotionless decision analysis (e.g., explicitly calculating expected utilities), over the last 10 years a collection of findings from various process tracing, decision support, and training studies are inconsistent with this view. Instead, the weight of the evidence suggests that the primary difference may be better explained as a function of Risk Literacy (i.e., the ability to evaluate and meaningfully understand risk information) as measured by statistical numeracy, in accord with Skilled Decision Theory. Moreover, statistical numeracy tests tend to be robust predictors of decision quality *because* they predict differences in risk comprehension, which then influence attitudes, intentions, decisions, and behaviors. To be clear, a naïve theoretical view is that increased numeracy skills tend to be important for decision making *because* numeracy helps people *calculate* and solve complex expected value/utility equations, perhaps as described by Dual Systems Theory (e.g., numeracy predicts better decisions because numerate people just do the math). Yet, in 2006, Peters and colleagues showed that higher scores on a numeracy test tend to be systematically associated with decision making biases, at least as measured by specific kinds of risky prospect evaluations involving extremely low amounts (e.g., hypothetical gambles involving less than $1; Peters et al., 2006). This finding revealed that more numerate individuals were (i) sometimes biased as compared to less numerate people and so (ii) could not be (primarily) relying on expected value calculations (e.g., ultimately suggesting that affect sometimes provided a cue in decision making for highly numerate people).

In 2009, more direct evidence on the relationship between risk literacy and statistical numeracy came from a process tracing study (Cokely & Kelley, 2009). The study directly traced cognitive processes to test assumptions about the role of calculation versus understanding using a set of relatively simple but wide-ranging risky prospect evaluations (e.g., lotteries) in a college undergraduate sample. The study ultimately modeled the relations among (a) decision latency (b) retrospectively reported decision strategies (i.e., protocol analysis) (c) choice patterns (d) quality of risky prospect decisions and (e) individual differences in general cognitive abilities, including numeracy, working memory, and cognitive impulsivity as measured by the Cognitive Reflection Test.

Findings revealed that cognitive abilities and superior decision making were *not* associated with objective calculations of expected values (e.g., “75% of $200 is $150, which is more than $100 for sure”). Less than 5% of the total sample expressed any processes that were consistent with expected value calculations. Even more ironically, the vast majority of participants who consistently made normatively superior decisions under risk later failed a math test requiring participants to explicitly solve the same level of math problem in accord with expected value calculations (e.g., which is greater: 3% of $7000 or $350). Instead, integrated choice and process modeling revealed that the interaction between cognitive abilities and superior decision making was almost entirely explained by differences in the amount of affective-charged, personally-relevant heuristic (narrative) deliberation expressed during risky prospect evaluation (e.g., “well, I sometimes make $200 bucks in one night but, $1,000 is getting close to my tuition, and like on a coin toss, totally not worth it”). More numerate decision makers also took more time to decide during the first phase of the study, despite being better at math and otherwise scoring higher on tests of other abilities. Results suggested that even the most numerate people often rely on heuristics and elaborative (i.e., meaningful and affectively charged) memory encoding processes, rather than solving even very basic expected value equations (Cokely & Kelley, 2009).

More recent work at the University of Oklahoma (Allan, 2018; Cokely et al., 2018) conducted one of the most comprehensive analyses of the relations between cognitive abilities tests (intelligence, numeracy) and superior decision making, including data from over 300 participants who completed a five-hour assessment battery. Participants completed multiple standardized measures of fluid and crystallized intelligence, including Ravens' Advanced Progressive Matrices, the Cattell Culture Fair Test, the Wonderlic Personnel Test, and the Employee Aptitude Survey (Cattell, 1973; Raven et al., 1988; Wonderlic, 1983, 2018). Using confirmatory factor analysis and structural equation modeling, a new model of general decision making skill was presented, where full-scale numeracy (i.e., objective numeracy measures for statistical and conventional numeracy) fully mediated the relationship between intelligence and decision making skill. Furthermore, a higher-order factor analytic model demonstrated that when numeracy and decision making skill are represented in a model of cognitive abilities, a new factor structure emerges, whereby (i) numeracy and decision making skill, (ii) crystallized intelligence, and (iii) fluid intelligence are three distinct factors. Moreover, when explaining general intelligence, *g*, no factor had more influence than that of numeracy and decision making skill (Allan, 2018; Huck, 2020). Research further suggest that statistical numeracy tests may generally double the predictive power of other standardized tests of general cognitive abilities, including tests of fluid intelligence (Ghazal, 2014).

Some of the strongest evidence of the causal relations between meaningful understanding and decision making come from randomized control experiments with visual aids such as research where the experimental condition group gets a visual aid (e.g., graph) to support their decision making. More direct evidence includes a large body of research on causal interventions using decision aids (e.g., transparent visual aids). This research shows that low ability decision makers can and often do match the decision making performance of higher ability participants, when they have help interpreting the risks (Garcia-Retamero & Galesic, 2010). A recent systematic review of visual aid studies involving more than 25,000 participants from several countries supports this finding (Garcia-Retamero & Cokely, 2017). Perhaps even more compelling, there is a growing body of work showing that training essential decision making skills, including specific risk literacy skills (i.e., graph literacy), can causally improve conceptually diverse decision making skills (e.g., ratio bias, sunk costs, framing effects). In turn, this training generalized further to improve metacognitive self-assessments and social comparisons, substantially reducing overconfidence vulnerability, independent of numeracy or other cognitive ability variables (Ybarra, 2021).

Current Study

Taken together, the studies reviewed in this chapter suggest that the relationship between cognitive abilities and superior decision making may generally be explained by differences in acquired skills that promote risk literacy (i.e., the ability to evaluate and meaningfully understand risk), such as statistical numeracy and graph literacy skills. Presently, however, there is no IRT based stand-alone Risk Literacy assessment (e.g., a latent trait model-based, norm-referenced direct test of Risk Literacy), nor is there a norm-referenced assessment of the original Berlin Numeracy Test linking it directly to other general risk literacy criteria, in a representative sample. Provided the opportunity to test and refine these measures in a representative sample, there is also an opportunity to address a major potential threat that has yet to be investigated with respect to numeracy and risk literacy, namely the potential influence of differential item functioning and measurement invariance on specific subgroups of participants (i.e., men and women). While this bias has been documented in other tests of intelligence, to date, limited research has examined the robustness of numeracy assessments using measurement equivalence methods. Lastly, the present study will also focus on developing three quantitatively precise metrics to predict, quantify, and communicate about *what* risk communication is likely to be difficult, *how much,* and *for whom* it will likely be difficult (e.g., how many people are likely to misunderstand, how do two different risk communications compare, and what is the minimum level of numeracy skill people will need to understand a risk communication). These questions about *what (item-specific difficulty), for whom (person-specific difficulty),* and *how difficult (norm-reference cumulative distributions)* a task is can then provide a basis for the development of specific metrics including: *risk literacy difficulty levels, decision vulnerability benchmarks,* and *minimum numeracy skill thresholds.*

General aims of the current project are:

(1) Develop, evaluate, and compare Norm-referenced tests of numeracy and risk literacy using IRT latent trait modeling and data from a probabilistically representative study of U.S. adults, in accord with modern psychometric standards, including:

   a. An investigation of subgroup differences in overall test achievement

   b. An analysis of the source of subgroup differences in test achievement (e.g., measurement invariance and differential item functioning analyses of test score differences between men and women).

(2) Develop and test some potentially useful metrics that can be derived from the numeracy and risk literacy norms, focusing on the estimation of decision vulnerability, difficulty levels, and skill thresholds, which may be particularly useful in the context of benchmarking or comparing various high-stakes risk communications (e.g., how much numeracy is required to understand one extreme weather risk communication vs. another).

Chapter 2

Risk Literacy Norms and Decision Vulnerability Metrics

Based on a probabilistically representative national sample, the current study was designed to develop test norms, evaluate measurement assumptions (i.e., differential item functioning), and to investigate a protocol for estimating three decision vulnerability metrics (i.e., difficulty, skill threshold, and probability of misunderstanding by group). The report of findings begins with the presentation of norm-referenced test scores and response patterns, including IRT parameters for the Berlin Numeracy Test and the newly developed Risk Literacy Test. I then report an analysis of measurement equivalence on these measures, focusing on a key demographic variable, gender, which is a theoretically noteworthy target for analysis (e.g., potential score bias could result from gender differences in math anxiety, stereo-type threat, etc.).

Next, I present an example framework and protocol for a decision vulnerability analysis method, developed as an initial proof of concept. The protocol was designed to accommodate the translation of out-of-sample response patterns into norm-referenced indices for benchmarking. As such, I developed methods for estimating three norm-referenced indices, namely (i) Risk Literacy Difficulty Level (estimated via the cumulative distribution of achievement), (ii) Decision Vulnerability Benchmarks (a general and subsample estimate of the probability of risk misunderstanding or bias), and (iii) Numeracy Skill Threshold (the minimum numeracy skill level required to have >50% likelihood of accurately interpreting risks relevant to the decision criterion, presented as a raw test score on the Berlin Numeracy Tests).

Finally, as an initial investigation into the robustness and validity of the protocol and indices, I present model recovery and hold-out type tests, examining the precision, reliability, and validity of the regression-based transformation methods. These methods were selected in part to be user-friendly for a wide range of risk communication and decision making researchers. Because most researchers are at least somewhat familiar with regression analyses, the logic of the framework should be transparent, which may also make the protocol less challenging, less likely to cause confusion or error, and easier to interpret and explain as required for peer review. For example, the reporting of results from a new study, could include comparative analyses based on norm-referenced benchmarks (e.g., the general risk literacy difficulty level of a new risk communication compared to a previously published risk communication). I then close with an out-of-sample practical application, providing an example of how the protocol could be used. This should help illustrate several limitations and highlight specific relevant concerns that should not be neglected (e.g., limitations as a function of criterion discriminability and need for skill stratification or performance validity assessments).

Participants

The sample was collected in Spring 2016 using a probability-based sampling procedure (KnowledgePanel® from GfK). GfK recruits panel members by using address-based sampling methods. Once on the panel, panelists are then contacted via email to participate in online surveys. Panelists without access to the appropriate technology were provided with access to the Internet and hardware as needed. Though the typical GfK survey is 10 to 15 minutes per survey, the current survey was unique in that it took roughly 1 hour to complete.

Of the 305 participants, 142 (46.5%) were female and 163 (53.4%) were male. Participants were between 18 and 86 years of age. Information on the representativeness of this sample is provided by comparing the sample to U.S. Census estimates of the resident population (See Table 2.1).

Table 2.1

Demographic Representativeness of U.S. Sample (2016)

	Current Sample Respondents n (%)	U.S. Adult Population* (%)
Gender		
Female	142 (46.6%)	51.3%
Male	163 (53.4%)	48.7%
Age		
18 to 34	84 (27.5%)	30.2%
35 to 54	123 (40.3%)	33.2%
55 and up	98 (32.1%)	36.4%
Ethnicity		
Hispanic	26 (8.5%)	15.8%
Non-Hispanic	279 (91.5%)	84.2%
Race		
White	245 (80.3%)	78.5%
African American	11 (3.6%)	12.8%
Other Race	49 (16.1%)	8.7%
Education		
Less than High School	36 (11.8%)	12.6%
High School	95 (31.1%)	27.7%
Some College	81 (26.6%)	31%
Bachelor and beyond	93 (30.5%)	28.7%
TOTAL	**305**	**-**

Note. Population estimates were obtained from the U.S. Census Annual Estimates of the Resident Population by Sex, Age, Race, and Hispanic Origin for the United States and States: April 1, 2010 to July 1, 2016 (PEPASR6H).

Measures and Methods

The data collected in this study included many measures for various other research purposes. In the present analysis, I focused on statistical numeracy and risk literacy assessments, with special attention to the psychometric validity and the development of standardized norms for the Berlin Numeracy Test and the Risk Literacy Test.

Berlin Numeracy Test & Schwartz Scale

The Berlin Numeracy Test (BNT) is one of the most efficient predictors of risk literacy and general decision making skills, especially for educated individuals from industrialized countries (Cokely et al., 2012, 2018). When studying a diverse population, the Berlin Numeracy Test is often paired with the three-item Schwartz et al. (1997) numeracy scale. The Schwartz test was one of the first published tests of statistical numeracy and it assesses individuals' understanding of proportions and probabilities. It is especially efficient for individuals with lower statistical numeracy ability. Taken together these two tests often provide a robust and efficient assessment for a wide range of skill. See Figure 2.1 and Table 2.2 for item and test statistics.

Risk Literacy Test

This set of items included both realistic risky decisions as well as paradigmatic risky prospect evaluations. Items assessed expected values, intertemporal choice, denominator neglect, and ecological risk literacy (e.g., medical and financial decisions that tend to be representative of the natural ecology; see Frederick, 2005; Okan et al., 2012; Pachur & Galesic, 2013). To develop an optimized set of risk literacy items, I used factor analysis and Item Response Theory to identify a unidimensional item set, with good psychometric properties (e.g., high discrimination and difficulty across the range of ability). Then, to assess measurement equivalence, I iteratively removed items that demonstrated differential item functioning. In the end, the optimized battery included seven items. Item text is presented in Appendix A. See Figure 2.1 and Table 2.2 for item and test statistics.

Risk Literacy and Numeracy Validation Items

An additional set of risk literacy and statistical numeracy items were included in the current study as a type of hold-out validation method for the decision vulnerability analysis (i.e., Analysis 3). Item text is presented in Appendix A.

Figure 2.1

Sample Distribution for (a) BNT-S (b) Risk Literacy Test

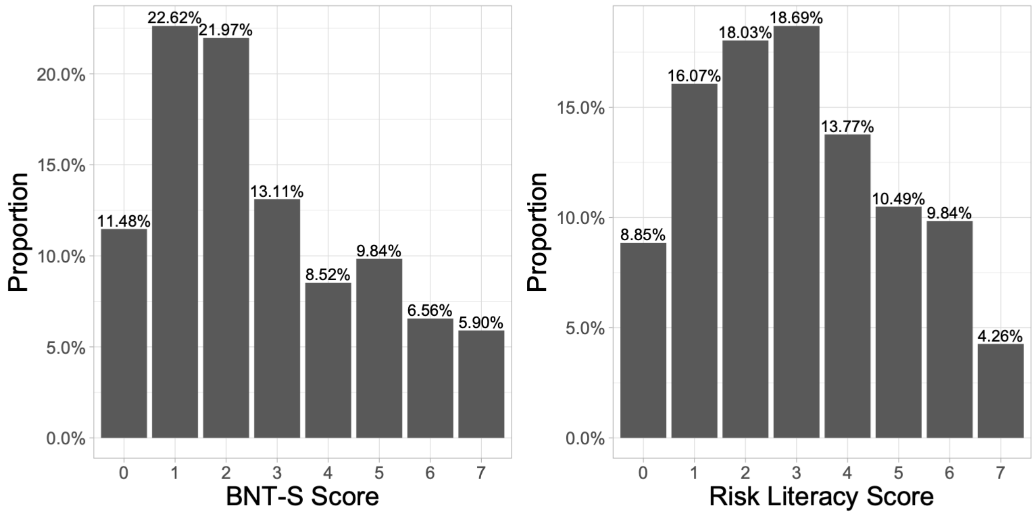


Table 2.2

Descriptive Statistics

	% Correct	SD	Skew	Factor Loadings	Difficulty	Discrimination
Berlin Numeracy Test-Schwartz (α = 0.77)						
S_cointoss	73.44	0.44	-1.06	0.41	-0.92	1.61
S_bigbucks	58.03	0.49	-0.33	0.45	-0.35	1.21
BNT_fivesided	41.31	0.49	0.35	0.73	0.21	3.44
S_acme	33.77	0.47	0.69	0.72	0.47	2.74
BNT_choir	27.54	0.45	1.01	0.62	0.78	1.96
BNT_sixsided	21.97	0.41	1.35	0.57	1.04	1.89
BNT_mushroom	13.77	0.35	2.10	0.47	1.61	1.60
Proportion of Variance				0.34		
Risk Literacy Test (α = 0.66)						
rlp_min	57.70	0.49	-0.31	0.55	-0.28	1.71
rlp_lose400	51.80	0.50	-0.07	0.52	-0.08	1.33
rlp_3400or3800	50.82	0.50	-0.03	0.40	-0.05	0.91
rlp_strokex1	49.18	0.50	0.03	0.62	0.02	2.02
rlp_beno	42.95	0.50	0.28	0.33	0.42	0.77
rlp_percentage	26.56	0.44	1.06	0.49	0.95	1.49
rlp_gain100	25.57	0.44	1.12	0.37	1.32	0.96
Proportion of Variance				0.23		

Note: FA using Oblimin rotation. Difficulty and Discrimination from a 2PL IRT model.

