UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

General Risk Literacy Skills are Causally Related to Improved Self-Evaluations: Training Graph Literacy Improves Decision Making and Lowers Overconfidence

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the Degree of

DOCTOR OF PHILOSOPHY

By

Vincent Thomas Ybarra

Norman, Oklahoma

2021

General Risk Literacy Skills are Causally Related to Improved Self-Evaluations:

Training Graph Literacy Improves Decision Making and Lowers Overconfidence

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF PSYCHOLOGY

BY THE COMMITTEE CONSISTING OF

Dr. Edward Cokely, Chair

Dr. Rocio Garcia-Retamero

Dr. Adam Feltz

Dr. Shane Connelly

Dr. Saeed Salehi

Acknowledgements

"One heuristic that has been of first importance…To make interesting scientific discoveries you should acquire as many good friends as possible, who are as energetic, intelligent, and knowledgeable as they can be."

-Herb Simon from *Models of My Life*

I would like to first thank all the people who have befriended and supported me over my graduate school career. Thank you to Dr. Edward Cokely for all your guidance and the opportunities you have given me. Thank you to my lab mates Jinan Allan, Madhuri Ramasubramanian, and Jinhyo Cho for being as energetic, intelligent, and knowledgeable as you can be. Thank you to my wife, Grace Williams, for your support and for keeping me afloat. Thank you to my parents, Ernesto, and Gloria Ybarra, for all your encouragement and for letting me make my own informed decision to pursue psychology. Thank you to all my committee members for all your time, support, and knowledge that you have given me over the years.

Table of Contents

Abstract

Overconfidence commonly refers to a cognitive bias in which people express more subjective confidence in their abilities than can be objectively justified. Similarly, the "Unskilled and Unaware" literature demonstrates that individuals who are lowest in experience or knowledge are often the most overconfident. Further, recent research on decision aids and training programs suggests that improving task-specific skills can causally reduce task-specific overconfidence biases. However, these training benefits do not appear to transfer or otherwise reduce overconfidence on untrained judgment tasks more generally. One important example of general training is Morewedge et al. (2015), where training specific decision biases (e.g., confirmation bias) led to reduced overconfidence for better-than-average overconfidence effects. While these findings are promising, research has not yet investigated the potential benefits of programs designed to train general decision making skills, such as risk literacy skills (i.e., the ability to evaluate and understand risk; See Cokely et al., 2018). Accordingly, I report results from the first experimental study to test the benefits of an online risk literacy tutor (tutor.lolevo.com) designed to improve graph literacy skills (e.g., interpreting bar, pie, line, icon arrays, and decision tree graphs). Statistical and structural modeling results revealed that the graph literacy tutor causally improved graph literacy skills, which in turn causally reduced biases on several untrained, conceptually distinct decision tasks (framing effects, ratio bias, and sunk costs). Moreover, improvements in graph literacy also transferred benefits to untrained overconfidence measures, causally improving self-evaluations on the untrained decision tasks. Consistent with Skilled Decision Theory, results indicated that risk literacy skills are trainable and are related to general decision making skills. Further, results suggested that training risk literacy skills could have implications for an abundance of risky situations (e.g., Covid-19 risk knowledge, heart attack symptom recognition) because training risk literacy can reduce many kinds of decision biases,

including multiple types of potentially trainable overconfidence biases (e.g., overestimation, overprecision, overplacement). In addition to other theoretical implications, it is noteworthy that the current study appears to be the first to discover a direct causal link between risk literacy skills and valuable metacognitive skills. Risk literacy skills causally promote skilled self-evaluation and accurate social comparisons.

Chapter 1: Introduction

Some individuals are overconfident, thinking they have more skills than reality justifies (Lichenstein, Fischhoff, & Phillips, 1977; Moore & Healy, 2008). One major factor in why people make biased decisions may be their overconfidence bias (Kahneman, 2011). For years, efforts to reduce overconfidence have often resulted in mixed outcomes (Moore et al., 2018). Some researchers also believe that overconfidence is not likely to improve through training (Kahneman, 2011). However, the research found in the "Unskilled and Unaware" literature showed that individuals could improve in overconfidence with gained experience for specific domains (Kruger & Dunning, 1999; Ehrlinger et al., 2008). Further, Morewedge et al. (2015) demonstrated that an individual could lower overconfidence for social comparisons (better-than-average effects) by training specific decision biases and providing feedback on confidence judgments. While Morewedge et al. (2015) are among the first to demonstrate the reduction of overconfidence by training specific decision making tasks, there is currently no evidence that overconfidence can be reduced via training general decision making skills like risk literacy skills (i.e., the ability to evaluate and understand risk). Growing evidence demonstrates that risk literacy skills, like statistical numeracy and graph literacy, are related to more calibrated self-evaluations and less overconfidence bias (Ghazal et al., 2014; Ybarra et al., in-prep). Evidence from risk literacy and previous overconfidence research leads to the idea that overconfidence could potentially be trained generally (Cokely et al., 2018; Ybarra et al., 2017). So, does training risk literacy skills help individuals avoid the overconfidence bias?

The following study took a previously validated risk literacy tutor and experimentally measured group differences in graph literacy, three decision making tasks (framing, ratio bias,

and sunk cost), and three distinct types of overconfidence (overestimation, overprecision, and overplacement). The results indicated that the risk literacy tutor tended to improve graph literacy and decision making task performance when controlling for statistical numeracy. Overconfidence in the decision making tasks tended to be more calibrated for those who completed the risk literacy tutor. The reduction in overconfidence tended to be for individuals at all levels of task performance, suggesting improved performance sensitivity for those who completed the risk literacy tutor. Implications suggest that risk literacy skills play one vital role in making accurate self-evaluations.

Chapter 2: Confidence and Risk Literacy

2.1 History of Self-Evaluations

How do we know what we know? The study of self-evaluations (comparisons of self-monitored information to a standard or a goal; Bandura, 1986) has a long history that dates to Socrates and has persisted through time. Many notable scientists and philosophers have touched on self-evaluations and the importance of knowing what we know and who we are (e.g., self-knowledge) – often focusing on knowing oneself and the consequences of not (Wilson, 2009). For example, some philosophers believe that accurate self-knowledge is crucial as it is 1) required for wisdom and 2) essential for knowledge (Renz, 2017). Additionally, some psychologists believe that accurate self-evaluations are essential for making good decisions related to health, education, social issues, and workplace matters (Dunning et al., 2004; Moore, 2020; Grant, 2021).

This discussion of self-evaluations finds its roots, like many others, in ancient Greece. In the town of Delphi, one can find the ruins of a temple built in dedication to the god Apollo and once inscribed with 147 sayings called the Delphic Maxims attributed to Apollo himself (Markopoulos & Vanharanta, 2017). This temple became a mainstay of some of the great philosophers of the time (470-399 BCE), including Socrates. The maxims' influence on Socrates' work is evident in one discussion found in Plato's *Apology* (399 BCE; Kamtekar, 2016; 2017), where Socrates recalled the maxim "Know thyself." In this recollection, Socrates spoke of the craftsmen who know only their crafts but falsely believed that they knew virtue. A modern cognitive psychology interpretation of this recollection could be that the craftsmen knew themselves in their self-evaluations related to their skill, but less so in other concepts, like virtue.

"Know thyself" appears in more of Socrates' teachings. For example, one of Socrates' students, Xenophon, wrote in *Memorabilia* (371 BCE; Bonnette, 1994) about the importance of the maxim "Know thyself." Xenophon quoted Socrates:

> [T}hose who know themselves, know what things are expedient for themselves and discern their own powers and limitations. And by doing what they understand, they get what they want and prosper: by refraining from attempting what they do not understand, they make no mistakes and avoid failure. And consequently, through their power of testing other men too, and through their intercourse with others, they get what is good and shun what is bad.

Thus, as quoted, Socrates described the benefits of knowing oneself and knowing others (for further information, see *Socrates and Self-Knowledge* by Moore, 2015).

Around the turn of the millennium, the idea of self-evaluations shifted to focus more on the human connection with and distinction from God (Renz, 2017). Saint Augustin (354-430 CE) wrote about the ethical implications for knowing oneself and how it relates to the individual's relationship with God (Brachtendorf, 2017). Augustin's interpretation implies that a sinner is aware of their actions but lacks accurate self-evaluations and loves themself too much. Thomas Aquinas (1225 – 1274 CE) carried on the religious line of thinking about self-knowledge. Aquinas wrote that God is omniscient. We as humans reflect God but do not possess complete knowledge of ourselves and must think about thinking when observing external objects or situations (Perler, 2017). For instance, Aquinas wrote that we as people use experiences to construct labels that define ourselves (e.g., "I am a parent" or "I am a good person"; Cory, 2014).

The idea of self-evaluations can be found in the writings of modern philosophers as well. One interpretation of the writings of Descartes (1596-1650 CE) is that knowing yourself may contribute to one's happiness (Renz, 2017). Further, knowing our limitations is important for knowing ourselves, and understanding our limitations is derived from our failures and understanding of others' failures. Thomas Hobbes (1588-1657 CE) wrote about the idea that our self-evaluations are built from our understanding of our actions and others' actions. Renz (2018) put forth the idea, based on Hobbes' work *The Leviathan*, that access to one's own experience is a prerequisite for self-knowledge. Garret (2017) believed that Hobbes wrote that comparisons of commonalities with others are essential for knowing thyself. Finally, Immanuel Kant's (1724-1804 CE) writings touched on the concept of self-knowledge. Emundts (2017) spoke about Kant's idea that we can deceive our self-knowledge into justifying our immoral actions. These ideas of modern philosophers set the underpinnings of theories of self-evaluations in psychology.

By the end of the 19th century, psychology was in its infancy, yet the idea of self-evaluations was present. For example, in William James' 1880 *Principles of Psychology,* he proposed an equation for self-esteem (a concept thought to be related to self-evaluations and overconfidence) where self-esteem is the ratio of one's success to their aspirations. James has an entire chapter dedicated to the conception of the self and how people self-identify. Sigmund Freud, around this time, explored self-evaluations with his psychoanalytic therapy, in which a patient could express their true selves (Reginster, 2017). The first formal self-evaluation measures began around this time**.** The first confidence studies began in 1884 with Pierce and Jastrow, and the concept continued in 1892 in a study by Fullerton and Cattell (Bjorkman, Juslin, & Winman, 1993). Psychology matured as a field around this time, which led to many research streams related to self-evaluations. Of these many branches, I am particularly interested in the inaccuracy of self-evaluations (e.g., overconfidence) and how it is related to decision making.

2.2 Overconfidence in Decision Making

For centuries, accurate self-evaluations have been considered significant for a variety of reasons. For instance, accurate self-evaluations tend to be one of the predictors of academic achievement, accounting for 46.3% of the total variance (Stankov et al., 2012). However, some individuals falsely believed they were better, exhibited more skill, or had more ability than they do (Kruger & Dunning, 1999; Griffin & Brenner, 2004). This phenomenon wherein individuals believe themselves to be better than they are or unjustly think they are better than others is called the *overconfidence bias*. One definition of the overconfidence bias from Moore and Schatz (2017) is "greater confidence than reality justifies." Moore and Schatz say that people display overconfidence bias when they compare incorrect self-beliefs with reality. For example, a student who believes they did well on a test but actually failed may be considered overconfident.