Analysis 1: Numeracy & Risk Literacy Norms

Norms are presented first for the general population, across each of the Berlin Numeracy Test and Risk Literacy Test forms (Table 2.3). Here, the cumulative percentile rank norms indicate the proportion of participants in the norming group who scored less than or equal to the given score. These tables can be used to interpret future scores, by providing a ranking relative to the current norming group. For example, a future score of 1 on the BNT-S would indicate that roughly 44% of the general U.S. population scores less than or equal to the test-taker. This alternatively would suggest that the test-taker scored worse than over 55% of the general sample – indicating that they may have a relatively lower level of statistical numeracy skills and may need extra assistance when interpreting complex probability information.

Table 2.3

U.S. Adult Population Norms (2016) Across Test Forms, in Cumulative Percentile Rank

Score	Schwartz	Full BNT	Adaptive BNT	Full BNT-S	Adaptive BNT-S	General Risk Literacy
0	15.41	45.90	-	11.48	15.41	8.85
1	44.26	70.49	52.13	34.10	44.26	24.92
2	75.08	85.57	72.46	56.07	60.66	42.95
3	100	93.44	84.92	69.18	70.16	61.64
4		100	100	77.70	81.31	75.41
5				87.54	85.57	85.90
6				94.10	90.49	95.74
7				100	100	100

Note. The Adaptive BNT is scored 1-4.

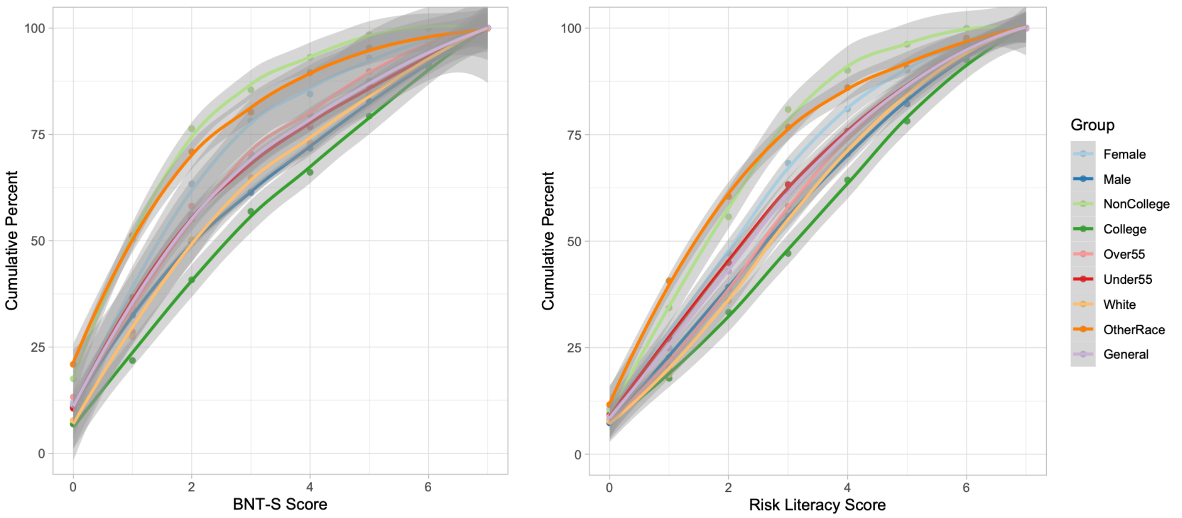
General Population and Demographic Group Norms

For the full scale BNT-S and Risk Literacy Test, norms are again presented for the general population, as well as for the four subgroups of interest, based on demographic characteristics. Using cumulative percentile ranks, the norms were stratified by gender (male and female), education (college educated and non-college educated), age (under 55 and 55+), and race (white and other race[1]).

For both the BNT-S and the Risk Literacy Test, the largest demographic difference existed between non-college and college educated samples (Figure 2.2). For the non-college educated sample, a BNT-S score of 4 ranked at the 93rd percentile. Conversely, amongst the college educated sample, a score of 6 ranked only in the 90th percentile. At lower BNT-S scores (e.g., BNT-S score of 1, 2 or 3) this translated to a 30-percentile point difference between non-college and college educated samples (See Figure 2.3 and Table 2.4). The same patterns held for the Risk Literacy Test (See Figure 2.4 and Table 2.5). Norms for the other test versions (Schwartz, BNT, Adaptive BNT, and Adaptive BNT-S) are also presented, stratified by gender, education, age, and race, and presented in cumulative percentile ranks (Tables 2.6 to 2.9).

Figure 2.2

Cumulative Distributions of Raw Scores on the (a) BNT-S and (b) Risk Literacy Test



[1] Other race includes Black/Non-Hispanic, Other/Non-Hispanic, 2+ Races/Non-Hispanic, Hispanic, as defined by the GfK KnowledgePanel®.

Figure 2.3

BNT-S Norms Stratified by Gender, Education, Age, and Race

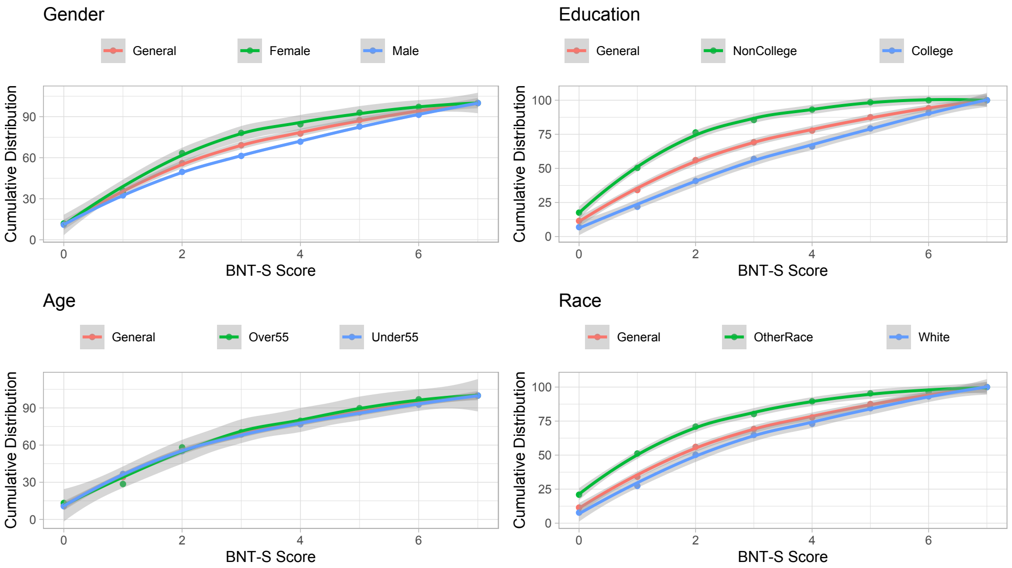


Table 2.4

BNT-S Norms for U.S. Adults (2016) Stratified by Gender, Education, Age, and Race

				Non		Under	Over		Other
Full BNT-S Score	General	Female	Male	College	College	55	55	White	Race
0	11.48	11.97	11.04	17.56	6.90	10.63	13.27	7.76	20.93
1	34.10	35.92	32.52	50.38	21.84	36.71	28.57	27.40	51.16
2	56.07	63.38	49.69	76.34	40.80	55.07	58.16	50.23	70.93
3	69.18	78.17	61.35	85.50	56.90	68.60	70.41	64.84	80.23
4	77.70	84.51	71.78	93.13	66.09	76.81	79.59	73.06	89.53
5	87.54	92.96	82.82	98.47	79.31	86.47	89.80	84.47	95.35
6	94.10	97.18	91.41	> 98.47*	90.80	92.75	96.94	93.15	96.51
7	100	100	100	100	100	100	100	100	100

Note. *No participant in the norming group's non-college subsample scored BNT-S = 6.

Figure 2.4

Risk Literacy Norms, Stratified by Gender, Education, Age, and Race

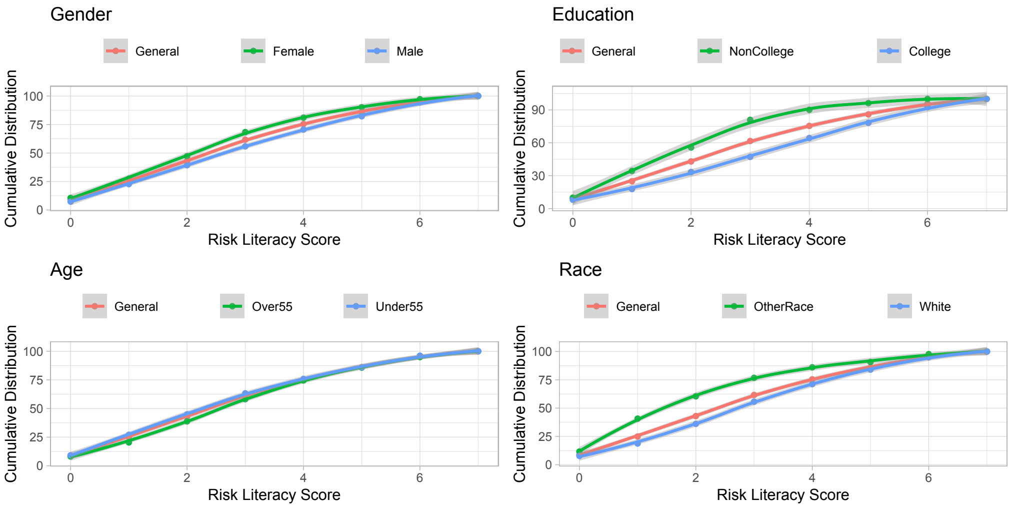


Table 2.5

Risk Literacy Norms for U.S. Adults (2016) Stratified by Gender, Education, Age, and Race

				Non		Under	Over		Other
Risk Literacy Test Score	General	Female	Male	College	College	55	55	White	Race
0	8.85	10.56	7.36	9.92	8.05	9.18	8.16	7.76	11.63
1	24.92	27.46	22.70	34.35	17.82	27.05	20.41	18.72	40.70
2	42.95	47.18	39.26	55.73	33.33	44.93	38.78	36.07	60.47
3	61.64	68.31	55.83	80.92	47.13	63.29	58.16	55.71	76.74
4	75.41	80.99	70.55	90.08	64.37	75.85	74.49	71.23	86.05
5	85.90	90.14	82.21	96.18	78.16	85.99	85.71	84.02	90.70
6	95.74	97.18	94.48	100	92.53	96.14	94.90	94.98	97.67
7	100	100	100	100	100	100	100	100	100

Note. No participant in the norming group's non-college subsample scored Risk Literacy = 7.

Table 2.6

Schwartz Test Norms for U.S. adults (2016) Stratified by Gender, Education, Age, and Race

				Non		Under	Over		Other
Schwartz Score	General	Female	Male	College	College	55	55	White	Race
0	15.41	14.79	15.95	25.19	8.05	14.49	17.35	11.42	25.58
1	44.26	47.89	41.10	62.60	30.46	47.83	36.73	38.36	59.30
2	75.08	83.10	68.10	92.37	62.07	73.43	78.57	69.86	88.37
3	100	100	100	100	100	100	100	100	100

Table 2.7

Full BNT Norms for U.S. adults (2016) Stratified by Gender, Education, Age, and Race

Full BNT Score	General	Female	Male	Non College	College	Under 55	Over 55	White	Other Race
0	45.90	51.41	41.10	60.31	35.06	43.96	50.00	38.81	63.95
1	70.49	78.17	63.80	84.73	59.77	70.53	70.41	67.12	79.07
2	85.57	91.55	80.37	96.95	77.01	84.54	87.76	82.65	93.02
3	93.44	97.18	90.18	97.71	90.23	92.27	95.92	92.69	95.35
4	100	100	100	100	100	100	100	100	100

Table 2.8

Adaptive BNT Norms for U.S. adults (2016) Stratified by Gender, Education, Age, and Race

Adaptive BNT Score	General	Female	Male	Non College	College	Under 55	Over 55	White	Other Race
1	52.13	59.86	45.40	68.70	39.66	50.24	56.12	46.12	67.44
2	72.46	76.06	69.33	87.02	61.49	71.01	75.51	68.95	81.40
3	84.92	91.55	79.14	95.42	77.01	83.57	87.76	81.74	93.02
4	100	100	100	100	100	100	100	100	100

Table 2.9

Adaptive BNT-S Norms for U.S. adults (2016) Stratified by Gender, Education, Age, and Race

Adaptive BNT-S Score	General	Female	Male	Non College	College	Under 55	Over 55	White	Other Race
0	15.41	14.79	15.95	25.19	8.05	14.49	17.35	11.42	25.58
1	44.26	47.89	41.10	62.60	30.46	47.83	36.73	38.36	59.30
2	60.66	68.31	53.99	80.15	45.98	59.90	62.24	54.34	76.74
3	70.16	77.46	63.80	90.08	55.17	67.63	75.51	65.30	82.56
4	81.31	88.03	75.46	95.42	70.69	80.19	83.67	77.17	91.86
5	85.57	91.55	80.37	96.18	77.59	84.54	87.76	82.19	94.19
6	90.49	94.37	87.12	98.47	84.48	88.41	94.90	88.13	96.51
7	100	100	100	100	100	100	100	100	100

Numeracy & Risk Literacy Norms by Quartiles

Following tradition from the Berlin Numeracy Test (Cokely et al., 2012), general numeracy and risk literacy quartile levels are presented. These quartiles provide less precise information, but support robust interpretability given that the quartiles between the two tests (BNT-S and Risk Literacy Test) are roughly equal (Table 2.10). For example, while a user of these norms should not venture to meaningfully interpret differences between an estimate of the 24th percentile versus the 26th percentile, it would be appropriate (i.e., reasonable, robust) to interpret differences between the first quartile and second quartile. Finally, given that the largest demographic difference was between college and non-college educated samples, distinct quartile levels for these two groups are provided.