In his 2011 book *Thinking Fast and Slow,* Kahneman called the overconfidence bias "the most significant of the cognitive biases,", especially for decision making. Kahneman gave several examples of how people are overconfident in their decisions. In one example, stock traders overconfidently picked stocks only to end up losing money. In another instance, Kahneman described how the most confident clinicians were wrong about patient diagnoses 40% of the time. Kahneman stated that overconfidence "contributes to an explanation of why people litigate, why they start wars, and why they open small businesses."

Kahneman's view on overconfidence is only one perspective of many from various fields that study overconfidence. Stankov and Kleitman (2008) gave an overview of different streams of confidence research, the first being related to self-confidence and personality. In this stream of research, confidence tends to be a personality trait like "assertiveness." The next stream focuses on education and psychological assessment, for which the example given is the use of overconfidence in cognitive tests and focuses on self-regulation (a metacognitive process thought to be necessary for learning). In this research, self-evaluations are thought to be an integral part of self-regulatory behaviors (Flavell, 1979; Zimmerman, 2000). The last stream of confidence research stems from psychologists studying decision making. This stream of overconfidence research is the primary focus of this dissertation.

The stream of research focused on overconfidence in decision making is often believed to have gained popularity from a 1977 study by Lichenstein and Fischhoff. The study aimed to see if individuals who knew nothing on a subject (picking stocks) would know if they were poor performing on that subject. They gave participants trend data on various stocks and asked which ones would increase in value over five weeks. Lichenstein and Fischhoff then asked, "on a scale from 50% to 100% what is the probability that your choice is correct?" They found that most participants were overconfident. On average, participants were correct only about 47% of the time but rated their confidence an average 18% higher. In an additional experiment, they found that individuals who were more knowledgeable on a subject tended to be less overconfident than less knowledgeable individuals.

The 1977 Fischhoff and Lichenstein study also formalized a way of measuring overconfidence using a calibration equation. The original equation is

$$\text{Calibration} = (1/N)\ T\ t=1\ nt\ (rt - ct)$$

Where N is the total number of responses, n_t is the number of times the response r_t was used, c_t is the proportion correct for all items assigned probability r_t, and T is the total number of different response categories used. In sum, calibration is the difference between subjective confidence judgments and the objective number of questions correct. A positive difference indicates overconfidence. This method of measurement of overconfidence persists in contemporary confidence studies.

Fischhoff and Lichenstein had two more published studies in 1977 on the topic of overconfidence. One completed with Paul Slovic asked participants to rate their confidence in answering general knowledge questions in various formats (e.g., open-ended vs. two alternatives). Like the previous study, participants tended to be overconfident in their degree of correctness. More overconfident participants were also more willing to gamble money on their correctness. The other of their 1977 overconfidence studies was an overview of previous studies on overconfidence. Findings included that participants believed themselves to be better than they actually were on a multitude of different tasks, that there is little research dedicated to training and overcoming overconfidence, and individuals like weather forecasters tend to be less overconfident when weather forecasting, but the reasons why are unclear.

Fischhoff and Lichenstein's series of studies in 1977 provided evidence for a systematic overconfidence bias in various tasks. However, Fischhoff and Lichenstein did not fully theorize why and how individuals were overconfident. Koriat et al. (1980) conducted the first study dedicated to exploring how individuals are biased in their confidence. The study again had individuals answer general knowledge questions and then subjectively estimate their proportion correct. However, this time participants had to provide explanations to support their decision. When participants gave less evidence to support contradicting alternatives, they were more likely to be overconfident. The thought being that overconfidence derives from neglecting evidence. Moreover, more salient evidence is associated with reductions in overconfidence.

The idea of making evidence more salient can be found in one of the first papers to theorize how confidence judgments are processed. The *Probabilistic Mental Models* (PMM) theory by Gigerenzer et al. (1991) is based on *probabilistic functionalism* by Egon Brunswik (1952; Hammond & Stewart, 2001). Probabilistic functionalism is the idea that an organism deals with uncertainty in the environment by evaluating its cues. Thus, probabilistic functionalism would suggest that overconfidence results when non-salient cues fail to be understood. Individuals then inaccurately assess their performance, such as the overconfident participants in Koriat et al. (1980).

To construct a probabilistic mental model, an individual creates a reference class, a target variable, probability cues, and cue validities. The example in the 1991 study detailed responses to a general knowledge question like "Which city in Oklahoma has more inhabitants? a) Oklahoma City or b) Tulsa." Suppose the participant answering the question does not know the answer. The participant would have to construct a PMM to solve the question. One of the first steps in constructing a PMM is selecting a reference class. The reference class in this example might be "all cities in Oklahoma." This reference class then determines which cues can function as probability cues for the target variable and their cue validities. For instance, a valid cue in the reference class might be NBA teams. Larger cities tend to have professional NBA teams, so which of the two cities has a professional NBA team? The cue is cognitively tested for correctness and compared to other cues. The cue validity then determines confidence judgments. For example, the NBA team cue may be strong, leading to the correct answer of Oklahoma City and a high confidence rating. Of course, cue generation depends on knowledge and frequentist inference.

The PMM model is one prevailing theory in how and why individuals exhibit the overconfidence bias, but it still leaves many questions unanswered. For example, how do individuals become less overconfident? Juslin et al. (1997) explained that for an individual to have perfect calibration under the theory of PMM, the individual would have to 1) have accurate mental representations of the ecological probabilities (*cognitive adjustment*, something akin to representative knowledge based on a reference class), 2) without error convert the ecological probabilities into an overt confidence judgment (*error-free translation*), and 3) encounter tasks that are representatively designed corresponding to their ecological probabilities. The third point, in essence, states that the individual's tasks should represent the knowledge that the individual would sample. Gigerenzer et al. (1991) gave the example of a typical general knowledge question that asked, "Which city is farther north? a) New York or b) Rome" as evidence to this point. The participant could generate a reference class of "general knowledge question," which may lead to an incorrect answer and overconfidence. The idea is that the generated reference class is incorrect due to the question being hard or misleading while selecting a more representative question like "Which city in Oklahoma has more inhabitants?" leads to the generation of the reference class "cities in Oklahoma." This reference class then may lead to a correct answer and less overconfidence. How, then, do individuals generate reference classes?

Koriat (2012) constructed a model called the *Self-Consistency Model* (SCM) that postulates that the primary cue that confidence judgments use is self-consistency. According to the SCM, individuals validate their knowledge or experiences by retrieving information from memories sampled from the outside world to form hypotheses of the population and determine the likelihood that the conclusion reached is correct. This model claims that confidence is approximately the assessment of

> “reproducibility – the likelihood that the same choice will be made in a subsequent encounter with the item. Although confidence judgments are construed as pertaining to validity – the probability that the chosen answer is correct, they are actually based on cues about reliability.”

The primary idea of this approach is that confidence is the feedback and sampling of information. Thus, a confidence judgment based on this model would draw from an individual's number of experiences, the "richness" of the experience, cognitive factors, bias, and the environment outlined by PMM (Gigerenzer et al., 1991). Notably, an individual would reduce overconfidence bias by increasing the sample size or the richness of the feedback from experiences (e.g., "I am less overconfident because I’ve done this before and gotten similar outcomes" or "I often do well on problems like this").

Koriat et al. (2020) demonstrated the SCM empirically by having participants answer true/false to personality statements (e.g., "I hate to change my plans at the last minute") and measuring deliberation time. After making a true/false judgment, participants then made a confidence judgment on a scale ranging from 0-100 (0% – Not confident at all; 100% – Completely confident). Results indicated that individuals responded faster and were less overconfident for repeated questions. Results also indicated that participants had higher and faster confidence judgments for items that other participants agreed on. Koriat believed that this reflected their personal experiences (perceptional decisions, social beliefs, social attitudes, personal preferences, and category membership decisions; see Koriat et al., 2016). For example, individuals gave faster and higher confidence judgments for questions more in line with their political thinking ("conservative" or "liberal"). These results indicated that confidence judgments were based not only on knowledge but also on personal experiences.

In sum, theories and models of overconfidence indicate that humans build knowledge-based mental models recalled from cues to help make decisions, evident from the PMM theory presented by Gigerenzer et al. (1991) and the SCM constructed by Koriat (2012). Still, growing evidence seems to indicate that not all confidence judgments are the same. It may be that different mental models are being constructed and perceived based on the type of overconfidence bias. Moore and Healy's 2008 article "The Trouble with Overconfidence" taxonomizes overconfidence bias into three distinct types: overestimation, overprecision, and overplacement.

2.3 The Three Types of Overconfidence

Recent advances in confidence research have focused on the measurement and taxonomy of confidence. Moore and Healy's 2008 paper "The Trouble with Overconfidence" classifies overconfidence into three distinct types (see Olsson, 2014). These three types of overconfidence are theoretically different, with separate measurements for each—table 1 breaks down each of the three types. The first type is the *overestimation* of one's actual ability, performance, level of control, or chance of success. Overestimation made up approximately 60% of the 645 empirical studies on overconfidence (numbers from Moore & Schatz, 2017).

Studies about overestimation often have participants rate their general abilities, skills, or forecasting concerning their domain-specific occupation or talent. For example, Garcia-Retamero et al. (2016) created a subjective graph literacy scale that asked individuals questions about their general skill at working with different types of graphs (e.g., "How good are you at working with bar charts?") and had participants rate their skill on a scale from 1 – *Not good at all* to 6 – *Extremely good*. These subjective ratings were then compared to their objective graph literacy. In short, overestimation asks a general confidence question on a skill.

The second type of overconfidence is *overplacement,* otherwise known as the *better-than-average* phenomenon in which people falsely believe themselves to be better than others. This type of confidence made up approximately 21% of empirical studies (this percentage goes up significantly when "better-than-average" is a search term along with "overconfidence"; Moore & Healy, 2008). Kruger and Dunning's 1999 study examined this type of overconfidence by having participants compare their score on the Law School Admissions Test to the average student's score on a scale from 0 ("I am at the very bottom of percentile rankings") to 99 ("I am at the very top"), with 50 being "I am exactly average." A more recent example of overplacement research is the study by Larrick et al. (2007), in which students had to judge their test performance in relation to given information on the mean, range, and distribution of performers. Each student then estimated how many students they were better than, on average, followed by an overestimation confidence judgment. Results indicated that a percentage of students did falsely believe themselves to be better than average. Furthermore, there was a positive within-subjects correlation between overplacement and overestimation. Moore (2007; see also Moore & Small, 2008) found that, generally, some individuals may be better at knowing themselves, but not others. In this case, individuals will not display overestimation but will show underplacement (i.e., worse-than-average). Note that most overconfidence studies do not measure more than one type of overconfidence, and even fewer measure all three (Olsson, 2014).

One way of measuring overplacement in decision making is the *Bias Blind Spot* (Pronin et al., 2002; Pronin & Kugler, 2007; Scopelliti et al., 2015; Ybarra, 2018, in-prep). The bias blind spot is an overplacement judgment wherein individuals assess their susceptibility to various decision biases and then judge others' susceptibility to each of the biases. Then the individual's bias judgment is subtracted from the other's bias judgment for each bias. The difference score for each bias is then aggregated to estimate how blind individuals are to their biases.

The third type of overconfidence is *overprecision*, or excessive certainty regarding the accuracy of one's belief, which made up 19% of empirical studies. Overprecision often compares an individual's estimated number of correct answers on a task to the actual number of correct answers. Adams and Adams (1961) created a specific way to measure overprecision. Participants were asked their confidence on a full scale from 0-100% in their ability to recall nonsense syllables. The confidence scale was defined to the participant as "the percent of questions believed to be correctly answered" (i.e., 90% means that the participant believed they were correct on 90% of the questions). A difference score was then calculated by subtracting the number of questions answered correctly from the percentage of questions the participant believed to be correct, with the remaining score being the participant's overconfidence (e.g., if a participant thought they answered 75% of questions correctly but scored 50% correct, the participant is thought to be overconfident by a degree of 25%). Often when measuring this type of overconfidence, it is translated into exact integers for clarity (e.g., "how many out of the ten asked questions did you answer correctly?").

To further learn about the three types of overconfidence, see Don Moore's (2020) book *Perfectly Confident*. Moore maintains that there are three types of overconfidence. However, Olsson (2014) states, "It is currently unknown to what extent these three different forms of overconfidence represent the same psychological construct, as only a handful of studies have investigated two or more ways of measuring overconfidence." Essentially, there may be three distinct types of overconfidence, but the evidence is sparse (Olsson lists only two studies that measured all three types of overconfidence). Further, what does this mean regarding training overconfidence? Does this mean that to lower various types of overconfidence bias, there needs to be different trainings? Unfortunately, there are currently zero training studies (to my reading) that compare all three types of overconfidence.

Table 1

Three Types of Confidence

Confidence	Definition	Typical Measurement	Example	Articles
Overestimation	Overrating your actual ability, skill, level of control, or chance of success.	Calibration = "Subjective confidence measure" – "Objective measure"	**Self-Evaluation**: "I am great at doing statistics" **Reality**: Low skill in statistics	Garvia-Retamero et al. (2016); Koriat et al. (2020)
Overplacement	Believing that you are better than others unjustly (better-than-average)	Calibration = "Subjective confidence measure compared to others" – "Objective measure"	**Self-Evaluation**: "I am better than average compared to my classmates in statistics" **Reality**: Below average grade in statistics	Kruger and Dunning (1999); Scopelliti et al., (2015)
Overprecision	Too sure you have the correct answer when wrong	Calibration = "Subjective estimation of the number of questions answered correctly" – "Objective number of questions answered correctly"	**Self-Evaluation**: "I am going to get 90 out of 100 questions correct on this statistics test" **Reality**: Gets 50 out of 100 questions correct on a statistics test	Gigerenzer et al. (1993); Adams and Adams (1961)

2.4 Training Overconfidence

In his 2011 book *Thinking Fast and Slow,* Daniel Kahneman concluded the section on overconfidence by saying, “Can overconfident optimism be overcome by training? I am not optimistic.” Continued, Kahneman wrote, “There have been numerous attempts to train people… with only a few reports of modest success.” Though results overall have been mixed, there are a few successful examples of training overconfidence. For instance, Royal Dutch Shell (commonly known as just Shell in the U.S.) oil and gas noticed that their newly hired geologists were often wrong and overconfident in predicting which wells would produce oil (Russo & Schoemaker, 1992). Shell then created a training program to correct overconfidence, ultimately saving time and money. The training involved studying past cases, making confidence judgments, and finally receiving feedback on their confidence judgments’ accuracy. The training tended to double the number of correct wells to be drilled (2/10 wells to 4/10 wells being correct on average).

Overconfidence is perhaps trainable in a highly specified domain, but can overconfidence be trained for general decisions? One important and novel example by Morewedge et al. (2015) was part of the Intelligence Advanced Research Projects Activity’s (IARPA) SIRIUS game project. The Morewedge et al. IARPA game had individuals solve the mystery of a missing individual by coming to terms with their own social biases. One of the most improved biases was overplacement measured via the bias blind spot, showing a reduction of approximately one standard deviation (F (1, 241) = 151.66, $p < .001$, $d_{\text{pre-post}} = .98$). While the results look promising, there is no evidence that the training would reduce the other two types of overconfidence.

Further, in a related study by Scopelliti et al. (2015) that developed the bias blind spot measure, they found that the bias blind spot is unrelated to general decision making skills. These results replicate conclusions by Dunning et al. (2004) that people's self-views tend to be weakly related to actual behavior and performance. Why is it that individuals skilled and knowledgeable in decision making are more/still overconfident? Unfortunately, there have been no studies that train all three types of overconfidence and measure generalization to wide-ranging decision making skills (e.g., risk literacy) that could help answer these questions.

2.5 Statistical Numeracy, Risk Literacy, and Skilled Decision Theory

One definition of statistical numeracy is the practical understanding of probabilistic and statistical problem solving (Cokely et al., 2012). Statistical numeracy tends to be one of the best predictors of superior judgment and general decision making (Cokely et al., 2018; Cokely & Kelley, 2009; Lipkus & Peters. 2009; Peters, 2020; Reyna et al., 2009; 2020). Numeracy tends to independently predict when other independent predictors of performance such as fluid intelligence, cognitive reflection, and attentional control are assessed (Allan, 2018). Research on numerical and non-numerical tasks indicated that statistical numeracy's predictive power goes beyond "just knowing the math." Numeracy tends to predict a cascade of metacognitive processes in which an individual evaluates and understands risk (Risk Literacy; see riskliteracy.org; Allan et al., 2017; Cokely et al., 2018; Petrova et al., 2017).

Risk literacy seems to be an essential component of skilled and informed decision making. In a study by Cokely et al. (forthcoming; Ghazal, 2014), structural models showed that risk literacy tended to mediate the relationship between statistical numeracy and general decision making performance. The demonstrated relationship between risk and decision making explains in part why statistical numeracy and risk literacy are both strong predictors of general decision making skills: decision making is fundamentally about reckoning risk and uncertainty (Cokely et al., 2018). If an individual understands probability theory and their personal limitations, skills, knowledge, and values, they should theoretically be able to make a good decision. Skilled Decision Theory outlines the reasons for this.

Skilled Decision Theory overviews the causal mechanisms for expert and skilled decision making in general (Cokely et al., 2018). The two primary mechanisms are 1) heuristic deliberation (e.g., metacognitive processing, Cokely & Kelley, 2009; Cokely et al., 2012) and 2) representative understanding (e.g., precise risk comprehension, Cokely et al., 2018). The first mechanism predicts skilled decision making via practical inductive reasoning skills (e.g., statistical numeracy and metacognition savviness) and elaborative heuristic deliberation. Both components of the mechanism theoretically help circumvent costly mistakes by employing metacognitive heuristics like double-checking an answer (Cokely et al., 2012; Ghazal et al., 2014). The second mechanism allows for the generation of representative understanding of the decision problem, allowing an individual to intuitively judge the weight and potential consequences of various options and outcomes (Peters, 2012; Petrova et al., 2015; Petrova et al., 2014). This understanding then informs the selection of adaptive heuristic strategies. For example, an individual potentially forms ecologically rational representations of cue validities and cue orderings that calibrate fast and frugal heuristic use to solve a decision task (Gigerenzer et al., 1999; Gigerenzer & Goldstein, 1996). What do the mechanisms of Skilled Decision Theory mean for overconfidence bias?

2.6 Risk Literacy and Overconfidence

Research has demonstrated that risk literacy skills are predictive of reduced vulnerability to overconfidence (Ghazal, 2014; Ybarra et al., in-prep). The study by Ghazal et al. (2014) had highly educated individuals complete a statistical numeracy measure (The Berlin Numeracy Test; see Cokely et al., 2012) and an array of decision making tasks like financial choices or medical judgments. Numeracy tended to be a predictor of decision making skills. Further, when assessing overprecision on a paradigmatic medical decision making problem, there was a positive correlation between numeracy and calibration. Sixty-five percent of highly numerate individuals were calibrated to their performance compared to forty-eight percent of low numerate individuals. Results revealed a curvilinear pattern of accuracy to confidence, resembling results like that in the “Unskilled and Unaware” literature (Dunning & Kruger, 1999). The authors concluded that the results were in line with the theoretical construct of numeracy (Skilled Decision Theory) that includes “1) a practical understanding of numbers and mathematical procedures, and 2) the skills necessary for effective problem solving and self-regulated learning”, meaning that general decision skills may help reduce overconfidence.

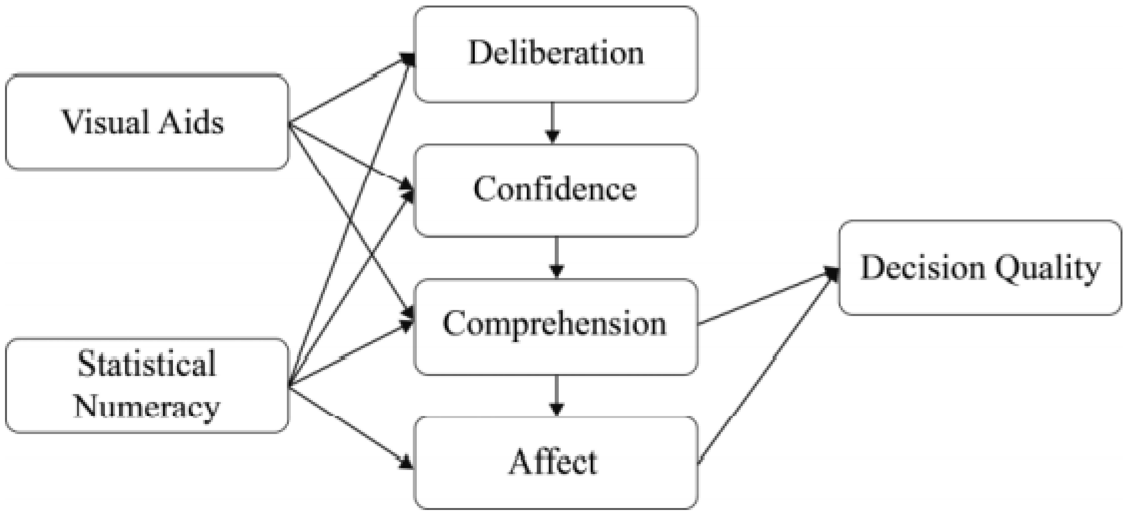
A later study conducted by Ybarra et al. (in-prep) found that statistically numerate individuals were less overconfident for both overestimation and overplacement. In the study, participants from around the world answered a general bias blind spot question in addition to a battery of decision making tasks. Results indicated that numerate individuals tended to perform better on the decision making tasks and were more calibrated to their performance. Results showed a curvilinear relationship between confidence and general decision making skills, meaning that those lowest in skill were the most overconfident, those with some amount of skill were slightly overconfident or calibrated, and those with the most skill were calibrated or slightly underconfident. Further, structure equation models revealed that decision making skills mediated the relationship between numeracy and confidence calibration. That risk literacy played an essential role in mitigating the overconfidence bias. Moreover, Ghazal et al. (2014) and Ybarra et al. (in-prep) point out that general decision making skills like risk literacy are trainable skills. Is it the case that training risk literacy could reduce overconfidence bias?