Table 2.10

BNT-S and Risk Literacy Quartiles

Full BNT-S Score	General BNT-S Levels	Non-College BNT-S Levels	College BNT-S Levels	Risk Literacy Score	General Risk Literacy Levels	Non-College Risk Literacy Levels	College Risk Literacy Levels
0	25%	25%	25%	0	25%	25%	25%
1		50%		1		50%	
2	50%	75%	50%	2	50%	75%	50%
3	75%	< 99.9%		3	75%	< 99.9%	
4			75%	4			75%
5	< 99.9%			5	< 99.9%		
6			< 99.9%	6			< 99.9%
7				7			

Analysis 2: Subgroup Measurement Equivalence Testing

One challenge that often arises when developing norms for cognitive abilities and psychological measurement (e.g., intelligence tests), is the interpretation of these norms without measurement equivalence. Said differently, mean differences between groups in cognitive abilities may demonstrate differences in achievement, or may suggest that there is measurement nonequivalence (i.e., that the test functions differently for one group than for another). Most measures of intelligence do not pass the strictest forms of measurement equivalence (Bowden et al., 2008; Daseking et al., 2017; Pezzuti et al., 2020; Wicherts, 2006), which has necessitated a need for different norms for different subsamples (e.g., men vs. women). This is because when there is measurement nonequivalence, the two groups cannot be rated on the same scale. It remains an open question whether measures of statistical numeracy and risk literacy will meet the measurement equivalence standards achieved by intelligence (i.e., strong measurement invariance).

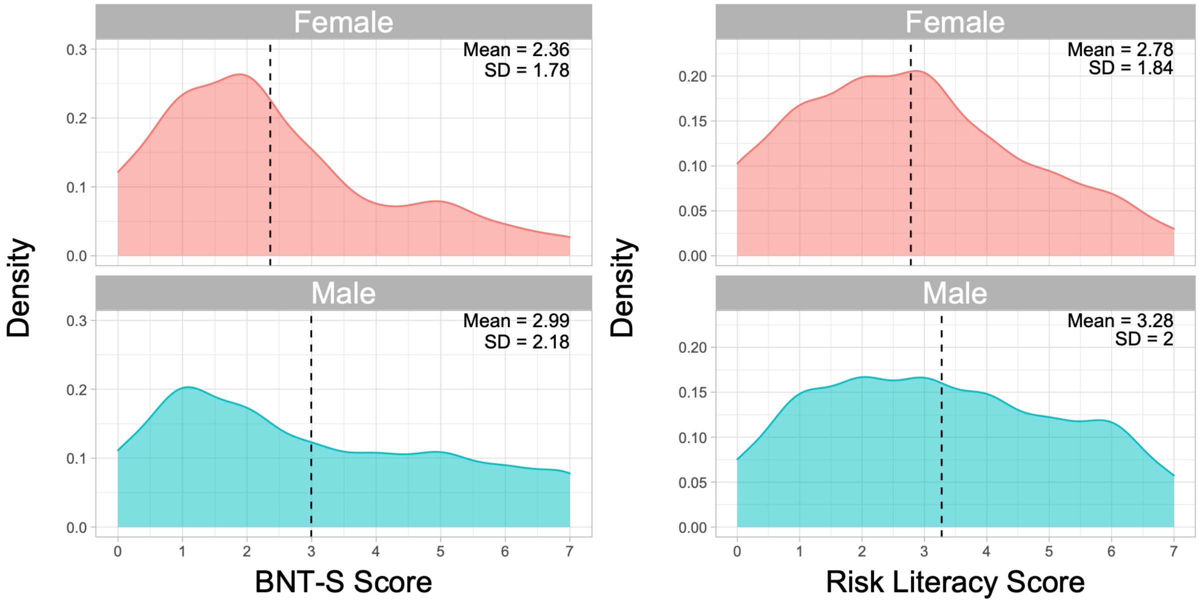
In the current study I focus on gender differences in measurement equivalence, in part because gender differences are often reported in numeracy and mathematical abilities (e.g., Galesic & Garcia-Retamero, 2010; Liu & Wilson, 2009; Pachur & Galesic, 2013). While this may be due to differences in achievement, reported differences in math anxiety and stereotype threat can exacerbate the effect (Betz, 1978; Brown & Josephs, 1999). Moreover, because the sample was relatively small (to account for the relatively long assessment participants engaged in), I expected to have sufficiently large groups and relatively equal subsamples for gender, but did not anticipate large enough samples for other subsamples (e.g., race, age).

First, to replicate previous findings, I tested mean differences for both the BNT-S and the Risk Literacy Test. As seen in Figure 2.5, men (M = 2.99, SD = 2.18) outperformed women

(M = 2.36, SD = 1.78) on the BNT-S, *t*(304) = -2.8, *p* < .05. Men (M = 3.28, SD = 2.0) also scored slightly higher than women (M = 2.78, SD = 1.84) on the Risk Literacy Test, *t*(304) = -2.25, *p* < .05. Given this gender difference, differential item functioning and measurement invariance in the BNT-S and the Risk Literacy Test are next examined. These procedures provide information on the extent to which the test measures the same construct for different groups (e.g., men vs. women), allowing for better interpretation of mean differences – e.g., are the mean differences due to a main effect, or are they artificial or substantively misleading?

Figure 2.5

Distribution by Gender (a) BNT-S (b) Risk Literacy



Berlin Numeracy Test – Schwartz

Differential Item Functioning. To test differential item functioning, I first confirmed that each measure fits a unidimensional structure (See Table 2.2 for item factor loadings). Then, I assessed the psychometric difficulty and discriminability using the ltm package in R (Rizopoulos, 2015), by fitting three two-parameter logistic models on (i) the full sample, (ii) males, and (iii) females (See Figure 2.6). I then used the difR package in R (Magis et al., 2020), which supplies a generic function (dichoDif) with nine different DIF detection methods. Here, I focused on the IRT-based methods, namely (i) Lord's chi-square test (Lord, 1980), (ii) Raju's area method (estimated based on a 2PL model with the z-statistic based on the unsigned area; see Raju, 1990; Raju et al., 1995), and the (iii) logistic regression technique, which tested for both uniform and nonuniform DIF (Swaminathan & Rogers, 1990).

Results suggested there was no gender-based DIF present in the BNT-S (See Table 2.11). First, Lord's chi-square method demonstrated that there was no significant difference between the item parameters in the two-parameter logistic (2PL) model. Second, the area between item response functions were not significantly different from zero, which implied there was no DIF present, according to the Raju area method. Third, the logistic regression method tested both uniform and nonuniform DIF, and following Nagelkerke's R^2 effect size guidelines from Zumbo & Thomas (1997) and Jodoin & Gierl (2001), all effect sizes were below the threshold for “negligible” effects – which suggested no DIF.

Figure 2.6

BNT-S 2PL Item Response Theory, by Gender

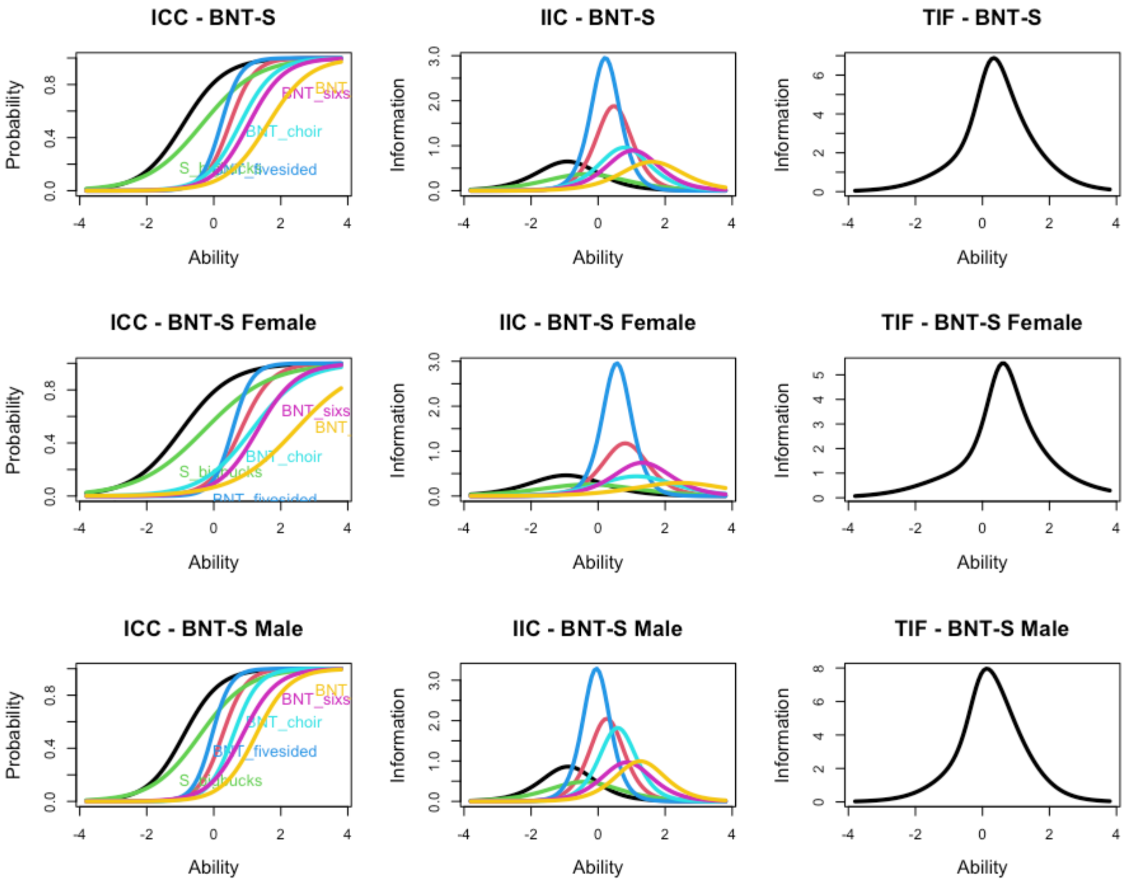


Table 2.11

BNT-S Differential Item Functioning

	Lord	Raju	Logistic
S_cointoss	0.08	0.21	0.53
S_acme	0.19	0.41	0.28
S_bigbucks	0.08	0.23	1.10
BNT_fivesided	2.36	1.54	5.90
BNT_choir	1.70	-1.05	3.38
BNT_sixsided	0.89	0.95	0.31
BNT_mushroom	0.32	-0.38	1.45

Note. There are no significant parameters, which suggests no DIF.

Measurement Invariance. Next, to assess Measurement Invariance, a series of Confirmatory Factor Analysis models were run where parameters were sequentially restricted. The first model tested, configural invariance, simply imposed the same factor structure on both groups (i.e., on both men and women; Horn & McArdle, 1992). This is a necessary but not sufficient condition for MI (Bauer, 2017). Then, in Model 2 weak invariance is tested, which constrained the factor loadings to be equal across groups. In Model 3, strong invariance constrained both the factor loadings and the intercepts to be equal across groups (Meredith, 1993). Though group differences are often reported when only strong measurement invariance is met (see Vandenberg & Lance, 2000), recent research suggests that it is necessary to also hold residual variances equal (DeShon, 2004; Lubke & Dolan, 2003; Wicherts & Dolan, 2010). As such, in Model 4 strict invariance is tested. This model constrained factor loadings, intercepts, and residual variances (See Table 2.12). Analyses were completed using the lavaan package in R (Rosseel et al., 2017).

Following conventions for confirmatory factor analysis, the model fit can be assessed using several different indices. Common fit indices include the Root Mean Square Error of Approximation (RMSEA), the Comparative Fit Index (CFI) and the Tucker Lewis Index (TLI). Standard convention suggests that a CFI and TLI greater than 0.95 and a RMSEA less than 0.05 suggest *good* fit, whereas a RMSEA less than 0.08 suggests *moderate* fit (Hu & Bentler, 1999; Kline, 2015).

CFAs were conducted separately for group 1, males (χ^2=33.22; p<.05, CFI=.935; TLI= 0.90; RMSEA=.092), and group 2, females (χ^2=21.67; p>.05, CFI=.943; TLI= 0.92; RMSEA=.062). Next, measurement invariance is tested (see Table 2.12 for the fit indices). As models were sequentially restricted, model fit did not get worse, as indicated by the nonsignificant **$\Delta\chi^2$**. The fourth model (Strict Invariance) had the lowest AIC value and therefore suggested the best trade-off between model fit and model complexity. The other fit indices of the strict model also indicated good fit (χ^2=72.2; *p*<.05, CFI=.94; TLI= 0.95; RMSEA=.06). This suggests that for the BNT-S, measurement invariance held, and it is reasonable to compare means between groups (gender).

Table 2.12

BNT-S Gender Measurement Invariance

Model	Restrictions	χ^2	df	$\Delta\chi^2$	RMSEA	AIC	CFI
Baseline Models							
Full Sample	-	30.11**	14	-	.061	2159.9	.963
Male	-	33.22**	14	-	.092	1182.68	.935
Female	-	21.67	14	-	.062	972.02	.943
Invariance Models							
Configural	-	54.896**	28	-	.079	2182.69	.938
Weak	loadings	59.296**	34	4.40	.070	2175.09	.941
Strong	loadings, intercepts	65.84**	40	6.55	.065	2169.64	.940
Strict	loadings, intercepts, residual variances	72.2*	47	6.36	.059	2161.99	.941

Risk Literacy

Differential Item Functioning. With the same procedure that was used for the BNT-S, differential item functioning between genders in the Risk Literacy scale was tested. First, the measure was confirmed to fit a unidimensional structure (See Table 2.2 for item factor loadings). Then, the psychometric difficulty and discriminability were assessed, by fitting three two-parameter logistic models on (i) the full sample, (ii) males, and (iii) females (See Figure 2.7). Finally, three DIF detection methods were implemented: Lord's chi-square method, Raju's area method, and the logistic regression method (Table 2.13). When comparing men and women on the Risk Literacy Test, no DIF was detected.