2.7 Improving Risk Literacy: Visual Aids and Training

Statistical numeracy tends to predict more resistance to a wide swath of decision making biases, including overconfidence bias. Unfortunately, a large proportion of individuals (including highly educated and intelligent working professionals) have relatively low levels of statistical numeracy (Garcia-Retamero & Cokely, 2017). Research has demonstrated that visual aids can help individuals better understand numerical expressions and probabilities (see Garcia-Retamero & Cokely, 2013, 2017). Figure 1 shows the general process model of skilled decision making and how visual aids and statistical numeracy relate. One study indicated that when participants completed a medical decision task that involved numerical information, highly numerate individuals could better solve the problem than less numerate individuals. However, providing less numerate individuals with a visual aid in addition to numerical information resulted in a similar performance to that of highly numerate individuals, increasing their accuracy to about 60%. Further, in a 2015 study by Garcia-Retamero et al., participants had to make confidence judgments about their performance on medical decision tasks with or without a visual aid. Participants given a visual aid tended to perform better on the tasks and were also less overconfident in their performance, no matter their statistical numeracy. People generally saw improvement from visual aids if the person had minimal graph interpretation skills (i.e., graph literacy).

Figure 1

Figure from Cokely et al. (2018)



Note. The generalized structural process model of skilled decision making from Cokely et al. (2018).

Some researchers have defined graph literacy as the ability to understand graphically presented information (Galesic & Garcia-Retamero, 2011). Graph literacy involves interpreting visual aids like bar charts, pie charts, line plots, decision trees, and icon arrays. Unfortunately, generally effective graphical aids may still be challenging for many people to understand. An estimated 15% of U.S. adults, or about 40 million, roughly equivalent to the population of California, misinterpret features like the height of simple bar charts (Galesic & Garcia-Retamero, 2011; see also: Okan et al., 2012, 2018; Terry & Ybarra, 2020), which means that tens of millions of potentially vulnerable individuals cannot understand important risk information from visual aids created to improve decisions related to high-stakes medical, educational, financial, and natural hazards.

To help individuals understand risk literacy an online tutor was constructed that trained important general decision making skills via graph literacy (Woller-Carter, 2015). The relatively brief (approximately 2-hour) training helped people understand and interpret graphs like the ones used for visual aids. A controlled experiment indicated that participants improved in graph literacy by about one standard deviation after taking the tutor, while participants in the control condition saw no improvements. Further, individuals who completed the risk literacy tutor also improved resistance to decision biases like framing effects, ratio bias, and sunk costs. To emphasize, after individuals completed the risk tutor, people improved in decision making for tasks that did not include any visual aid. The question remains whether participants who took the tutor also built resistance to the overconfidence bias.

To summarize, avoiding overconfidence is essential for behaviors and processes related to decision making, skill attainment, education, ethics, and self-regulation. Efforts to reduce overconfidence have led to mixed results, and some researchers believe that training overconfidence may not be possible (Kahneman, 2011). Additionally, there is evidence that overconfidence is not improved by nor related to general skills, experience, or knowledge (Dunning et al., 2004, Scopelliti et al., 2015). However, these efforts tended not to use measures related to risk literacy skills. Evidence has shown that risk literacy is a trainable skill predictive of better decision making and less overconfidence bias. Can causally training risk literacy skills lead to better self-evaluations?

Chapter 3: Methods

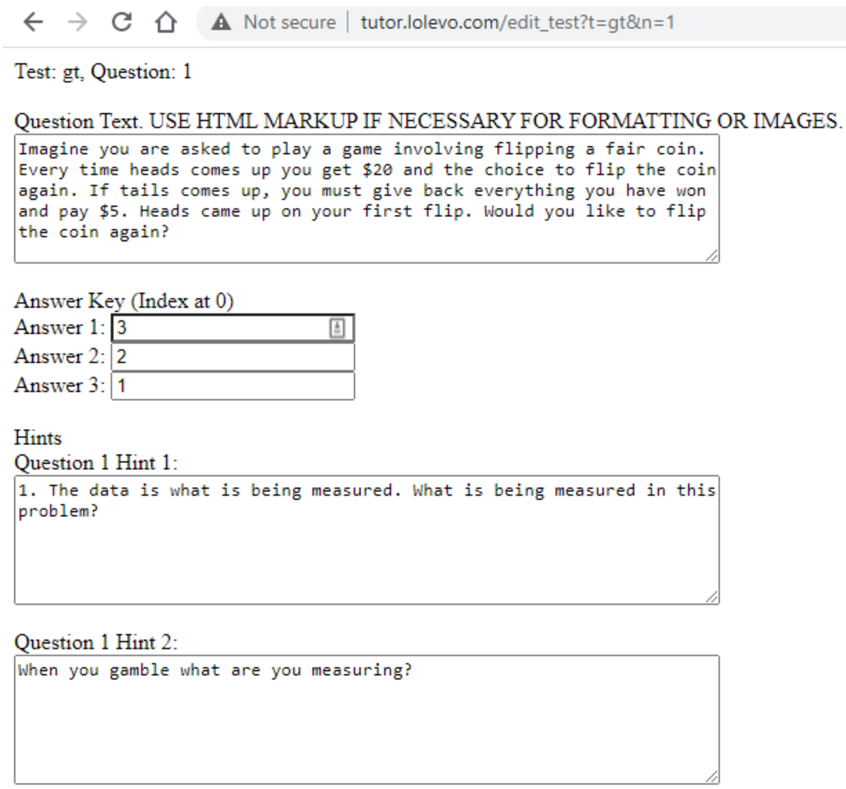
3.1 Tutor Construction

The tutor construction and reconstruction were part of an NSF-funded project designed to assess and improve general decision making skills. Previously, the team at riskliteracy.org created a suite of brief online risk literacy and risk communication training modules intended for use by diverse adults from industrialized countries. The first completed module, validated via scientific control trials, is a graph literacy training program (for a complete review, see Woller-Carter, 2015). The initial tutor's construction followed well-established standards in intelligent tutoring and was programmed using Adobe Flash and Carnegie Mellon's Cognitive Tutor Authoring Tools (CTAT) (Aleven et al., 2006, 2009). The same principles and standards were applied during the tutor's reconstruction, though it was rebuilt from the ground up in Python rather than in Flash and CTAT (tutor.lolevo.com; Ybarra et al., 2018).

One primary goal in renovating the risk tutor was to increase the graph tutor's flexibility without designing wholly new projects. The tutor was programmed in Python using Flask as the web framework to achieve desired flexibility and reduce programming time. There are many potential benefits of using Python and Flask. Flask is a web framework that does not require specific libraries for form validation, database connectivity, or other relevant tools associated with developing and using a web application. Flask allowed for a lightweight backend that could be customized easily compared to heftier web frameworks like Django. One outcome of these advantages was the created browser-based question editor that allowed instantaneous changes to the tutor questions and associated hints (Figure 2).

Figure 2

In-Browser Question Editor



Note. The created in-browser question editor. This tool can be accessed from any online browser. The editor allows for quick question and hint editing. The editor also paved the way for an in-browser tool to allow anyone to make online risk interventions. This tool is still in development.

3.2 Procedure

The design of the experiment was a one-way independent samples design.[1] Random assignment placed all participants in either the control or tutor condition. Both conditions consisted of two phases, the first taken in the laboratory and the second completed at home. All pre-testing and training were in the lab and lasted approximately three hours. The tutor condition had selected

[1] Due to programming issues and significant attrition of participants from noncompliance, the originally conducted third condition was excluded. For more information see appendix.

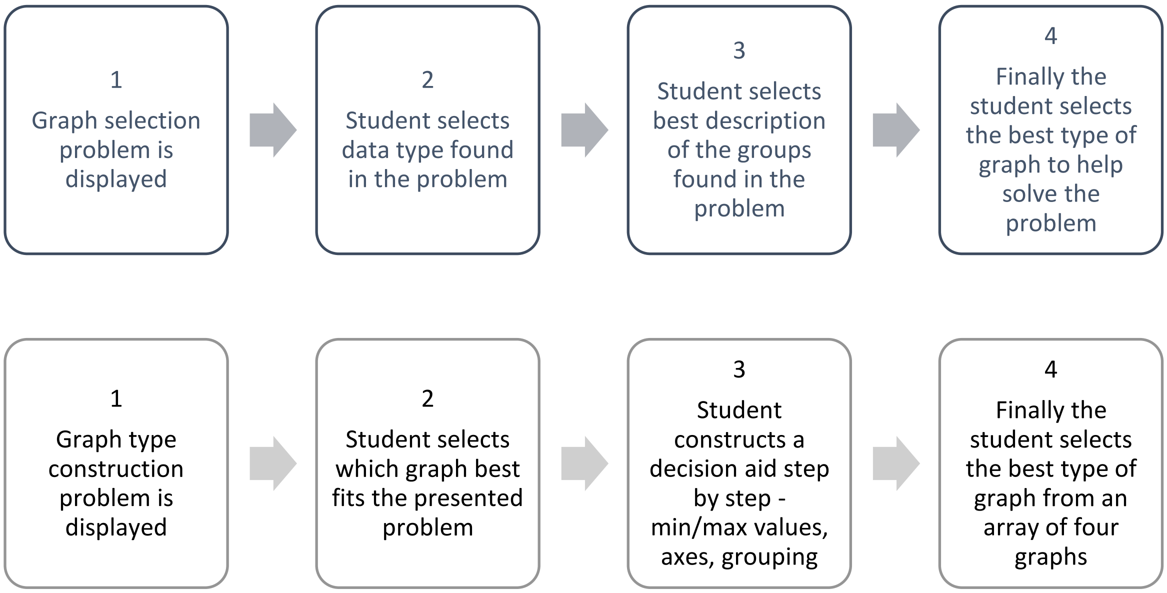
post-test measures completed immediately after training. Phase 2 was completed approximately 2-4 weeks after Phase 1. The participants were sent the link for Phase 2 two weeks after Phase 1 but could complete the post-test at their own pace. Instructions asked participants to complete Phase 2 in one sitting after opening the link. Debriefing occurred after Phase 2.

The control condition was a financial literacy video tutor series that educated participants on long-term investing, retirement savings, and financial decision making. The videos were all from khanacademy.org's investments and retirement unit. Selection criteria for the control condition content included proven relevance to future life outcomes (e.g., healthy retirement) and domain-specific elements related to decision making and numeracy. In other words, the control was not to interfere with training general risk literacy skills but needed to be one of those domains that the risk literacy tutor could improve (Skagerlund et al., 2018).

The risk tutor condition was the online risk literacy tutor aimed at training specific risk literacy skills (e.g., graph literacy). The participant first completed a pre-test that asked for necessary graph information (e.g., pie charts add up to 100%). Next, the participant completed two different graph training modules. The first module of the tutor trained participants on interpreting data by displaying a problem involving probability that could be visualized and having the participant choose the appropriate visual aid. The second had participants construct a visual aid for several types of probabilistic problems and then choose the best constructed visual aid from an array of graphs. At each stage of the tutor, the participant had to select the correct answer to proceed. Exhaustive hints were available for the participants at their request; for a detailed procedure, see Figure 3.

Figure 3

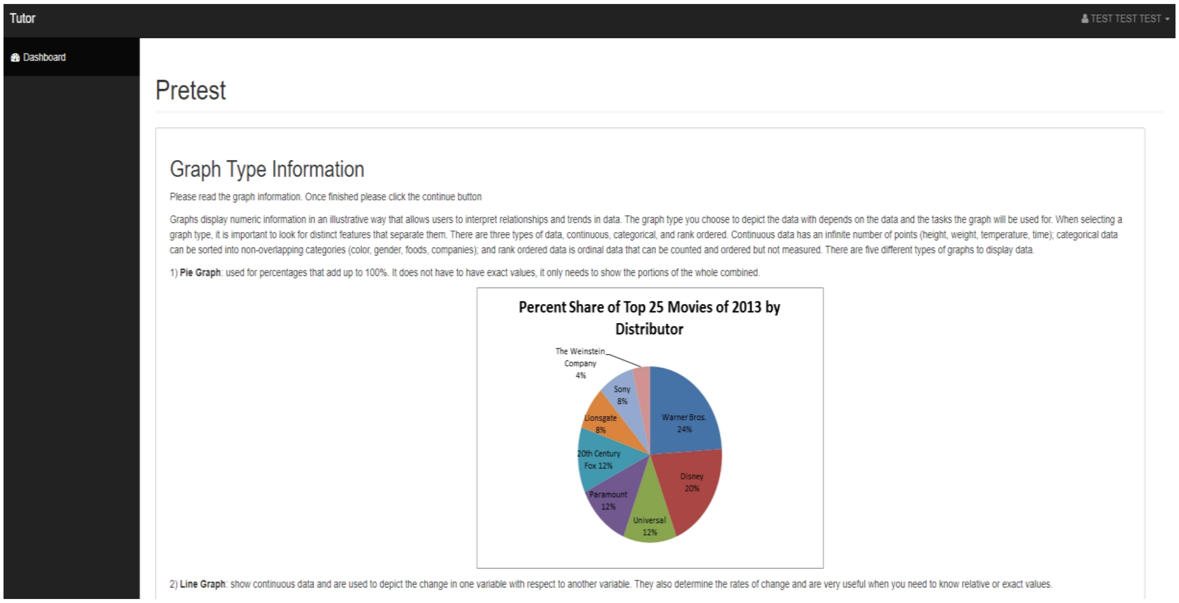
Procedure for the Tutor Condition



Note. The top blue flowchart shows how students solved the first type of problem presented in the risk tutor. This part of the tutor emphasized identifying the type of data in the presented problem (e.g., continuous vs. categorical) and having the participant match that to the correct graph. The bottom green flowchart shows how students solved the second type of problem, which emphasized constructing graphs and identifying poorly created graphs.

Figure 4

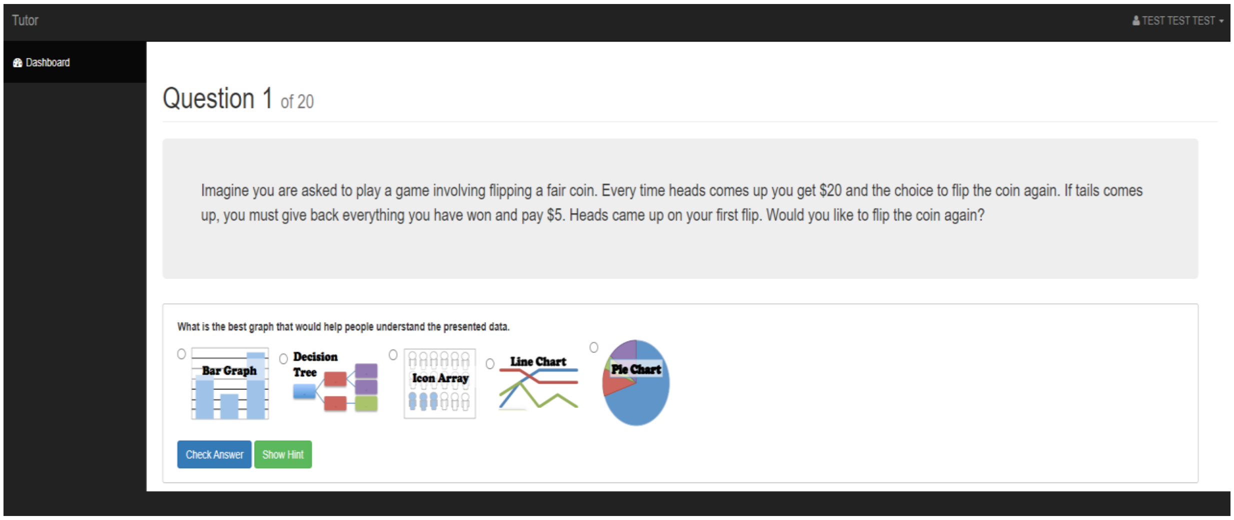
Initial Graph Overview



Note. Screenshot of the initial graph overview. The pre-test explained each graph type, and then base knowledge was assessed in a multiple-choice format (e.g., "What type of graph uses percentages that add up to 100%?").

Figure 5

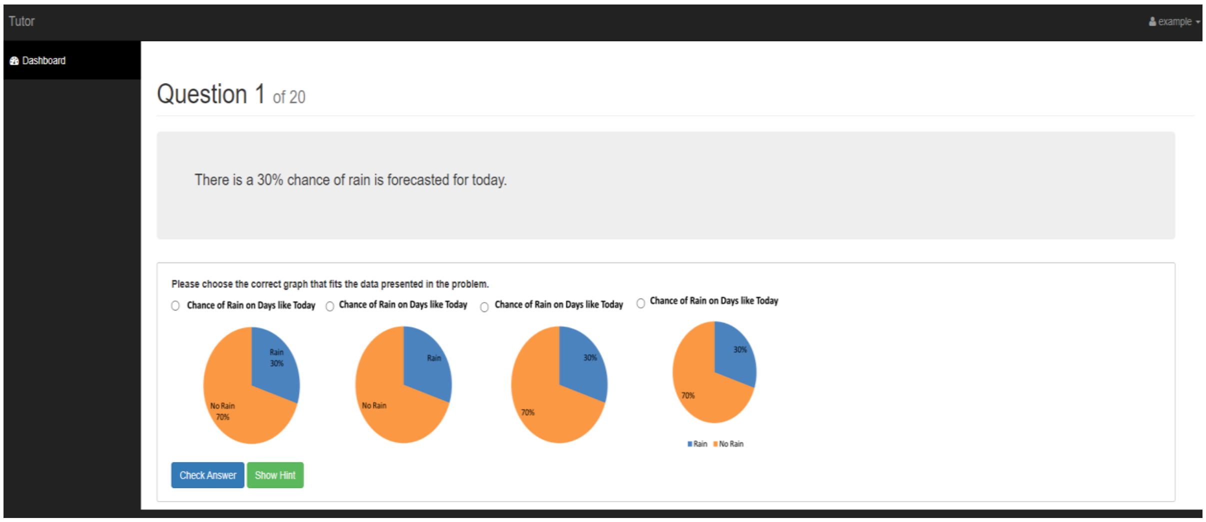
First Module of Risk Tutor



Note. Screenshot of a problem from the first module of the risk literacy tutor. This module had individuals interpret and classify data found in a decision making problem. The participant then used that information to choose an appropriate visual aid to help solve the problem.

Figure 6

Second Module of Risk Tutor



Note. Screenshot of the second module of the risk literacy tutor. This module had participants construct a visual aid for a decision making problem by choosing the graph's values (e.g., axis value, pie chart values). Finally, the participant selected the best constructed visual aid from an array of visual aids, all similar.

3.3 Pre-test measures

Berlin Numeracy Test-S. Statistical numeracy was measured using the Berlin Numeracy Test (see RiskLiteracy.org; Cokely et al., 2012). Following best practice recommendations, the non-adaptive, 7-item Berlin Numeracy Test-S (BNT-S) – which includes three questions from Schwartz et al. (1997) – was used to increase sensitivity measuring those who may be less skilled. One example of a question includes, "Imagine we are throwing a five-sided die 50 times... out of 50 throws, what proportion will result in an odd number?"

Subjective Numeracy Scale. Fagerlin and colleagues (2007) created an 8-item scale that asked participants to rate their skills working with numeric information subjectively (e.g., "How good are you at working with fractions?") and their preferences for risk information (e.g., "When reading the newspaper, how helpful do you find tables and graphs that are parts of a story?"). The measure often has correlations with objective numeracy (of varying sizes) and is sometimes used as a substitute for objective measures. The pre-test observed participants' proclivity to be overconfident in (specifically to overestimate) their numerical and risk literacy skills.

Graph Construction Knowledge. The Graph Construction Knowledge Scale attempted to create a graph literacy pre-test focused on assessing the correct use of graphs in different situations. Item construction used research on graph literacy and appropriate construction of visual aids (Galesic & Garcia-Retamero, 2011; Garcia-Retamero & Cokely, 2017). Questions aimed to measured knowledge related to the correct application of specific types of graphs to different situations (e.g., "When using a bar graph which situation/goal is most appropriate for its use?").

Subjective Graph Literacy. Garcia-Retamero et al. (2016) created the Subjective Graph Literacy scale to robustly predict objective graph literacy and quickly assess risk communication preferences. Questions asked participants subjectively how skilled they are with various types of graphs (e.g., "How good are you at working with bar charts?"). The scale's inclusion in the pre-test allowed for observation of participants' proclivity to be overconfident in their graph literacy. This measure was also used in the post-test to see if an individual's subjective skill in using graphs changed after completing the graph literacy tutor.

3.4 Post-test measures

Graph Literacy Scale. The Graph Literacy Scale developed by Galesic and Garcia-Retamero (2011) consisted of 12 questions that presented participants with a visual aid and had them 1) read the data, 2) read between the data (compare parts of the visual aid), and 3) read beyond the data (make inferences from the visual aid). As detailed in Woller-Carter (2015), an additional four questions from various sources were asked to add sensitivity for those skilled in graph literacy due to the use of an educated sample.