Figure 2.7

Risk Literacy 2PL Item Response Theory, by Gender

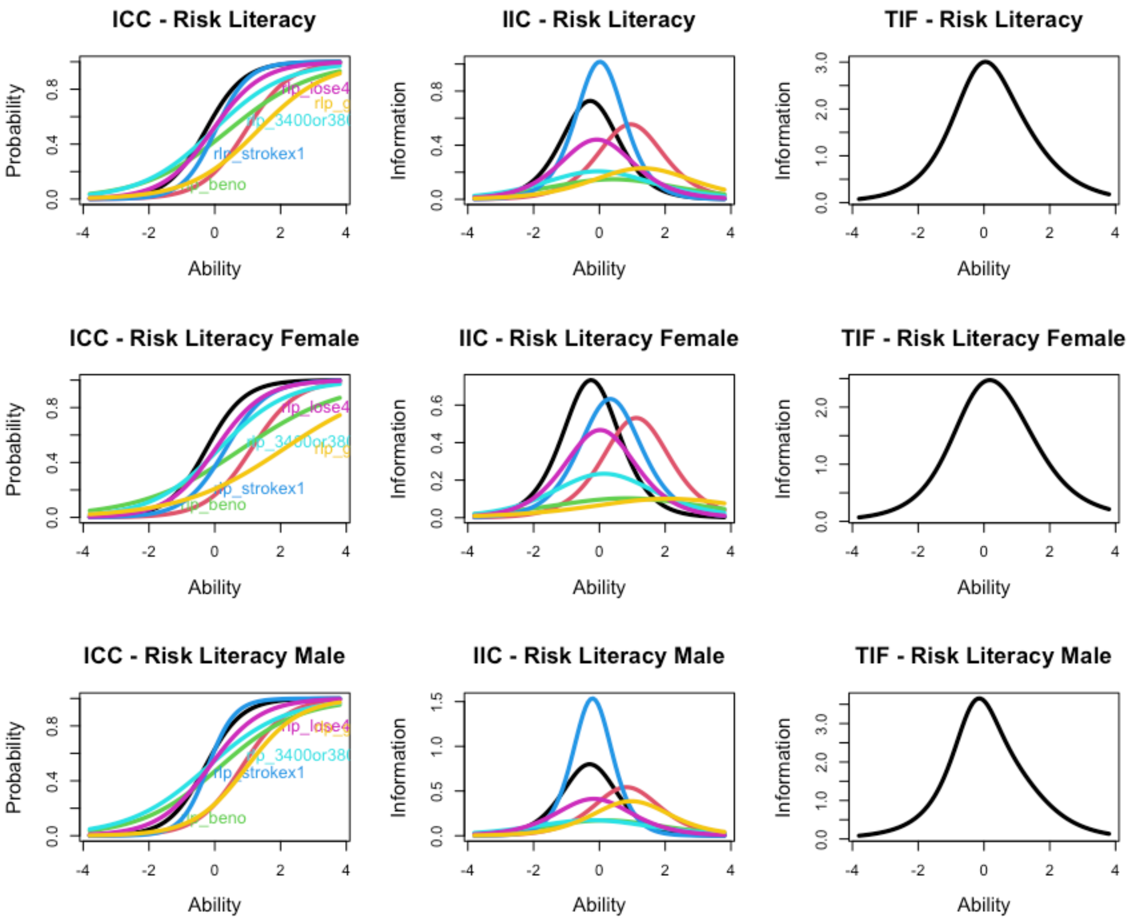


Table 2.13

Risk Literacy Differential Item Functioning

	Lord	Raju	Logistic
rlp_min	1.57	1.13	1.89
rlp_percentage	1.63	1.23	0.06
rlp_beno	0.54	0.44	0.38
rlp_strokex1	1.15	-1.07	4.01
rlp_3400or3800	1.45	1.24	1.51
rlp_lose400	1.31	1.21	1.48
rlp_gain100	0.32	-0.38	1.33

Note. There are no significant parameters, which suggests no DIF

Measurement Invariance. CFAs were first conducted separately for group 1, males (χ^2=18.85; *p*>.05, CFI=.96; TLI= .94; RMSEA=.046), and group 2, females (χ^2=24.39; *p*<.05, CFI=.875; TLI= .81; RMSEA=.072). As models were sequentially restricted, model fit did not get worse, as indicated by the nonsignificant **Δ**χ^2 (see Table 2.14 for fit indices). The fourth model (Strict) had the lowest AIC value and therefore suggests the best trade-off between model fit and model complexity. The other fit indices of the strict model also indicated good fit (χ^2=53.2; *p*>.05, CFI=.97; TLI= 0.97; RMSEA=.03). This suggested that for the Risk Literacy Test, measurement invariance held, and it is reasonable to compare means between groups.

Table 2.14

Risk Literacy Gender Measurement Invariance

Model	Restrictions	χ^2	df	$\Delta\chi^2$	RMSEA	AIC	CFI
Baseline Models							
Full Sample	-	33.98**	14	-	.068	2741.20	.912
Male	-	18.85	14	-	.046	1477.49	.963
Female	-	24.39*	14	-	.072	1275.75	.875
Invariance Models							
Configural	-	43.24*	28	-	.06	2781.25	.93
Weak	loadings	46.41	34	3.17	.049	2772.42	.942
Strong	loadings, intercepts	51.56	40	5.15	.044	2765.57	.946
Strict	loadings, intercepts, residual variances	53.19	47	1.63	.029	2753.20	.971

Overall, results suggest that the BNT-S and Risk Literacy Test not only meet but exceed typical standards of measurement equivalence (i.e., strict invariance). This supports the interpretation of the norms presented in Analysis 1, because it suggests that differences are likely due to differences in achievement, as opposed to unequal features of the assessment battery (i.e., measurement nonequivalence). Had measurement invariance failed, there would be important implications regarding the fairness and interpretation of these tests.

Analysis 3: Decision Vulnerability Metrics

Here, *Decision Vulnerability* was defined as the estimated proportion of people (adults) in a well-defined (sub)population who are likely to misinterpret, misunderstand, or otherwise experience cognitive errors and biases on a specified task (e.g., making a decision, interpreting a risk communication). Based on numeracy and risk literacy norms, I sought to develop a method to predict *risk literacy difficulty levels*, or the estimated difficulty associated with the evaluation and understanding of a risk or task. To do so, I estimated a *numeracy skill threshold*, a minimum raw score on the standardized BNT-S, associated with achievement of at least 50% accuracy on the specified task. This numeracy skill threshold indicates the numeracy level at which a typical individual is more likely than not to accurately interpret a risk communication independently.

The first step to developing a methodology for decision vulnerability analysis, was to predict risk literacy cumulative percentile rank from numeracy score, while controlling for major demographic groups (i.e., dichotomized age, race, education, and gender). Risk Literacy prediction equations were then developed using multiple linear regression. Five separate regression equations, one for each of the five different numeracy test versions were developed (i.e., Full BNT-S, Schwartz, Full BNT, Adaptive BNT and Adaptive BNT-S; See Table 2.15 for Eq. 3.1-3.5). In each regression equation, the numeracy score and the four demographic groups were treated as independent variables. The dependent variable was the cumulative percentile rank on the raw score of the risk literacy test. Though I demonstrated measurement equivalence for gender and had reason to believe the same would follow for the other demographics (i.e., education, race, age), these factors are included in the model to account for main effects (i.e., achievement differences in numeracy between men and women).

Table 2.15

Risk Literacy Difficulty Level Prediction Equations

Test Form	Equation	
Full BNT-S	$y_{rl} = .041x_{age} + .044x_{race} + .035x_{education} + .017x_{gender} + .083x_{BNT-S} + 0.274$	(3.1)
Schwartz	$y_{rl} = .033x_{age} + .053x_{race} + .053x_{education} + .039x_{gender} + .138x_{Schwartz} + 0.244$	(3.2)
Full BNT	$y_{rl} = .047x_{age} + .063x_{race} + .07x_{education} + .018x_{gender} + .12x_{BNT} + 0.336$	(3.3)
Adaptive BNT	$y_{rl} = .052x_{age} + .064x_{race} + .068x_{education} + .026x_{gender} + .129x_{BNTadap} + 0.21$	(3.4)
Adaptive BNT-S	$y_{rl} = .044x_{age} + .048x_{race} + .035x_{education} + .023x_{gender} + .073x_{BNTSadap} + 0.305$	(3.5)

The risk literacy prediction equations estimated the risk literacy cumulative percentile rank among the general U.S. population. The next step was to select a method to estimate risk literacy percentile ranks for specific subsamples (e.g., women, highly educated people, etc.). In an effort to develop a simple, yet robust method that would provide a starting point (e.g., proof of concept) and would be easy to communicate to and understood by a wide range of researchers, I again developed linear regression models. In this case, the independent variable was the risk literacy cumulative percentile rank (as estimated in Eq. 3.1-3.5). Based on the demographic group of interest, Equations 3.6 to 3.13 were developed to estimate the risk literacy cumulative percentile rank specific to a subgroup (i.e., the decision vulnerability benchmarks; the dependent variable).

Table 2.16

Decision Vulnerability Translation Equations for Subsamples

Demographic Subgroup	Equations	
Female	$y_{RLfemale} = 1.019x_{RLgeneral} + 0.03$	(3.6)
Male	$y_{RLmale} = .999x_{RLgeneral} - 0.033$	(3.7)
Non-College	$y_{RLnoncollege} = 1.12x_{RLgeneral} + 0.64$	(3.8)
College	$y_{RLcollege} = 1.012x_{RLgeneral} - 0.085$	(3.9)
Less 55 Years	$y_{RLunder55} = .983x_{RLgeneral} + 0.021$	(3.10)
55+ Years	$y_{RLover55} = 1.037x_{RLgeneral} - 0.045$	(3.11)
Other Race	$y_{RLotherrace} = .936x_{RLgeneral} + 0.15$	(3.12)
White	$y_{RLwhite} = 1.044x_{RLgeneral} - 0.067$	(3.13)

Decision Vulnerability Analysis Protocol

A researcher interested in using the decision vulnerability analysis method can follow the protocol provided in Table 2.17 to estimate three metrics (i) a *numeracy skill threshold* (Step 3), (ii) the *risk literacy difficulty level* in the general population (Step 4), and (iii) *decision vulnerability benchmarks* for specific subgroups (Step 5). To investigate the decision vulnerability associated with a new risk communication (or artifact), a researcher could use data they have collected and simple linear regressions to generate these metrics. First, using their data and a simple linear regression, where accuracy on the new risk communication serves as the dependent variable, and numeracy as the independent variable, a *numeracy skill threshold* is generated (i.e., the estimated minimum raw score associated with achievement of at least 50% accuracy). Then, using the appropriate risk literacy prediction equation provided (Eq. 3.1

to 3.5), they can next estimate the *risk literacy difficulty level* (i.e., the percentile rank in the general population; Step 4). Finally, to the extent a particular demographic subgroup is of interest (e.g., women, highly educated individuals), they can use Eq. 3.6 to 3.13 to estimate the *decision vulnerability benchmark* for specific subgroups (Step 5).

Table 2.17

General Protocol for a Decision Vulnerability Analysis

Step	
1	Collect data assessing the new risk communication (e.g., comprehension). Include a measure of statistical numeracy (e.g., BNT-S, Adaptive BNT, etc.), as well as any relevant demographics.
2	Using data, predict (*y*) accuracy on the given artifact (i.e., risk communication), with (*x*) the selected numeracy test, as well as necessary demographic covariates (e.g., gender, age, race, education).
3	Solve for the numeracy score (*x*) at the point where a participant has a 50% chance of answering the artifact correctly (i.e., $y = 0.5$). *
4	Select the relevant equation from Eq. 3.1-3.5. This will depend on the type of statistical numeracy test used during data collection (Step 1). Using the selected equation, solve for the general risk literacy level (*y*) at the numeracy threshold score found in Step 2 (i.e., *x* = numeracy score at 50% accuracy on the artifact).
5	Use Eq. 3.6-3.13 to translate the general risk literacy level to the decision vulnerability benchmark for the desired subsample (e.g., men, college educated)

Note. *When the artifact includes more than one item, the 50% accuracy mark will be *k*/2, where *k* = the number of items.

Demonstration and Validation of Decision Vulnerability Analysis

Utilizing additional statistical numeracy and risk literacy items in the survey, I selected five items to serve as example artifacts (i.e., probabilistic problems), as a test of the decision vulnerability analysis methodology. These items all exhibited relatively high discriminability (i.e., > 1) and spanned the range of difficulty (e.g., roughly -1, 0 and +1 theta, θ; Table 2.18).

Taking the item *prob_burn* as the first artifact of interest, I next followed each step, as described in the general protocol (See Table 2.17). First, I identified the data set for analysis and selected the anchor numeracy test (i.e., full BNT-S). Second, I regressed accuracy on the artifact (i.e., *prob_burn*) on the numeracy scale (i.e., BNT-S).[2] Third, I solved for x (BNT-S score) at a 50% accuracy level (i.e., $y = 0.5$). This provides the *numeracy skill threshold*, which for *prob_burn* is 3.45. Fourth, because the data included the full-scale BNT-S, I selected Eq. 3.1, and solved for the risk literacy difficulty level at the BNT-S numeracy skill threshold (BNT-S = 3.45). The estimated *risk literacy difficulty level* for *prob_burn* was 0.628. This result suggests that an estimated 62.8% of the general population is likely to misinterpret or inaccurately answer this artifact. Finally, to translate the general *risk literacy difficulty level* (62.8%) to *decision vulnerability benchmarks* for different subgroups, I used Eq. 3.6 to 3.13. Results are provided for *prob_burn* in Table 2.19.

[2] The current analysis used linear multiple regression; however logistic analysis may also be appropriate to consider in future analyses.

Table 2.18

Single Item Examples

Item	Item Difficulty	Item Dscrmn	BNT-S Score at 50% Accuracy*	General Risk Literacy Difficulty Level
oper_fieldtrip. A school is having a field trip and many parents are going on the field trip with the children. What is the child to parent ratio if there are 20 children and 5 parents?	-1.20	2.20	-2.77	0.114
prob_dice5. People often roll dice when playing games. Most dice have 6 sides and each side has a different number on it ranging from 1-6. If you rolled one of the dice, on average what is the probability that it would land on 5?	-0.79	2.29	0.38	0.373
prob_burn. Imagine that the probability of a child getting sunburned at the beach is 65% while the probability of an adult getting sunburned at the beach is 15%. If there were 300 people who spent a day at the beach, and 60% of the people were children, how many people are likely to get a sunburn?	0.11	1.87	3.45	0.628
oper_goods. Imagine that goods imported into a country increased by 40% and exports decreased by 30% during a certain year. What was the ratio of imports to exports at the end of the year compared to the beginning of the year?	0.96	1.41	5.25	0.777
prob_diceeven. Imagine that you are throwing 2 regular 6-sided dice up in the air. If each side has a different number on it ranging from 1-6, on average what is the probability that both of them land on even numbers?	1.10	1.11	4.98	0.754

Note. *For single item artifacts, the 50% accuracy midpoint is $y = 0.5$. Full multiple choice item text is presented in Appendix A.

Table 2.19

Decision Vulnerability Results for prob_burn

Group	Predicted Decision Vulnerability Benchmark	Proportion Correct, Weighted Sample	Predicted Levels	True Score Levels	Deviation = Predicted Value - Weighted Proportion	> 10% Deviation
General	0.63	0.57	III	II	0.06	F
Female	0.67	0.64	III	III	0.03	F
Male	0.59	0.50	II	II	0.09	F
Non-College	0.77	0.71	III	III	0.06	F
College	0.55	0.47	II	II	0.08	F
Under 55	0.64	0.58	III	II	0.06	F
Over 55	0.61	0.56	II	II	0.04	F
Other Race	0.74	0.75	III	III	0.01	F
White	0.59	0.48	II	II	0.11	T

Note. Proportion correct refers to the proportion of the sample that answered *prob_burn* correctly. This proportion was estimated while accounting for sample weights to estimate the U.S. Census (2016).