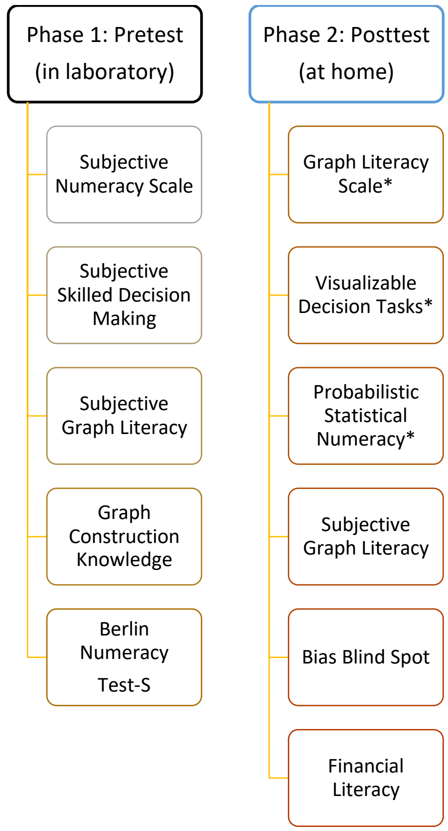
Visualizable Decision Tasks. The tasks from Ybarra et al. (2018) measured three different biases that are theoretically related to graph literacy skills, according to the results of Woller-Carter (2015). The three biases measured were framing effects, sunk cost bias, and ratio bias. Overestimation confidence questions ("How confident are you that you are correct in your answers?") were included for each of the three biases and used a 1 (not confident at all) to 10 (completely confident) Likert scale (Ybarra, 2018). Overprecision confidence questions ("Out of the five problems, how many did you answer correctly?") were included for the sunk cost and ratio bias questions (Ybarra, 2018).

Financial Literacy. The financial literacy questions created by Lusardi and Mitchell (2007) measure financial knowledge related to economic behaviors like saving for retirement. The measure included assessing basic math skills related to investing, such as calculating the amount accumulated from interest rates. The measure also included more sophisticated financial literacy measures from Lusardi et al. (2014) that assess specific knowledge related to investing, such as knowing that stocks are generally riskier than bonds.

Bias Blind Spot. Twenty-three questions assessing bias blind spot consisted of fourteen questions constructed by Scopelliti et al. (2015) and nine questions created by Ybarra (2018). Ybarra (2018) found that the Bias Blind Spot scale created by Scopelliti et al. (2015) was a measure of social bias overconfidence (specifically overplacement) that was independent of more cognitive items. Question structure consisted of describing what the bias is, then measured an individual's propensity to believe themselves less biased than others by asking, "How biased are you?" and "How biased is the average individual?", then having participants rate each question on a Likert scale from 1 (*not at all*) to 7 (*very much*) (Scopelliti et al., 2015). The measure was scored by taking the difference between the subjective rating of average individuals' bias and subtracting the subjective rating of one's own bias. Next, the differences for the three biases were averaged to give an individual's final bias blind spot score indicative of an individual's potential vulnerability to decision making overconfidence.

Figure 7

Pre- and Posttest Measures.



Note. A chart of all pre-test and post-test measures.[2]

3.5 Participants

A total of 273 University of Oklahoma students completed a 4-hour experiment in exchange for class credit (66% Female; age ranged from 18-22 years old). The final sample is derived from individuals who completed both the in-lab and at-home portions, completed the full tutor, and completed over 50% of the at-home portion.

[2] All scales marked with an * indicate that in the tutor condition these measures were given immediately after the tutor.

Chapter 4: Analyses and Results

4.1 Pre-test Results

The purpose of this first analysis is to determine if there exists any significant difference in pre-test measures between conditions. The sample showed average statistical numeracy levels compared to previous studies for college-aged participants (Table 2; Allan, 2018; Ybarra, 2018). On average, participants answered 3 out of 7 numeracy questions correctly. A one-way ANOVA found no significant differences between all conditions for all pre-test measures.

Table 2

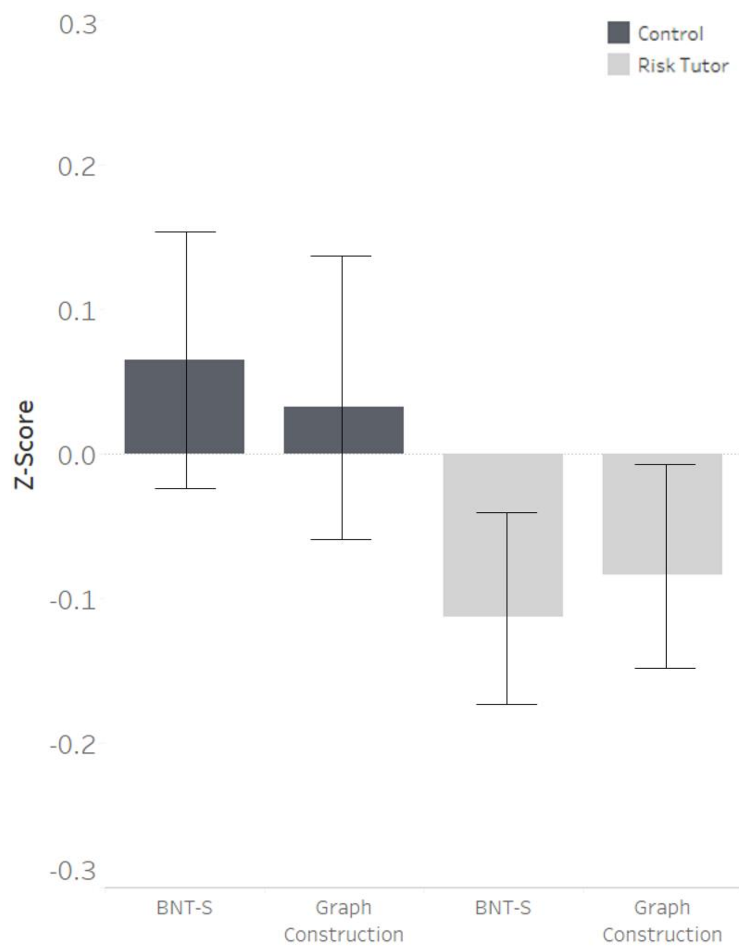
Pre-test Measures

		Risk Tutor (n = 138)	Control (n = 135)
	Maximum	Mean (SD)	Mean (SD)
Berlin Numeracy Test - S	7	3.08 (1.41)	3.36 (1.77)
Graph Construction Knowledge	11	4.46 (1.63)	4.66 (1.80)
Subjective Graph Literacy	6	4.34 (0.98)	4.45 (0.85)

Note. Descriptive statistics for pre-test measures by condition

Figure 8

Pretest Measures



Note. Shows bar charts of pre-test measures. There are no significant differences between groups. Measures are z-scored and the error bars represent standard error.

4.2 Post-test Results

The purpose of the following analyses is to determine if there is any significant difference in post-test measures. Table 2 shows the means and standard deviations for post-test measures. First, a one-way ANOVA analysis revealed no significant differences for basic financial literacy, though there was a significant difference in sophisticated financial literacy where the control condition performed better (F (1, 272) = 7.79, p = .006; Figure 9).

Next, group differences in graph literacy were assessed (Figure 10). A difference score of pre-and post-test graph literacy was computed using the difference in scores for both the graph construction scale and graph literacy scale at pre-and post-testing. There was a significant between-group difference for graph literacy, via one-way ANOVA, in that the risk tutor condition had significantly larger pre-post differences (F (1, 272) = 4.03, p = .046, d = .42). One-way ANCOVA results determined a statistically significant difference between the risk tutor and control conditions on the post-test Graph Literacy Scale controlling for statistical numeracy (BNT-S). There was a significant increase in Graph Literacy Scale scores after completing the risk tutor (F (2, 270) = 5.54, p = .019, R^2 = .27).

Group differences in visualizable decision making task performance were analyzed (Figure 11). Table 3 shows the means and standard deviations between conditions for post-test measures. A composite score was created by adding standardized values for ratio bias, framing effects, and sunk cost bias. A one-way ANCOVA determined a statistically significant difference between the risk tutor and control conditions on the visualizable decision making tasks controlling for statistical numeracy. There was a significant increase in performance on visualizable decision tasks after completing the risk tutor (F (2, 270) = 6.58, p = .011, R^2 = .13).

Table 3

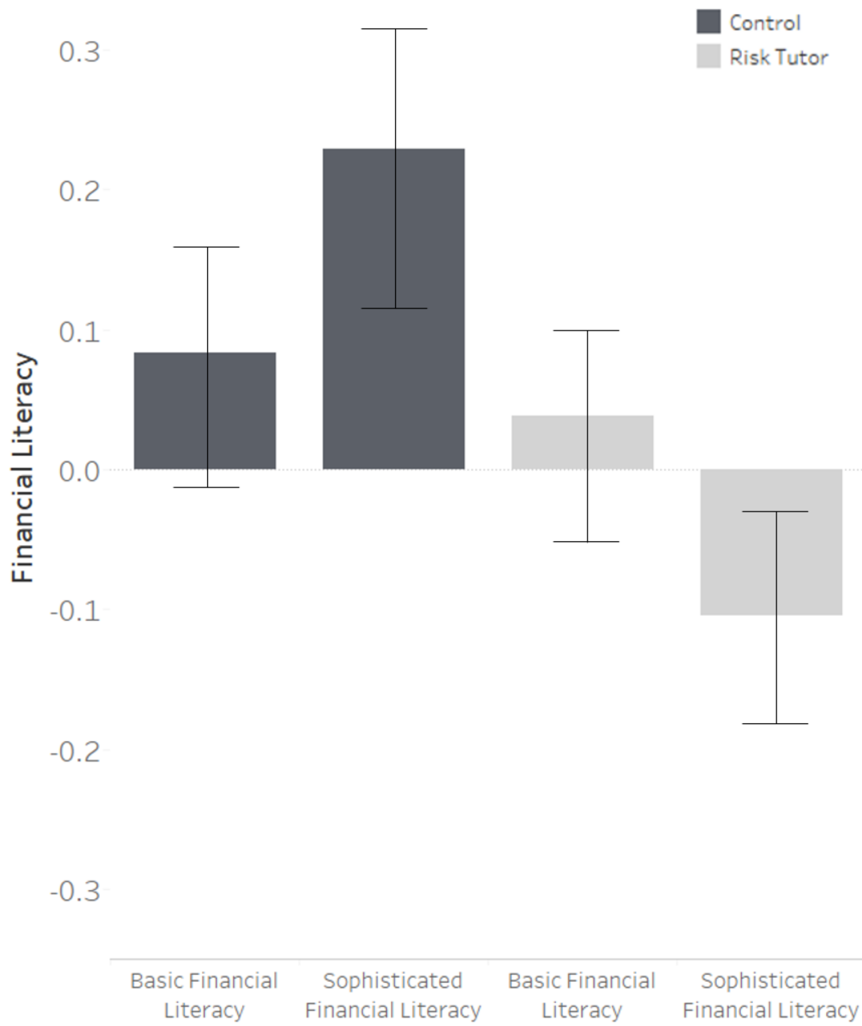
Post-test Measures Means and Standard Deviations

		Risk Tutor (*n* = 138)	Control (*n* = 135)
	Maximum	Mean (SD)	Mean (SD)
Berlin Numeracy Test - Components	9	3.08 (1.74)	3.50 (2.38)
Graph Literacy	16	12.40 (2.09)	12.05 (2.45)
Graph Literacy - Hard Items	4	2.68 (0.98)	2.47 (1.21)
Visualizable Decision Tasks			
Ratio Bias	5	2.93 (1.59)	2.77 (1.58)
Sunk Cost	5	3.18 (1.15)	3.04 (1.29)
Framing	3	1.35 (0.87)	1.13 (0.91)
Financial Literacy			
Basic	4	2.29 (1.20)	2.35 (1.31)
Sophisticated	8	2.66 (1.92)	3.31 (1.94)

Note. Descriptive statistics for post-test measures by condition.

Figure 9

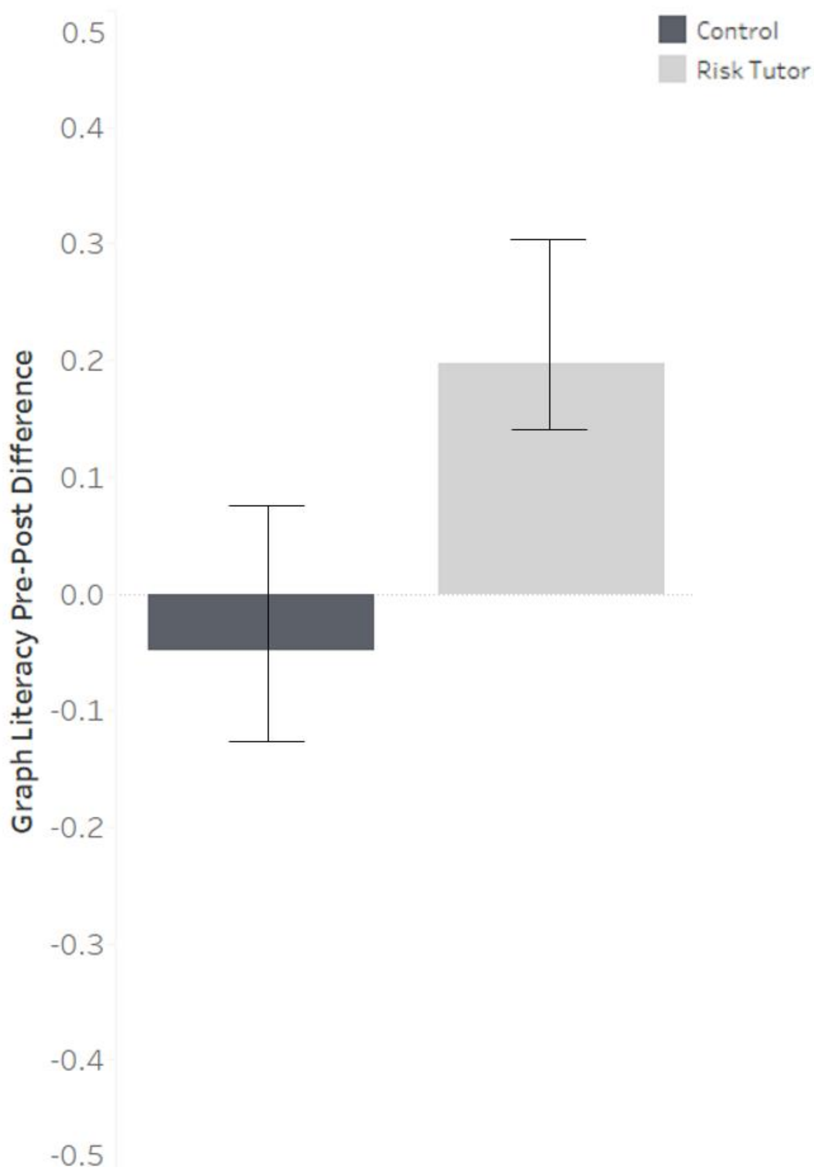
Condition Differences in Financial Literacy



Note. Posttest between-group differences on sophisticated financial literacy. There was a significant group mean difference for sophisticated financial literacy between the control condition and the risk tutor condition via one-way ANOVA (F (1, 272) = 7.79, p = .006). Participants in the control financial literacy condition tended to score better on sophisticated literacy questions, which asked questions about domain-specific financial knowledge (e.g., "What are riskier of the two? Stocks or bonds?"). Measures are z-scored and the error bars represent standard error.

Figure 10

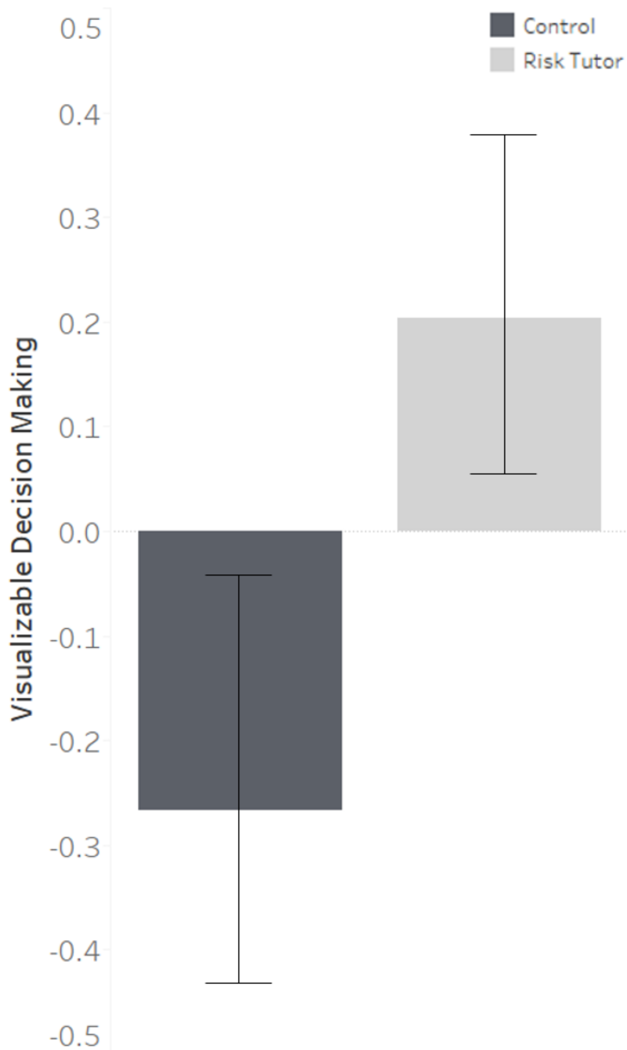
Condition Differences in Graph Literacy



Note. The image displays within-group pre-post differences between conditions on the pre-test graph construction measure and the post-test Graph Literacy Scale. There was a significant group mean difference for changes in pre-post graph literacy. Results indicated that participants who completed the two-hour risk literacy tutor causally improved in graph literacy compared to the control condition. Measures are z-scored and the error bars represent standard error

Figure 11

Condition Differences in Visualizable Decision Making



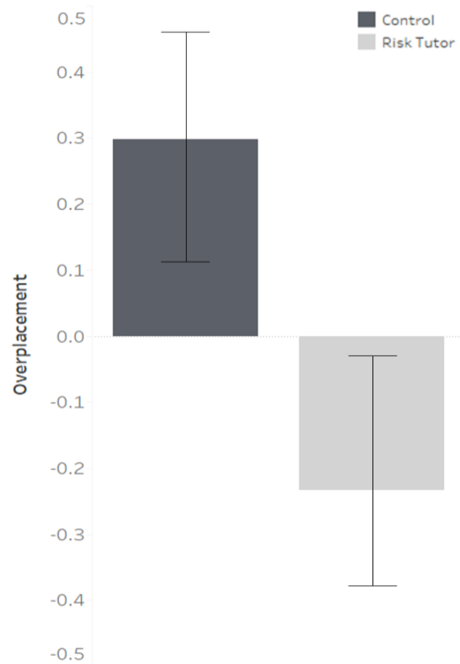
Note. The graph shows group post-test differences between conditions on visualizable decision tasks. Results indicated that participants who completed the risk tutor performed better on the visualizable decision making tasks (Sunk cost, ratio bias, and framing effects) than participants in the control condition. Results indicated that training general risk literacy skills improve untrained decision making skills.

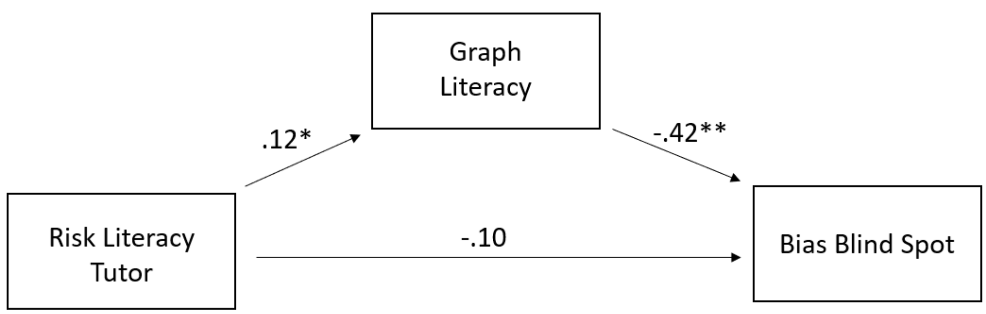
4.3 Post-test confidence results

Finally, group differences for the three types of overconfidence bias were analyzed. Overplacement was first analyzed. A one-way ANCOVA found no statistically significant difference between the risk tutor and control conditions on the post-test bias blind spot measures for the visualizable decision making biases controlling for statistical numeracy (BNT-S). Ybarra et al. (in-prep) showed that better decision making performance was positively related to the bias blind spot. A similar multiple regression path analysis was constructed to observe if the tutor led to differences in the bias blind spot (Figure 12). Path analyses indicated a significant positive relationship between visualizable decision making task performance and cognitive bias blind spot scores.

Figure 12

Risk Literacy Training and the Bias Blind Spot





Note. Top: There was a significant, but small effect for the differences in overplacement for the visualizable decision making tasks as seen in the bar chart. Bottom: A structural equation model estimated the amount that the tutor and risk literacy skills mediated the relationship, controlling for numeracy, with bias blind spot (e.g., overplacement for the visualizable decision making tasks). The model is in line with previous research with the bias blind spot and risk literacy (see Ybarra et al., in-prep) and has good fit $\chi 2$ (1) = 4.13, CFI = .97, TLI = .89, RMSEA = .10 (0.01 – 0.22), SRMR = 0.03, R^2 = .19.

[a]Within figure, * = $p < .05$, ** = $p < .01$.

A one-way ANCOVA determined a statistically significant difference between the risk tutor and control conditions on the post-test overestimation measures for the visualizable decision making tasks controlling for statistical numeracy (BNT-S). There was a significant decrease in posttest overestimation after completing the risk tutor controlling for pretest numeracy (F (2, 269) = 7.94, p = .005, R^2 = .03). A second one-way ANCOVA indicated a significant decrease in posttest overprecision for the sunk cost and ratio bias visualizable decision making tasks (F (2, 270) = 5.91, p = .016, R^2 = .02). Figure 13 shows a summary of group differences for overestimation and overprecision for both conditions.