To assess the robustness and accuracy of model predictions as part of a validity assessment for the estimated decision vulnerability predictions, I then calculated the *actual* proportion of participants in the sample who correctly answered each artifact (e.g., *prob_burn*, etc.). I then used two methods to compare the predicted value to the true value. First, I identified the risk literacy quartile levels provided with the norms in Study 1A (i.e., Table 2.10), in order to assign broad "levels" to the true and predicted decision vulnerability benchmarks. Following analyses from the Cokely et al. (2012) paper, I identified the four levels around the quartile mark, Level I was set at ranks between 0% and 37.5%, Level II at 37.5-62.5%, Level III at 62.5-87.5% and Level IV set at risk literacy levels above 87.5%. As a test of the validity and robustness of model predictive accuracy, I next compared the true score levels to the predicted levels. As can be seen in Table 2.19, only two predictions were miscategorized, namely the general risk literacy difficulty level, and the decision vulnerability benchmark for the under 55 group. Further comparison of all five of the single-item artifacts, revealed that 38 of the 45 predicted levels matched the true value levels (84.4%; See Appendix B).

A second validity test next compared predicted and actual results by calculating the difference between the predicted value and the true weighted proportion. The total number of decision vulnerability estimates that deviated from the true value by more than 10% were counted. For *prob_burn*, only one predicted value deviated by more than 10%, namely that for the white race group. Across the five artifacts, results revealed relatively high accuracy with only 4 of the 45 predicted values deviating by more than 10% (i.e., 9% of tested observations).

Following this analysis which focused on only single item artifacts, I next conducted a similar decision vulnerability analysis, but this time applied to multi-item/multi-criteria artifacts. Specifically, I considered the sum of the five single-item artifacts (Table 2.20). The analysis method and results for multi-item artifacts followed the same form as that used for single item artifacts (Table 2.17) The primary difference was that for multiple item artifacts, the 50% accuracy point is the number of items, *k*, divided by 2. See Appendix B for complete decision vulnerability subgroup metrics.

Table 2.20

Multi-Item Artifact Example

Artifact Set	BNT-S Score at 50% Accuracy*	General Risk Literacy Difficulty Level	Proportion Correct in Weighted Sample
prob_burn + prob_dice5 + oper_goods	3.16	0.60	0.54
prob_burn + prob_dice5 + oper_goods + oper_fieldtrip + prob_diceeven	2.88	0.58	0.52

To further validate, I next examined the final set of all seven artifacts (i.e., five single items as well as two combined multi-item artifacts). Fifty-five of the 63 predicted risk literacy levels matched the true weighted proportion (87.3%). For the second validation method, only 4 of the 63 predicted values were off by more than 10% (i.e., 93.7% accuracy). With a more stringent cutoff point that considered any deviations greater than 5%, the categorization error rate increased to 42 out of 63, indicating that 66.7% of the predicted values were off by more than 5%. Regardless of the margin of error, nearly all of the predicted estimates were found to be conservative, such that predictions were somewhat likely to (slightly) overestimate the proportion of people expected to misunderstand. While bias in prediction is a limitation, to the extent this pattern generally holds, it appears both modest in magnitude and better than the alternative (e.g., it seems generally riskier to *underestimate* how many people will understand than to *overestimate*). Said differently, given potential stakes, it is likely better to false alarm (e.g., it might be too hard) than to miss (e.g., everyone will understand).

While generally promising, some initial evidence suggests that the discriminability of the artifact (as estimated by Item Response Theory) is likely a factor that will constrain the quality of estimates for risk literacy difficulty level, particularly when there is only one criterion (e.g., as compared to averaging over several criteria). For example, this limitation is tested using two risky gamble items (*rlp_400or100* and *rlp_1000or2400;* see Table 2.21). When the decision vulnerability method is tested using an item with low discriminability (*rlp_400or100* has difficulty = 0.62 and discriminability = 0.56), the risk literacy prediction equation suggested that the BNT-S 50% accuracy level (i.e., the numeracy skill threshold) was 14.13 – more than double the standard BNT-S scale. This further translated to a risk literacy difficulty level of 1.51. The interpretation of this value seems nearly meaningless, as it suggested over 150% of people are likely to inaccurately answer this artifact (gamble). However, if this result is compared to another risk literacy gamble with higher discriminability (*rlp_1000or2400* has difficulty = 0.86 and discriminability = 1.02), the estimated numeracy skill threshold was only 5.36 (well within the bounds of the scale range), and the risk literacy difficulty level was 0.786. To the extent this pattern holds, future researchers should note that the discriminability or reliability of an item could (and likely often does) influence the accuracy of the numeracy skill threshold and decision vulnerability benchmark estimates (e.g., perhaps especially when the item discriminability < 1).

Table 2.21

Examples with Varied Discrimination Values

Item	Item Difficulty	Item Discrimination	BNT-S Score at 50% Accuracy	General Risk Literacy Difficulty Level
rlp_400or100. Imagine you were offered the following choices. Which option would you select? - $400 now - $100 every year for 10 years*	0.62	0.56	14.13	1.51
rlp_1000or2400. Imagine you were offered the following choices. Which option would you select? - $1000 in six months - $2400 in two years*	0.86	1.02	5.36	0.786

Analysis 4: Practical Out-of-Sample Application for Decision Vulnerability

The decision vulnerability approach could support research on extreme weather risk communication, especially those developed and distributed by NOAA and the NHC. Given that understanding risk is an important step in informed decision making and taking protective action, by providing mechanisms to predict the risk literacy difficulty level, as well as estimating the proportion of people likely to misunderstand, the decision vulnerability approach could provide useful metrics to support the weather service's mission to protect lives and property.

Using a survey of members of the public from across the U.S. (N = 1,000), with an oversample from coastal regions that are more likely to experience hurricanes (N = 2,000), I conducted an out-of-sample application of the decision vulnerability approach. Participants were presented with probabilistic forecast products issued by the National Weather Service (National Hurricane Center) and were then asked a series of questions regarding their familiarity, use, and comprehension of the products. For each of the products (i.e., windspeed probabilities and storm surge inundation graphic; Figure 2.8), participants were asked to interpret the meaning of the product, using a multiple-choice format, where options varied in terms of both the spatial and temporal probabilities (e.g., There is a 10% chance that Location A will get more than 9 ft of storm surge above ground level).

Following the decision vulnerability approach, I provided a first pass out-of-sample validation test, using both risk communication artifacts (i.e., the windspeed probabilities and storm surge inundation graphic). This data (N = 2,747) included the Adaptive BNT-S, so the decision vulnerability analyses utilized Eq. 3.5. The numeracy skill threshold and general risk literacy difficulty level were estimated (See Table 2.22). The estimated numeracy skill threshold for the windspeed probabilities graphic was 3.9, and the predicted risk literacy difficulty level was 0.667. This suggests that roughly 66.7% of the general population are likely to misunderstand this risk communication or are otherwise likely to exhibit bias. Similarly, the numeracy skill threshold for the storm surge inundation graphic was 2.2, and the predicted risk literacy difficulty level was 0.54, suggesting 54% of adults in the general public are likely to misinterpret the storm surge graphic.

Figure 2.8

(a) Hurricane Windspeed Probabilities and (b) Storm Surge Inundation

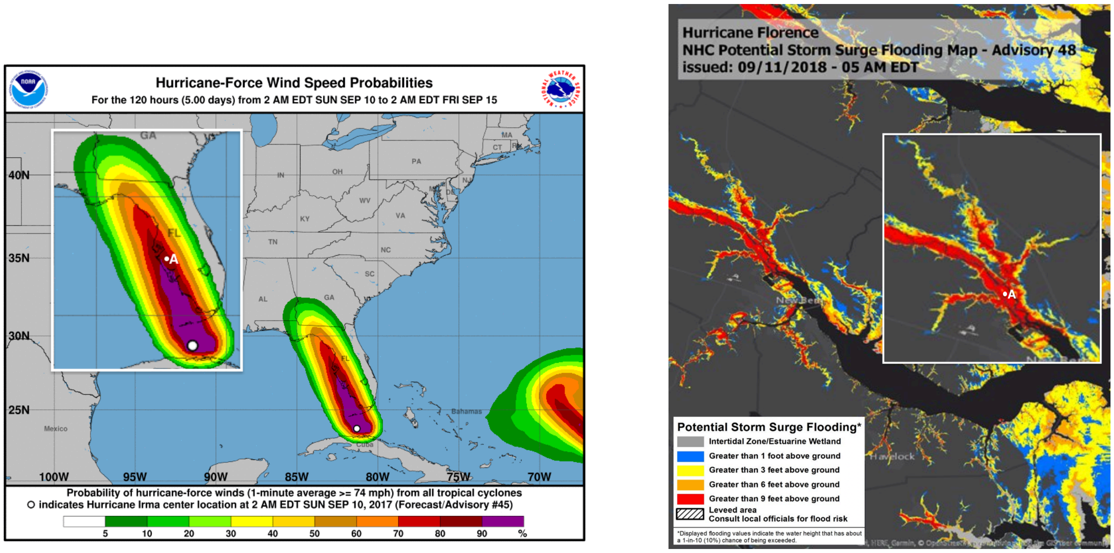

Hurricane-Force Wind Speed Probabilities
For the 120 hours (5.00 days) from 2 AM EDT SUN SEP 10 to 2 AM EDT FRI SEP 15
GA
FL
A
IN
OH
PA
MA
CT
RI
NJ
MD
DE
WV
VA
KY
TN
NC
SC
AL
GA
FL
Bahamas
Cuba
Mexico
40N
35N
30N
25N
100W
95W
90W
85W
80W
75W
70W
Probability of hurricane-force winds (1-minute average >= 74 mph) from all tropical cyclones
○ indicates Hurricane Irma center location at 2 AM EDT SUN SEP 10, 2017 (Forecast/Advisory #45)
5
10
20
30
40
50
60
70
80
90
%
Hurricane Florence
NHC Potential Storm Surge Flooding Map - Advisory 48
issued: 09/11/2018 - 05 AM EDT
A
New Bern
Havelock
Potential Storm Surge Flooding*
Intertidal Zone/Estuarine Wetland
Greater than 1 foot above ground
Greater than 3 feet above ground
Greater than 6 feet above ground
Greater than 9 feet above ground
Leveed area
Consult local officials for flood risk
*Displayed flooding values indicate the water height that has about a 1-in-10 (10%) chance of being exceeded.
HERE, Garmin, © OpenStreetMap contributors, and the GIS user community

Table 2.22

Hurricane Decision Vulnerability Analysis

	Full Data Set (N= 2,747)		Reduced Sample with Performance Validity (N=319)	
Item	Numeracy Skill Threshold	General Risk Literacy Difficulty Level	Numeracy Skill Threshold	General Risk Literacy Difficulty Level
Windspeed comprehension	3.90	0.67	1.94	0.52
Storm Surge comprehension	2.22	0.54	0.50	0.42
Windspeed comprehension + Storm Surge comprehension	2.97	0.60	1.10	0.46

While these analyses provided a first proof of concept in an out-of-sample and applied context, they also raise a few open questions for future research. For example, as is seen in Table 2.22, the validity of the results may depend on the reliability of the sample. The hurricane dataset included a performance validity item, intended to screen participants for careless responding. When the decision vulnerability approach is conducted using this reduced sample (i.e., only including participants who passed the performance validity item), there are marked differences in the numeracy skill threshold and the risk literacy difficulty level. For both the windspeed probabilities and the storm surge inundation graphic, the estimated risk literacy difficulty level decreased when participants who may not be paying attention were removed. This raised an important question to consider: are there mechanisms to control for or correct for careless responding, within the decision vulnerability framework? Future research will need to further investigate what other factors may influence the reliability and validity of the decision vulnerability metrics.

Chapter 3

General Discussion

Latent Trait Measurement Modeling and Norms. There is a great deal of research on numeracy and its measurement, including different components and numeric competencies. But what really *is* numeracy and how would we know? Outside of experimental and cognitive process tracing studies, a growing number of analyses suggest that numeracy may be very well characterized as unidimensional (i.e., fits a one-factor solution; see Allan, 2018 for an example). Leveraging modern quantitative techniques developed over nearly two centuries of research on psychological testing and measurement (e.g., Embretson & Reise, 2013; Holzinger & Swineford, 1939; Tukey, 1969), the findings presented in this report are novel in many ways. This report is the first to develop and distill precise measurement models of statistical numeracy and risk literacy, by linking the quantitatively estimated latent traits that are empirically grounded in cognitive theory (i.e., Skilled Decision Theory). Based on probabilistically representative testing of all published Berlin Numeracy Tests (e.g., BNT-S, BNT-Adaptive), and the new Risk Literacy Test, the current study is the first to establish norm-referenced test scores for the general adult population of the United States, as well as for key subgroups (e.g., gender, age, race, and education). Consistent with previous findings, results confirmed that among the general U.S. population a large proportion of adults have relatively low overall levels of numeracy and risk literacy skills. These scores appeared so low that it seems likely that they could qualify as functionally innumerate or risk illiterate, had they been associated with specific criterion-referenced indices (e.g., skill Level 1 as defined by NCES PIAAC, 2020), To further illustrate, the current study estimates that roughly 67% of all U.S. adults are likely to find it difficult to translate “1 in 1,000” into a percent (0.1%). In comparing demographic groups, the largest differences existed between college educated and non-college educated samples. Furthermore, the measurement models for the college educated sample demonstrated sensitivity across the range of scores, whereas for the non-college educated sample, there was a reduction in sensitivity among lower skill levels, such that roughly 75% of non-college educated adults are expected to answer less than or equal to 2 items correct on both the full Berlin Numeracy Test-Schwartz (7 items) and the Risk Literacy Test (7 items).

More specifically, measurement models were developed to link theory, tests, and norm-references for both numeracy and risk literacy via IRT latent trait modeling. To do this I (i) validated the Berlin Numeracy Test and (ii) developed an optimized Risk Literacy Test using IRT latent trait modeling. To do so, I confirmed the unidimensional structure of the Berlin Numeracy Test (e.g., all factor loadings > 0.3, and the proportion of variance explained by the unidimensional factor was 34%). The reliability and validity of the test was further evidenced by the high Cronbach's alpha (α = 0.77), and by high discriminability parameters across the range of difficulty, which is presented in the 2PL IRT latent trait measurement model (see Figure 2.6). Next, I followed a similar approach for the development and validation of a norm-referenced test of Risk Literacy. The measure of risk literacy included both paradigmatic risky choice tasks as well as ecological decisions (e.g., medical and financial choices). In validating this measure, analyses suggested the risk literacy skill was indeed reasonably well-characterized by a unidimensional latent trait (i.e., all factor loadings > 0.3, and the proportion of variance explained by the unidimensional factor was 23%) and appeared at least moderately reliable (α = 0.66). It is worth noting that this alpha is likely to provide a particularly conservative estimate, given that the test was designed to be both short and address multiple components of risk literacy (i.e., 7 items, including both ecological and paradigmatic choices). Nevertheless, Item

Response Theory analyses indicated that the items rather symmetrically spanned the difficulty range, while still exhibiting relatively high discriminability. Analyses also revealed a relatively high-test information function, which suggests the instrument is likely to be sensitive for both high-ability and low-ability test-takers (see Figure 2.7).

Measurement Equivalence and Norm Subgroups. The current study is also the first and only to test measurement equivalence on the popular Berlin Numeracy Test, and the new Risk Literacy Test, with its unique emphasis on practical, general decision making skills and biases (e.g., evaluating loans, risky prospects, selecting medical treatments). Analyses indicated that the BNT-S and the Risk Literacy Test passed some of the most rigorous standards of measurement equivalence (e.g., Lubke & Dolan, 2003; Mellenbergh, 1989; Wicherts & Dolan, 2010). This suggests that mean differences in test scores of men and women are not plausibly explained by differences in test bias (e.g., differences in math anxiety or stereotype threat that could differentially affect men and women). Even in the context of several other novel findings and analyses presented in this study, finding strict measurement equivalence is remarkable and potentially quite valuable, considering that even the most extensively validated and trusted tests of adult intelligence—such as the WAIS-IV (as commonly used for clinical, neuropsychological, industrial, and educational and other diagnostic purposes)—often fail to achieve strict invariance (Bowden et al., 2008; Daseking et al., 2017; Pezzuti et al., 2020; Wicherts, 2006). That is, although intelligence tests can and sometimes do achieve this standard, they often only meet lower standards of invariance (e.g., weak, strong), ultimately indicating the presence of bias (e.g., to some extent, men and women at the same latent trait of ability have different propensities to answer an item correctly, due to their group membership). In contrast, in our representative and demographically diverse sample of adult men and women from the United States, gender differences on the Berlin Numeracy Test and the new Risk Literacy Test are best explained as differences in overall skill achievement, independent of bias.

Decision Vulnerability Analyses and Metrics. Beyond the development of integrated measurement models, the creation of a new Risk Literacy test, estimation of nationally representative and subgroup test score norms, and tests of strict measurement invariance in response patterns for men and women, the current study also developed and presented a novel framework for estimating *decision vulnerability*. This framework includes a practical protocol that aims to keep complexity for researcher or end-users to a minimum, without making large simplicity/accuracy trade-offs. In other words, the norms enable meaningful interpretations and translations of test scores. As such, high-quality norms based on rigorous, integrated latent trait measurement models naturally lend themselves to many potentially valuable quantitative predictions, including diagnostic uses, the standardization of comparisons from independent studies, and a potential basis for the development of common benchmarks. Just as invariant measurement in physics facilitates discovery and scientific integration via standardization and unification of measures, so too can norms provide relatively invariant, robust, and common currencies for the systematization of theory and research. Regardless of whether these norms can more generally empower scientific research (e.g., via standardizing measurements and their interpretations), as part of the decision vulnerability framework, a simple protocol enables quantitatively robust predictions of decision making and risk metrics. More specifically, in Analysis 3 and 4, a method for *decision vulnerability analysis* is developed and tested. This provided a proof-of-concept process for:

(i) **Risk Literacy Difficulty Levels,** predicting the risk literacy difficulty level for a specified task (i.e., the estimated difficulty associated with understanding a risk communication represented as a cumulative distribution score, yet generally interpretable as akin to the overall ability (e.g., theta) level associated with some criterion.

(ii) **Decision Vulnerability Benchmarks**, estimating the proportion of people who are likely to misinterpret a risk or experience cognitive errors and biases during decision making, for the general adult population of the United States and for demographic subgroups (e.g., men, women).

(iii) **Numeracy Skill Thresholds,** estimating the minimum score on the BNT-S that is associated with at least 50% accuracy. This threshold indicates the numeracy level at which a typical individual is likely to accurately interpret a risk communication independently, or otherwise avoid cognitive errors and biases.

Although there is more work to be done, given the noted and seemingly impressive psychometric performance of the norms, latent trait models, and other results reviewed in chapter two, perhaps it is not surprising that over 90% of the decision vulnerability predictions fell within 10 percentile points of the true values in the model recovery and hold-out item analyses.

Decision Making Theories. *Dual Systems Theory* suggests the differences in decision making biases result from differences in largely stable and abiding cognitive capacities (i.e., intelligence; Frederick, 2005; Kahneman, 2003). As such, only *a small proportion of* individuals might ever be able to engage in superior and unbiased decision making. In contrast, *Skilled Decision Theory* suggests that individual differences in decision making quality are not (primarily) constrained by differences in intelligence or other working memory and attentional control capacities, but rather are primarily constrained by the kinds of skills and knowledge one has acquired and can bring to bear on the task. In turn, statistical numeracy tests tend to be efficient and robust predictors of decision quality because they predict differences in risk comprehension, which then influence attitudes, intentions, decisions, and behaviors (Cokely et al., 2018). With evidence from recent training studies which suggest that risk literacy and related skills can be trained (e.g., Neth et al., 2018; Ybarra, 2021), the current research suggests that statistical numeracy may also be an efficient, useful, and unbiased estimator of *decision vulnerability*. Similarly, while Dual Systems Theory suggests decision making is constrained by heritable capacities, and as such cannot be meaningfully trained or improved, the present study advances the science of informed decision making (i.e., Skilled Decision Theory) by developing standardized norm-referenced metrics of statistical numeracy and risk literacy. In contrast, the Dual Systems approach does not currently have a systematic way to conduct analyses like the current study, due to the lack of systematized individual difference measurement tools to compare abilities (or capacities) across people, tasks, or time (e.g., "*If you cannot measure it, you cannot improve it.*").

Future Research

The current study provides many new opportunities for future research. While the current study was designed to examine measurement equivalence between genders, the promising evidence that numeracy demonstrated strict measurement invariance encourages future development with respect to assessing equivalence more holistically, across demographic groups (e.g., race, age). For example, a larger-sample study could support studying more complex demographic subgroups (e.g., white males vs. Hispanic women, or the role of education on gender differences). Moreover, many previous studies on norms and measurement equivalence have focused on cultural or country differences (e.g., PISA; see Breakspear, 2012;

Kutner et al., 2006). Given that numeracy and risk literacy skills depend in part on reading fluency, future research can and should consider translating the BNT-S and Risk Literacy Test into other languages. These translated assessments could also be validated using measurement equivalence to support the development of standardized norms, across countries.

Another way to think about invariance is with respect to causality. In Simon's (1990) view "the fundamental goal of science is to find invariants" (p. 1). Simon (1990) suggested that human behavior is a function of the person and the environment (i.e., two blades of scissors). However, a major problem with identifying invariants in psychological research is that "people are adaptive systems, whose behavior is highly flexible" (p. 16). Recent research has started to demonstrate that tests of statistical numeracy are invariant (i.e., a statistical property of a measurement indicating that the same construct is being measured across groups). Future research in this area holds promise for determining other indicators of invariance, and thus causality.

Many researchers and practitioners utilize norms for clinical and diagnostic purposes. In contrast, the Decision Vulnerability framework aims to consider what implications might exist at the group or population level. Said differently, unlike neuropsychology, the decision vulnerability approach is not trying to diagnose individuals (i.e., give one person a score), but rather is more interested in advancing the interpretation of group differences.

To support this goal, future research will be aimed at developing a web-based platform, accessible to researchers, practitioners, and laypeople. This platform will provide a way for researchers to use the decision vulnerability analysis approach (e.g., input data and receive estimates for numeracy thresholds and risk literacy difficulty levels). By developing this open-science framework, the decision vulnerability research project will also have an opportunity to gain valuable information on the bounds/limits of the decision vulnerability approach (e.g., when, under what circumstances, or for whom, is numeracy particularly useful for understanding risk communications?). Furthermore, as norms continue to be developed and updated (e.g., every 5-10 years), a related web-based platform could be created to provide individuals performance feedback (e.g., see RiskLiteracy.org).

Conclusion

The world is getting more complex, and with intensified polarization and cognitive demands, science has a responsibility to empower diverse individuals to reckon with complex risks. Numeracy norms seem to provide meaningful decision vulnerability metrics in applied contexts. By assessing the extent to which measures of numeracy are less likely to give biased estimates across some levels of diversity (e.g., gender), as well as examining the different distributions of subgroups, the current study provides initial steps for psychometrically robust metrics of decision vulnerability. Taken together, robust norms (i) could allow for the prediction of the comprehensibility of different risk communications (e.g., treatment risks, financial products), (ii) may help speak to the generalizability of basic laboratory findings and (iii) may inform the design of decision support technologies (e.g., risk communications, intelligent tutors).

The metrics, models, and norms developed here are far from perfect, but perhaps they begin to provide methods that allow us to answer meaningful scientific questions, by more precisely measuring how much, when, why, and to what extent we can predict and explain fundamental and practical questions. Perhaps in turn, this will provide mechanisms to more precisely track changes in decision making skill across time and space, and to demonstrate that

(i) incremental improvements have occurred (due to interventions) and that (ii) these improvements have caused some economic benefits (e.g., reduced loss of life and property).

Imagine a world in which individuals are better able to make autonomous, informed decisions that are consistent with their goals, preferences, and values. Perhaps advances in the decision vulnerability approach could have supported Doug deSilvey and his family. In the future, how many lives could be saved by improved metrics and assessments of decision vulnerability?

References

Allan, J. N. (2018). Numeracy vs. Intelligence: A model of the relationship between cognitive abilities and decision making.

Allan, J. N., Ripberger, J. T., Wehde, W., Krocak, M., Silva, C. L., & Jenkins-Smith, H. C. (2020). Geographic distributions of extreme weather risk perceptions in the United States. *Risk analysis*, *40*(12), 2498-2508.

Allan, J. N., Ripberger, J. T., Ybarra, V. T., & Cokely, E. T. (2017, June). Tornado risk literacy: Beliefs, biases, and vulnerability. In *13th Bi-Annual Int. Conf. on Naturalistic Decision Making and Uncertainty* (pp. 284-290). Bath, United Kingdom: University of Bath.

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

Antonios, N., & Raup, C. (2012). Buck v. Bell (1927). *Embryo Project Encyclopedia*.

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods, 22*(3), 507–526.

Betz, N. (1978). Prevalence, distribution, and correlates of math anxiety in college students. Journal of Counseling Psychology, 25, 441--448.

Binet, A. (1903). *L'étude expérimentale de l'intelligence*. Schleicher frères & cie.

Boin, A., Brown, C., & Richardson, J. (2019). Analysing a Mega-Disaster: Lessons from Hurricane Katrina.

Bowden, S. C., Gregg, N., Bandalos, D., Davis, M., Coleman, C., Holdnack, J. A., & Weiss, L. G. (2008). Latent mean and covariance differences with measurement equivalence in college students with developmental difficulties versus the Wechsler Adult Intelligence Scale–III/Wechsler Memory Scale–III normative sample. *Educational and psychological measurement*, *68*(4), 621-642.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791-799.

Breakspear, S. (2012). The policy impact of PISA: An exploration of the normative effects of international benchmarking in school system performance.

Brown, R. P., & Josephs, R. A. (1999). A burden of proof: Stereotype relevance and gender differences in math performance. *Journal of Personality and Social Psychology, 76*(2), 246–257.

Bruine de Bruin, W., Parker, A. M., & Fischhoff, B. (2007). Individual differences in adult decision-making competence. *Journal of personality and social psychology*, *92*(5), 938.

Cattell, R. B. (1941). Some theoretical issues in adult intelligence testing. *Psychological Bulletin*, *38*(7), 592.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of educational psychology*, *54*(1), 1.

Cattell, R. B. (1971). Abilities: Their structure, growth, and action.

Cattell, R. B. (1973). *Culture-fair intelligence test*. Institute for personality and ability testing.

Cattell, R. B. (1987). *Intelligence: Its structure, growth and action* (Vol. 35). Elsevier.

Cattell, R. B., Feingold, S. N., and Sarason, S. B. (1941). A culture-free intelligence test: II. Evaluation of cultural influence on test performance. *J. Educ. Psychol.* 32, 81–100.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press.

Centers for Disease Control and Prevention. (2021, March 29). *Understanding Health Literacy*. Centers for Disease Control and Prevention. https://www.cdc.gov/healthliteracy/learn/Understanding.html.

Cho, J. (2020). Numeracy Predicts Accurate Climate Change Knowledge and Beliefs: A Model of Factors That Influence Biases and Polarization.

Cokely, E. T., Feltz, A., Ghazal, S., Allan, J. N., Petrova, D., & Garcia-Retamero, R. (2018). Skilled Decision Theory: From intelligence to numeracy and expertise. *Cambridge handbook of expertise and expert performance*, 476-505.

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: the berlin numeracy test. *Judgment and Decision making*.

Cokely, E. T., & Kelley, C. M. (2009). Cognitive abilities and superior decision making under risk: A protocol analysis and process model evaluation, *Judgment and Decision making, 4(1), 20-33*.

Crowther Report. (1959). Education in England: The History of Our Schools. Retrieved from: http://www.educationengland.org.uk/documents/crowther/crowther1959- 1.html

Daseking, M., Petermann, F., & Waldmann, H. C. (2017). Sex differences in cognitive abilities: Analyses for the German WAIS-IV. *Personality and Individual Differences*, *114*, 145-150.

Del Missier, F., Mäntylä, T., & De Bruin, W. B. (2012). Decision-making competence, executive functioning, and general cognitive abilities. *Journal of Behavioral Decision Making, 25(4),* 331-351.

DeShon, R. P. (2004). Measures are not invariant across groups without error variance homogeneity. *Psychology Science*, *46*, 137-149.

Dhami, M. K., Schlottmann, A., & Waldmann, M. R. (Eds.). (2012). *Judgment and decision making as a skill: Learning, development and evolution*. Cambridge University Press.

Dikkers, R. D., Marshall, R. D., & Thom, H. C. (1971). Hurricane Camille-August 1969 (NIST TN 569).

Embretson, S. & Reise, S. (2013). *Item Response Theory*. Psychology Press.

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, *8*(3), 223-241.

Fechner, G. T. (1948). Elements of psychophysics, 1860.

Feltz, S., & Feltz, A. (2019). Consumer Accuracy at Identifying Plant-based and Animal-based Milk Items. *Food Ethics*, *4*(1), 85-112.

Ferguson, A., Myers, C. S., Bartlett, R. J., Banister, H., Bartlett, F. C., Brown, W., & Tucker, W. S. (1940). Final report of the committee appointed to consider and report upon the possibility of quantitative estimates of sensory events. *Report of the British Association for the Advancement of Science*, *2*, 331-349.

Flanagan, B. E., Gregory, E. W., Hallisey, E. J., Heitgerd, J. L., & Lewis, B. (2011). A social vulnerability index for disaster management. *Journal of homeland security and emergency management*, *8*(1).

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic perspectives*, *19*(4), 25-42.

Galesic, M., & Garcia-Retamero, R. (2010). Statistical numeracy for health: a cross-cultural comparison with probabilistic national samples. *Archives of internal medicine*, *170*(5), 462-468.

Galton, F. (1890). Kinship and correlation. *The North American Review*, *150*(401), 419-431.

Garcia-Retamero, R., & Galesic, M. (2010). Who profits from visual aids: Overcoming challenges in people's understanding of risks. *Social Science & Medicine, 70*, 1019–1025.

Garcia-Retamero, R., & Cokely, E. T. (2017). Designing visual aids that promote risk literacy: a systematic review of health research and evidence-based design heuristics. *Human factors*, *59*(4), 582-627.

Ghazal, S. (2014). *Component numeracy skills and decision making*. Michigan Technological University.

Ghazal, S., Cokely, E. T., & Garcia-Retamero, R. (2014). Predicting biases in very highly educated samples: Numeracy and metacognition. *Judgment and Decision Making, 9*(1), 15–34.

Gigerenzer, G. (2015). *Simply rational: Decision making in the real world*. Evolution and Cognition.

Ginsburg, L., Manly, M., & Schmitt, M. J. (2006). The components of numeracy. *NCSALL Occasional Paper]. Cambridge, MA: National Center for Study of Adult Literacy and Learning.*

Goddard, H. (1912) The Kallikak Family: A Study in the Heredity of Feeble-Mindedness. New York: Macmillan.

Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography.

Gould, S. J. (1996). *The mismeasure of man*. WW Norton & Company.

Hanushek, E. A., & Woessmann, L. (2010). *The high cost of low educational performance: The long-run economic impact of improving PISA outcomes*. OECD Publishing. 2, rue Andre Pascal, F-75775 Paris Cedex 16, France.

Herrnstein, R. J., & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life.* New York: Free Press.

Heckman, J. J. (1995). Lessons from the bell curve. *Journal of Political Economy, 103*, 1091–1120.

Holzinger, K. J., & Swineford, F. (1939). A study in factor analysis: The stability of a bi-factor solution. *Supplementary educational monographs*.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental aging research, 18*(3), 117-144.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, *6*(1), 1-55.

Huck, J. (2020). Who makes better decisions? The relative importance of numeracy and cognitive abilities for elements of decision making.

Huff, D. (1954). How to lie with statistics. New York, NY: WW Norton & Company Inc

Hunt, E. (2010). *Human intelligence*. Cambridge University Press.

Hunt, E., & Wittmann, W. (2008). National intelligence and national prosperity. *Intelligence*, *36*(1), 1-9.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Shah, P. (2011). Short- and long-term benefits of cognitive training. *Proceedings of the National Academy of Sciences USA, 108*, 10081–10086.

Jenny, M. A., Keller, N., & Gigerenzer, G. (2018). Assessing minimal medical statistical literacy using the Quick Risk Test: a prospective observational study in Germany. *BMJ open*, *8*(8).

Jodoin, M. G. and Gierl, M. J. (2001). Evaluating Type I error and power rates using an effect size measure with logistic regression procedure for DIF detection. *Applied Measurement in Education, 14,* 329-349.

Johnson, B. B. (2008). Public views on drinking water standards as risk indicators. *Risk Analysis: An International Journal*, *28*(6), 1515-1530.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*(2), 183-202.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*(4), 409-426.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American economic review*, *93*(5), 1449-1475.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: a latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, *133*(2), 189.

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.

Krenzke, T., Ren, W., Paul, A., Schneider, A., Karne, V., Abbott, M., Kemmerer, J., Mohadjer, L., and Hogan, J. (2020). U.S. PIAAC Skills Map: State and County Indicators of Adult Literacy and Numeracy (NCES 2020-047). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Kutner, M., Greenburg, E., Jin, Y., & Paulsen, C. (2006). The Health Literacy of America's Adults: Results from the 2003 National Assessment of Adult Literacy. NCES 2006-483. *National Center for Education Statistics*.

Liu, O. L., & Wilson, M. (2009). Gender differences in large-scale math assessments: PISA trend 2000 and 2003. *Applied Measurement in Education*, *22*(2), 164-184.

Lord, F. M. (1980). Applications of item response theory to practical testing problems (pp. 181-223). Hillsdale NJ: Erlbaum.

Lubke, G. H., & Dolan, C. V. (2003). Can unequal residual variances across groups mask differences in residual means in the common factor model?. *Structural Equation Modeling*, *10*(2), 175-192.

Luce, R. D., & Suppes, P. (2002). Representational measurement theory. In H. Pashler & J. Wixted (Eds.), *Stevens' handbook of experimental psychology: Methodology in experimental psychology* (pp. 1–41). John Wiley & Sons Inc.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of mathematical psychology*, *1*(1), 1-27.

Magis, D., Beland, S., Raiche, G., & Magis, M. D. (2020). Package ‘difR’.

Mahmoud-Elhaj, D., Tanner, B., Sabatini, D., & Feltz, A. (2020). Measuring objective knowledge of potable recycled water. *Journal of Community Psychology*, *48*(6), 2033-2052.

McCabe, D. P., Roediger III, H. L., McDaniel, M. A., Balota, D. A., & Hambrick, D. Z. (2010). The relationship between working memory capacity and executive functioning: evidence for a common executive attention construct. *Neuropsychology*, *24*(2), 222.

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International journal of educational research*, *13*(2), 127-143.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525-543.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational measurement: Issues and practice*, *14*(4), 5-8.

Miron-Shatz, T., Hanoch, Y., Doniger, G., Omer, Z., & Ozanne, E. M. (2014). Subjective but not objective numeracy influences willingness to pay for BRCA1/2 genetic testing. *Judgment and Decision making, 9(2), 152-158*.

Murray, C. (2021). *Facing Reality: Two Truths about Race in America*. Encounter Books.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78*, 691-692.

National Center for Education Statistics. (2006). The Health Literacy of America's Adults: Results from the 2003 National Assessment of Adult Literacy. Washington, DC: U.S. Department of Education.

NBCUniversal News Group. (2005, September 19). *Katrina forecasters were remarkably accurate*. NBCNews.com. https://www.nbcnews.com/id/wbna9369041.

Neisser, U., Boodoo, G., Bouchard Jr, T. J., Boykin, A. W., Brody, N., Ceci, S. J., ... & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American psychologist*, *51*(2), 77.

Nelson, W., Reyna, V. F., Fagerlin, A., Lipkus, I., & Peters, E. (2008). Clinical implications of numeracy: theory and practice. *Annals of behavioral medicine*, *35*(3), 261-274.

Neth, H., Gaisbauer, F., Gradwohl, N., & Gaissmaier, W. (2018). riskyr: a toolbox for rendering risk literacy more transparent.

Nisbett, R. E. (2009). *Intelligence and how to get it: Why schools and cultures count*. WW Norton & Company.

Okan, Y., Garcia-Retamero, R., Cokely, E. T., & Maldonado, A. (2012). Individual differences in graph literacy: Overcoming denominator neglect in risk comprehension. *Journal of Behavioral Decision Making*, *25*(4), 390-401.

Organization for Economic Cooperation and Development (OECD) (2013). OECD Skills Outlook 2013: First Results From the Survey of Adult Skills. Paris: OECD Publishing.

Pachur, T., & Galesic, M. (2013). Strategy selection in risky choice: The impact of numeracy, affect, and cross-cultural differences. *Journal of Behavioral Decision Making*, *26*(3), 260-271.

Paulos, J. A. (1988). Innumeracy: Mathematical illiteracy and its consequences. Macmillan.

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science*, *31*(7), 770-780.

Peters, E. (2012). Beyond comprehension: The role of numeracy in judgments and decisions. *Current Directions in Psychological Science*, *21*(1), 31-35.

Peters, E. (2017). Educating good decisions. *Behavioural Public Policy*, *1*(2), 162-176.

Peters, E. (2020). *Innumeracy in the wild: Misunderstanding and misusing numbers*. Oxford University Press.

Peters, E., Baker, D. P., Dieckmann, N. F., Leon, J., & Collins, J. (2010). Explaining the effect of education on health: A field study in Ghana. *Psychological science*, *21*(10), 1369-1376.

Peters, E., & Bjalkebring, P. (2015). Multiple numeric competencies: When a number is not just a number. *Journal of personality and social psychology*, *108*(5), 802.

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological science*, *17*(5), 407-413.

Petrova, D., Garcia-Retamero, R., Catena, A., Cokely, E., Heredia Carrasco, A., Arrebola Moreno, A., & Ramírez Hernández, J. A. (2017). Numeracy predicts risk of pre-hospital decision delay: A retrospective study of acute coronary syndrome survival. *Annals of Behavioral Medicine*, *51*(2), 292-306.

Petrova, D., Kostopoulou, O., Delaney, B. C., Cokely, E. T., & Garcia-Retamero, R. (2018). Strengths and gaps in physicians' risk communication: a scenario study of the influence of numeracy on cancer screening communication. *Medical Decision Making*, *38*(3), 355-365.

Pezzuti, L., Tommasi, M., Saggino, A., Dawe, J., & Lauriola, M. (2020). Gender differences and measurement bias in the assessment of adult intelligence: Evidence from the Italian WAIS-IV and WAIS-R standardizations. *Intelligence*, *79*, 101436.

Program for the International Assessment for Adult Competencies (PIAAC) - PIAAC Proficiency Levels for Numeracy. National Center for Education Statistics (NCES) Home Page, a part of the U.S. Department of Education. (n.d.). https://nces.ed.gov/surveys/piaac/numproficiencylevel.asp.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. Applied Psychological Measurement, 14, 197-207

Raju, N.S., van der Linden, W., & Fleer, P. (1995). An IRT-based internal measure of 22 22 test bias with applications for differential item functioning. Applied Psychological Measurement, 19, 353-368.

Ramasubramanian, M. (2020). Individual Differences and Risk Perception: Numeracy Predicts Differences in General and Specific Risk Perceptions.

Ramasubramanian, M., Allan, J. N., Retamero, R. G., Jenkins-Smith, H., & Cokely, E. T. (2019, November). Flood Risk Literacy: Communication and Implications for Protective Action. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 63, No. 1, pp. 1629-1633). Sage CA: Los Angeles, CA: SAGE Publications.

Raven, J. C. (1938). *Raven's progressive matrices*. Los Angeles, CA: Western Psychological Services.

Raven, J.C., Court, J.H., & Raven, J. (1988). Manual for Ravens' Progressive Matrices and Vocabulary Scales (Section 4). London: H. K. Lewis.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*(3), 552–566.

Reyna, V. F., Nelson, W. L., Han, P. K., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychological Bulletin, 135*(6), 943–973.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of statistical software*, *17*(5), 1-25.

Rosseel, Y., Oberski, D., Byrnes, J., Vanbrabant, L., Savalei, V., Merkle, E., Hallquist, M., Rhemtulla, M., Katsikatsou, M., Barendse, M. and Chow, M. (2017). Package 'lavaan'. *Retrieved June*, *17, 2017.*

Roozenbeek, J., Schneider, C. R., Dryhurst, S., Kerr, J., Freeman, A. L., Recchia, G., ... & Van Der Linden, S. (2020). Susceptibility to misinformation about COVID-19 around the world. *Royal Society open science*, *7*(10), 201199.

Ruch, F. L., & Ruch, W. W. (1963). Employee aptitude survey: Technical report. Los Angeles: Psychological Services.

Savage, L. J. (1954). *The foundations of statistics*. Courier Corporation

Schapira, M. M., Walker, C. M., Miller, T., Fletcher, K. E., Ganschow, P. S., Jacobs, E. A., Imbert, D., O'Connell, M., & Neuner, J. M. (2014). Development and Validation of the Numeracy Understanding in Medicine Instrument Short Form. *Journal of health communication*, *19*(2), 240-253.

Schapira, M. M., Walker, C. M., & Sedivy, S. K. (2009). Evaluating existing measures of health numeracy using item response theory. *Patient education and counseling*, *75*(3), 308-314.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*(2), 262–274.

Schwartz, L. M., Woloshin, S., Black, W. C., & Welch, H. G. (1997). The role of numeracy in understanding the benefit of screening mammography. *Annals of internal medicine*, *127*(11), 966-972.

Simon, H. A. (1990). Invariants of human behavior. *Annual review of psychology*, *41*(1), 1-20.

Spearman, C. (1927). The measurement of intelligence. *Nature*, *120*(3025), 577-578.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and brain sciences*, *23*(5), 645-665.

Sternberg, R. J. (1977). *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities*. Lawrence Erlbaum.

Sternberg, R. J. (2003). *Wisdom, intelligence, and creativity synthesized*. Cambridge University Press.

Stevens, S.S. (1946). On the theory of scales of measurement. *Science, 103*, 677-80.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement*, *27*(4), 361-370.

Tate, E. (2013). Uncertainty analysis for a social vulnerability index. *Annals of the association of American geographers*, *103*(3), 526-543.

Tay, L., Meade, A. W., & Cao, M. (2015). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods*, *18*(1), 3-46.

Terman, L. M. (1916). The uses of intelligence tests. *The measurement of intelligence*, 3-21.

Terman, L. M. (1921). Intelligence and its measurement: A symposium--II. *Journal of Educational Psychology, 12*(3), 127–133.

Terman, L. M., & Oden, M. H. (1947). *The gifted child grows up: twenty-five years' follow-up of a superior group.* Stanford Univ. Press.

Thorndike, E. L. (Ed.). (1903). *Heredity, Correlation and Sex Differences in School Abilities: Studies from the Department of Educational Psychology at Teachers College, Columbia University* (Vol. 11, No. 2). Macmillan.

Thorndike, R. L. (1949). Personnel selection; test and measurement techniques.

Thurstone, L. L. (1938). Primary mental abilities.

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist, 24*(2), 83–91.

U.S. Census Bureau. (2016). Annual County Resident Population Estimates by Age, Sex, Race, and Hispanic Origin: April 1, 2010 to July 1, 2016. Retrieved from https://www2.census.gov/ programs-surveys/popest/datasets/2010-2016/counties/asrh/

U.S. Census Bureau. (2016). American Community Survey Table S1501 2016: ACS 1-Year Estimates Subject Tables for Educational Attainment. Retrieved from https://data.census.gov/cedsci/table?q=education&g=0100000US&y=2016&tid=ACSST1Y2016.S1501&hidePreview=false

U.S. Department of Education, National Center for Education Statistics, Program for the International Assessment of Adult Competencies (PIAAC), U.S. PIAAC 2012/2014.

Van De Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Measurement invariance. *Frontiers in psychology*, *6*, 1064.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational research methods*, *3*(1), 4-70.

Wicherts, J. M., & Dolan, C. V. (2010). Measurement invariance in confirmatory factor analysis: An illustration using IQ test performance of minorities. *Educational Measurement: Issues and Practice*, *29*(3), 39-47.

Wicherts, J. M., & Dolan, C. V. (2010). Measurement invariance in confirmatory factor analysis: An illustration using IQ test performance of minorities. *Educational Measurement: Issues and Practice*, *29*(3), 39-47.

Weschler, D. (1955). Manual for the Weschler adult intelligence scale. *New York: Psychological Corporation*.

Wonderlic, E. F. (1983) Wonderlic Personnel Test manual. NorthField, IL: E. F. Wonderlic & Assoc.

Wonderlic. (2018). The Official Wonderlic Test Online. Retrieved April 06, 2018, from https://www.wonderlic.com/

Ybarra, V. (2021). General Risk Literacy Skills are Causally Related to Improved Self-Evaluations: Training Graph Literacy Improves Decision Making and Lowers Overconfidence

Ybarra, V., Cokely, E. T., Adams, C., Woller-Carter, M., Allan, J., Feltz, A., & Garcia-Retamero, R. (2017). Training Graph Literacy: Developing the RiskLiteracy.org Outreach Platform. In *CogSci*.

Zumbo, B. D. and Thomas, D. R. (1997). A measure of effect size for a model-based approach for studying DIF. Prince George, Canada: University of Northern British Columbia, Edgeworth Laboratory for Quantitative Behavioral Science.

Appendix A: Measurements

Schwartz Numeracy Test

You will now be asked to solve a few problems. Please note that you are allowed to enter numbers that include up to 2 decimal points (for example, 1.11). You are also welcome to use a calculator to help solve these problems.

Imagine that we flip a fair coin 1,000 times. What is your best guess about how many times the coin would come up heads in 1,000 flips?

____

In the BIG BUCKS LOTTERY, the chance of winning a $10 prize is 1%. What is your best guess about how many people would win a $10 prize if 1,000 people each buy a single ticket to BIG BUCKS?

____

In ACME PUBLISHING SWEEPSTAKES, the chance of winning a car is 1 in 1,000. What percent of tickets to ACME PUBLISHING SWEEPSTAKES win a car?
______________________ percent

Berlin Numeracy Test

Out of 1,000 people in a small town 500 are members of a choir. Out of these 500 members in a choir 100 are men. Out of the 500 inhabitants that are not in a choir 300 are men. What is the probability that a randomly drawn man is a member of the choir? Please indicate the probability as a percent.
______________________ percent

Imagine we are throwing a five-sided die 50 times. On average, out of these 50 throws how many times would this five-sided die show an odd number (1, 3 or 5)?

____

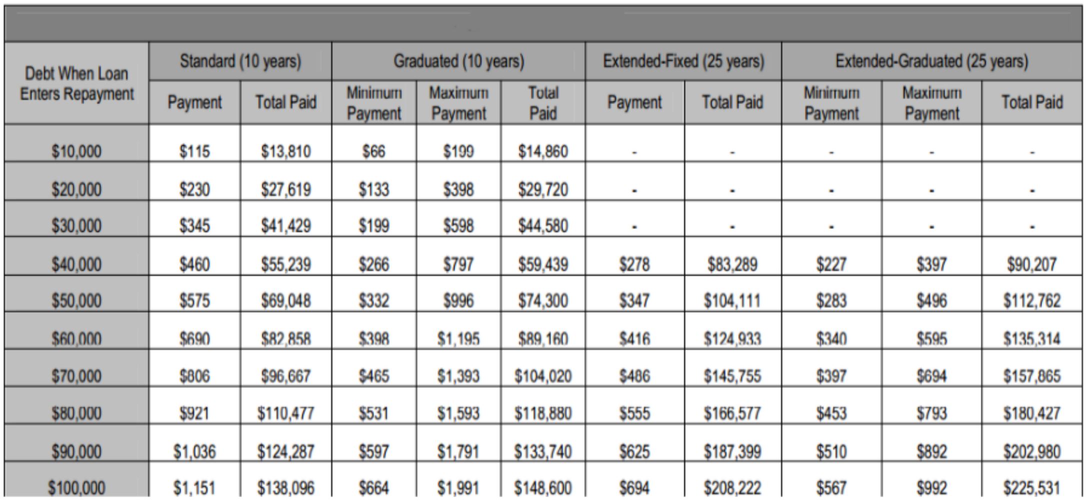
Imagine we are throwing a loaded die (6 sides). The probability that the die shows a 6 is twice as high as the probability of each of the other numbers. On average, out of these 70 throws how many times would the die show the number 6?

____

In a forest, 20% of the mushrooms are red, 50% are brown, and 30% are white. A red mushroom is poisonous with a probability of 20%. A mushroom that is not red is poisonous with a probability of 5%. What is the probability that a poisonous mushroom in the forest is red? Please indicate the probability as a percent.

______________________ percent

Imagine that you take out a $50,000 federal student loan to help pay for college. You are offered four possible repayment plans. The table below provides examples of the monthly repayments for each plan. Note: For the Graduated (10 years) plan, you would start by paying the minimum amount; the payment amount then increases every two years up to the maximum amount.
Look at the table carefully and answer the following questions.



Debt When Loan Enters Repayment	Standard (10 years)		Graduated (10 years)			Extended-Fixed (25 years)		Extended-Graduated (25 years)		
	Payment	Total Paid	Minimum Payment	Maximum Payment	Total Paid	Payment	Total Paid	Minimum Payment	Maximum Payment	Total Paid
$10,000	$115	$13,810	$66	$199	$14,860	-	-	-	-	-
$20,000	$230	$27,619	$133	$398	$29,720	-	-	-	-	-
$30,000	$345	$41,429	$199	$598	$44,580	-	-	-	-	-
$40,000	$460	$55,239	$266	$797	$59,439	$278	$83,289	$227	$397	$90,207
$50,000	$575	$69,048	$332	$996	$74,300	$347	$104,111	$283	$496	$112,762
$60,000	$690	$82,858	$398	$1,195	$89,160	$416	$124,933	$340	$595	$135,314
$70,000	$806	$96,667	$465	$1,393	$104,020	$486	$145,755	$397	$694	$157,865
$80,000	$921	$110,477	$531	$1,593	$118,880	$555	$166,577	$453	$793	$180,427
$90,000	$1,036	$124,287	$597	$1,791	$133,740	$625	$187,399	$510	$892	$202,980
$100,000	$1,151	$138,096	$664	$1,991	$148,600	$694	$208,222	$567	$992	$225,531

rlp_min. Which option has the minimum interest payment (least expensive overall)? [See Table; Fill in the blank]

rlp_percentage. What is the total interest paid in percentage if you have borrowed $50,000 and returned $69,048? [See Table; Fill in the blank]

rlp_beno. With the new drug BENOFRENO, the risk of death from a heart attack may be reduced for people with high cholesterol. A study of 900 people with high cholesterol showed that 80 of the 800 people who have not taken the drug died after a heart attack, compared with 16 of the 100 people who did take the drug. How beneficial was the Benofreno? [1-7 Scale]

rlp_strokex1. Mrs. Jones is told she has a 28 in 1,000 chance of dying from cancer and a 59 in 1,000 chance of dying from a stroke. Mrs. Jones's doctor now tells her that a new pill, STROKEX, will lower her chance of dying from stroke by 50%. Another pill, CANCERX will lower her chance of dying from cancer by 50%.
Assume she can only take 1 pill. Assuming the 2 pills are equally safe and cost the same, which should she take to minimize her risks of death?

a) STROKEX pill
b) CANCERX pill
c) Both are equally effective
d) Neither pill is effective

rlp_3400or3800. Imagine you were offered the following choices. Which option would you select:

a) $3400 this month
b) $3800 next month*

rlp_lose400. Imagine you were offered the following choices. Which option would you select:

a) 50% chance to lose $400
b) Lose $50 for sure*

rlp_gain400. Imagine you were offered the following choices. Which option would you select:

a) Gain $100 for sure
b) 75% chance to win $200*

oper_fieldtrip. A school is having a field trip and many parents are going on the field trip with the children. What is the child to parent ratio if there are 20 children and 5 parents?

a) 2 children for every one parent
b) 20 children for every 1 parent
c) 1 child to every 5 parents
d) 5 children to every 1 parent
e) 4 children to every 1 parent*

prob_dice5. People often roll dice when playing games. Most dice have 6 sides and each side has a different number on it ranging from 1-6. If you rolled one of the dice, on average what is the probability that it would land on 5?

a) 1 time out of 6 rolls of the dice*
b) 5 times out of 6 rolls of the dice
c) 1 time out of 2 rolls of the dice
d) 1 out of 5 rolls of the dice
e) 6 out of 1 roll of the dice

prob_burn. Imagine that the probability of a child getting sunburned at the beach is 65% while the probability of an adult getting sunburned at the beach is 15%. If there were 300 people who spent a day at the beach, and 60% of the people were children, how many people are likely to get a sunburn?

a) About 195
b) About 150
c) About 135*
d) About 80
e) About 64

oper_goods. Imagine that goods imported into a country increased by 40% and exports decreased by 30% during a certain year. What was the ratio of imports to exports at the end of the year compared to the beginning of the year?

a) 1/2
b) 3/2
c) 4/3
d) 2/1*
e) 1

prob_diceeven. Imagine that you are throwing 2 regular 6-sided dice up in the air. If each side has a different number on it ranging from 1-6, on average what is the probability that both of them land on even numbers?

a) 1 out of 36 rolls of the dice
b) 3 out of 6 rolls of the dice
c) 1 out of 4 rolls of the dice*
d) 2 out of 6 rolls of the dice
e) 2 out of 36 rolls of the dice

rlp_400or100. Imagine you were offered the following choices. Which option would you select?

a) $400 now
b) $100 every year for 10 years*

rlp_1000or2400. Imagine you were offered the following choices. Which option would you select?

a) $1000 in six months
b) $2400 in two years*

windspeed comprehension. Which of the following statements best describes the probability of hurricane-force winds in **Location A**?

a) There is an 80-90% chance of hurricane-force winds in Location A during the next 5 days.
b) There is an 80-90% chance of hurricane-force winds in Location A in each of the next 5 days.
c) There is an 80-90% chance of hurricane-force winds occurring somewhere along the southern Gulf coast of Florida during the next 5 days.
d) Not sure

storm surge comprehension. Which of the following statements best describes the storm surge forecast for **Location A**?

a) There is a 10% chance that Location A will get more than 9 ft of storm surge above ground level.
b) There is a 10% chance that Location A will get less than 9 ft of storm surge above ground level.
c) There is a 10% chance that Location A will get approximately 9 ft of storm surge above ground level.
d) Not sure

Appendix B: Validation Results

Table B1.

Validation Results

Artifact	Group	Predicted Risk Literacy Percentile	Proportion Correct, Weighted Sample	True Score Levels	Predicted Value Levels	Deviation = Predicted Value - Weighted Proportion	> 10% Deviation?
prob_dice5	General	0.374	0.295	I	I	0.079	F
prob_dice5	Female	0.411	0.357	I	II	0.055	F
prob_dice5	Male	0.341	0.228	I	I	0.113	T
prob_dice5	NonCollege	0.483	0.439	II	II	0.044	F
prob_dice5	College	0.294	0.191	I	I	0.103	T
prob_dice5	Under55	0.389	0.312	I	II	0.076	F
prob_dice5	Over55	0.343	0.262	I	I	0.081	F
prob_dice5	OtherRace	0.502	0.426	II	II	0.075	F
prob_dice5	White	0.322	0.223	I	I	0.099	F
prob_burn	General	0.628	0.572	II	III	0.056	F
prob_burn	Female	0.670	0.639	III	III	0.031	F
prob_burn	Male	0.595	0.499	II	II	0.095	F
prob_burn	NonCollege	0.768	0.713	III	III	0.055	F
prob_burn	College	0.551	0.470	II	II	0.081	F
prob_burn	Under55	0.638	0.577	II	III	0.062	F
prob_burn	Over55	0.606	0.562	II	II	0.044	F
prob_burn	OtherRace	0.740	0.749	III	III	0.009	F
prob_burn	White	0.588	0.476	II	II	0.112	T
oper_goods	General	0.777	0.757	III	III	0.020	F
oper_goods	Female	0.822	0.792	III	III	0.030	F
oper_goods	Male	0.744	0.718	III	III	0.025	F
oper_goods	NonCollege	0.935	0.848	III	IV	0.087	F
oper_goods	College	0.702	0.691	III	III	0.011	F
oper_goods	Under55	0.785	0.766	III	III	0.020	F
oper_goods	Over55	0.761	0.740	III	III	0.020	F
oper_goods	OtherRace	0.879	0.858	III	IV	0.021	F
oper_goods	White	0.743	0.702	III	III	0.041	F
oper_fieldtrip	General	0.114	0.202	I	I	0.089	F
oper_fieldtrip	Female	0.146	0.239	I	I	0.092	F
oper_fieldtrip	Male	0.081	0.163	I	I	0.082	F
oper_fieldtrip	NonCollege	0.192	0.315	I	I	0.123	T
oper_fieldtrip	College	0.030	0.121	I	I	0.091	F

oper_fieldtrip	Under55	0.133	0.229	I	I	0.096	F
oper_fieldtrip	Over55	0.073	0.153	I	I	0.080	F
oper_fieldtrip	OtherRace	0.258	0.335	I	I	0.076	F
oper_fieldtrip	White	0.051	0.131	I	I	0.080	F
prob_diceeven	General	0.754	0.748	III	III	0.006	F
prob_diceeven	Female	0.799	0.787	III	III	0.012	F
prob_diceeven	Male	0.721	0.706	III	III	0.015	F
prob_diceeven	NonCollege	0.909	0.811	III	IV	0.098	F
prob_diceeven	College	0.679	0.703	III	III	0.024	F
prob_diceeven	Under55	0.763	0.750	III	III	0.013	F
prob_diceeven	Over55	0.737	0.745	III	III	0.008	F
prob_diceeven	OtherRace	0.858	0.805	III	III	0.053	F
prob_diceeven	White	0.720	0.718	III	III	0.002	F
BNTC_3item	General	0.604	0.541	II	II	0.063	F
BNTC_3item	Female	0.646	0.596	II	III	0.050	F
BNTC_3item	Male	0.571	0.482	II	II	0.089	F
BNTC_3item	NonCollege	0.741	0.666	III	III	0.074	F
BNTC_3item	College	0.526	0.451	II	II	0.076	F
BNTC_3item	Under55	0.615	0.552	II	II	0.063	F
BNTC_3item	Over55	0.581	0.521	II	II	0.060	F
BNTC_3item	OtherRace	0.717	0.678	III	III	0.039	F
BNTC_3item	White	0.563	0.467	II	II	0.096	F
BNTC_5item	General	0.581	0.515	II	II	0.066	F
BNTC_5item	Female	0.622	0.563	II	II	0.060	F
BNTC_5item	Male	0.548	0.463	II	II	0.085	F
BNTC_5item	NonCollege	0.715	0.625	III	III	0.090	F
BNTC_5item	College	0.503	0.435	II	II	0.068	F
BNTC_5item	Under55	0.592	0.527	II	II	0.066	F
BNTC_5item	Over55	0.557	0.493	II	II	0.065	F
BNTC_5item	OtherRace	0.696	0.635	III	III	0.061	F
BNTC_5item	White	0.539	0.450	II	II	0.089	F

Note. BNTC_3item is the sum of prob_dice5 + prob_burn + oper_goods.
BNTC_5item adds an easier item (oper_fieldtrip) and a harder item (prob_diceeven), to sum:
prob_dice5 + prob_burn + oper_goods + oper_fieldtrip + prob_diceeven