Table 4

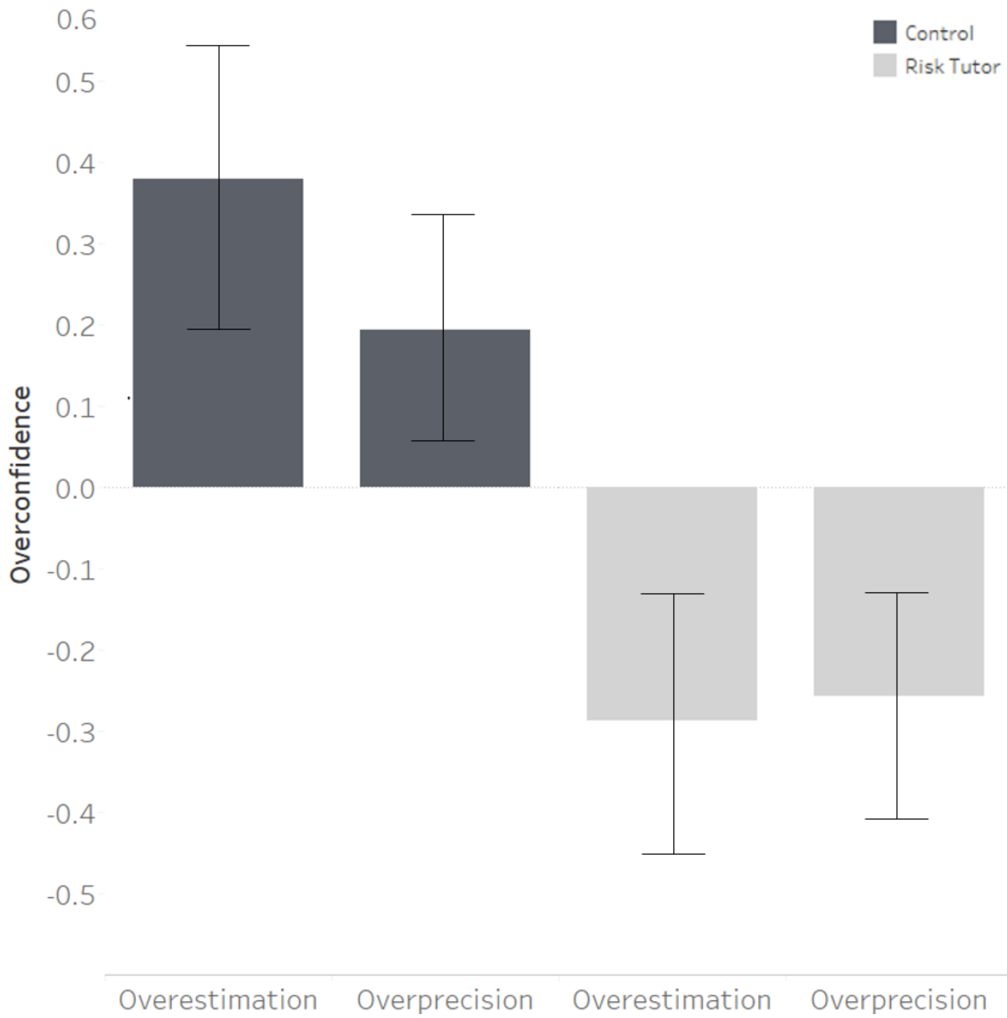
Overprecision and Overestimation on the Visualizable Decision Making Tasks

		Risk Tutor (n = 138)	Control (n = 135)
Overconfidence	Max.	Mean (SD)	Mean (SD)
Ratio Bias Overprecision	5	0.77 (1.42)	1.10 (1.71)
Ratio Bias Overestimation	10	0.99 (2.95)	1.59 (3.46)
Sunk Cost Overprecision	5	1.39 (1.85)	1.86 (1.88)
Sunk Cost Overestimation	10	1.87 (3.01)	2.21 (2.85)
Framing Overestimation	10	2.60 (2.88)	3.75 (3.08)

Note. Descriptive statistics for post-test visualizable decision making task confidence measures by condition.

Figure 13

Condition Differences in Overconfidence

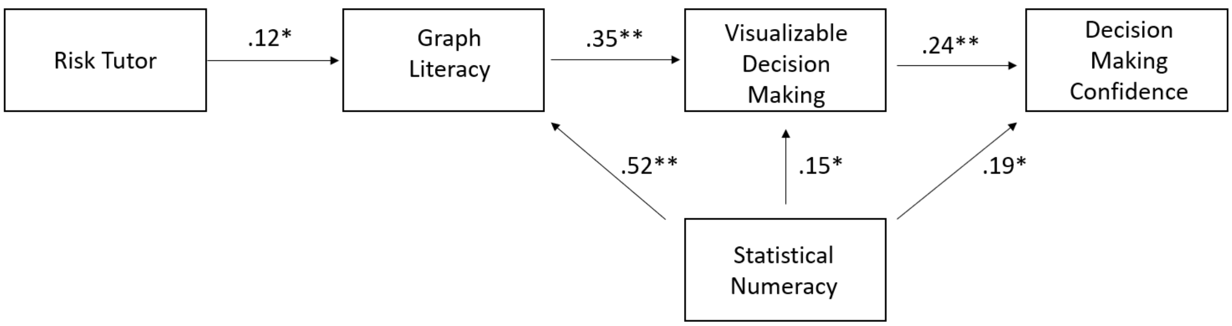


Note. The top image shows between-condition differences in overconfidence for the visualizable decision tasks. The left, middle, and bottom two images show the significant group mean differences between the control and the risk tutor conditions while controlling for statistical numeracy for overprecision and overestimation. The risk tutor tends to causally predict lower overconfidence for all visualizable decision biases (ratio bias, sunk cost, and framing). The right middle and bottom two images show overprecision and overestimation ratings across levels of statistical numeracy. The risk tutor condition shows predicted means of lower overconfidence for both overprecision and overestimation for all three distinct visualizable decision making tasks.

A path analysis evaluated how the risk tutor changes confidence judgments. The path analysis predicted confidence using the risk tutor, graph literacy, and objective visualizable decision skill as predictors while controlling for statistical numeracy (Figure 14 shows the whole model).

Figure 14

Path Analysis of How the Tutor Predicts Confidence Judgments



Note. The above structural equation model shows the tutor's effect on graph literacy, visualizable decision making task performance, and confidence judgments. The model showed good fit: $\chi 2$ (3) = 6.60, CFI = .98, TLI = .94, RMSEA = .06 (0.00 – 0.13), SRMR = 0.03. Those in the risk tutor condition tended to improve in graph literacy, $R^2 = .28$. Graph literacy (measured via the graph literacy scale) predicted better visualizable decision making task performance, $R^2 = .20$. Higher visualizable decision making task performance mediated higher confidence judgments ("How confident are you that you are correct?"), indicating better calibration, $R^2 = .13$. Statistical numeracy predicted higher graph literacy, visualizable task performance, and confidence. The model implies that training graph literacy leads to better decision making task performance and accurate confidence judgments. Further, general risk literacy skills are predictive of superior decision making performance but also better self-evaluations.[3]

[a]Within figure, * = $p < .05$, ** = $p < .01$.

[3] The model shown above was chosen for both theoretical and statistical reasons. Graph literacy must come after the risk tutor as it is the trained effect. It was believed from Skilled Decision Theory and previous studies (Woller-Carter, 2015) that untrained, objective decision skills would be improved by risk literacy skills like graph literacy. The improved skill then would, theoretically, then change self-evaluations. For competing models see Appendix D.

Chapter 5: Discussion and Conclusion

5.1 Discussion

The study presented here is the first experiment to test if training general risk literacy skills (e.g., graph literacy) helps individuals avoid the overconfidence bias on various decision making and self-evaluation tasks (framing effects, ratio bias, sunk cost, overconfidence bias). The results presented here suggested that training graph literacy skills causally improves performance on conceptually distinct decision making tasks that involve visualizable, graph-relevant elements. Individuals who complete training via the risk literacy tutor also tended to be less overconfident in their decision making task performance across many classes of self-evaluation judgments, consistent with Skilled Decision Theory (Cokely et al., 2018). Findings extend upon similar research by Morewedge et al. (2015) that showed improvements in one type of overconfidence, overplacement, via training specific decision biases (Symborski et al., 2014; 2016). Findings from the Morewedge et al. (2015) study and this study are inconsistent with previous results that appeared to suggest that overconfidence was untrainable and therefore was not likely to be related to individual differences in self-evaluation judgments more generally (Kahneman, 2011). In contrast, the current results of the first experimental test of risk literacy skills training revealed that general decision making skills are often robustly causally related to accurate metacognitive judgments (e.g., cognitive self-evaluations and social comparisons). In sum:

1. Risk literacy skills are self-evaluation skills,

2. Self-evaluation is a general, trainable skill, and

3. Risk literacy is a trainable skill that is causally related to general decision making skills.

5.2 Risk Literacy Skills Promote Accurate Self-Evaluations

How are individuals improving in their self-evaluations? One theory is that the tutor provides individuals with strategies and general decision making skills to understand better and evaluate the knowledge that helps inform self-evaluations. The tutor could help individuals create better problem representations in their long-term memory (Ghazal et al., 2014). Having a deeper representative understanding may help avoid biased intuitions, allowing for accurate cognitive monitoring (see Mitchum & Kelley, 2010; Cokely, Kelley, and Gilchrist, 2006). This process may result in the individuals having better strategy selection which is crucial for task performance (Gigerenzer et al., 1999). The better strategy could result in a more accurate confidence judgment as it theoretically indicates the validity of the chosen strategy as postulated by the probabilistic mental model theory (Gigerenzer et al., 1991).

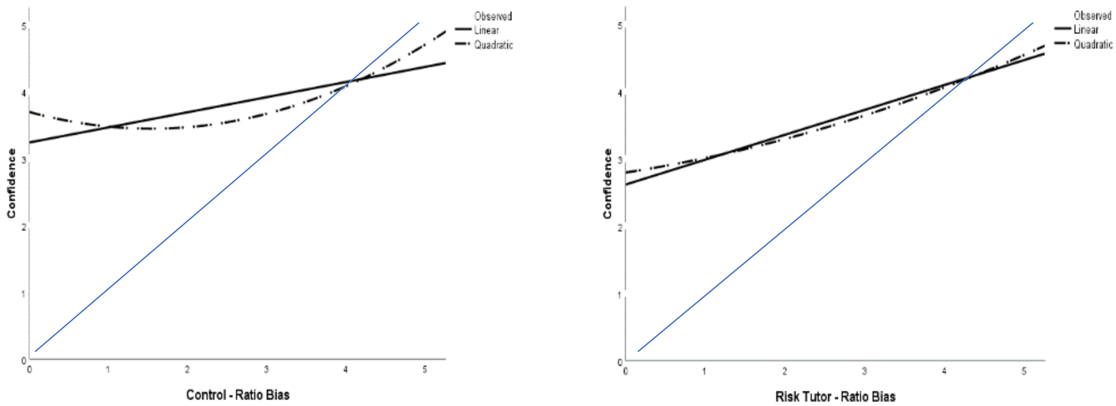
Another possible theory is that individuals incorporate feedback from the tutor into their self-evaluations (Arkes et al., 1987; Lichtenstein & Fischhoff, 1980; Sitzman, Rhodes, & Tauber, 2014). The tutor provides individuals feedback on their performance on problems that may share identical elements with the visualizable decision making tasks (framing effects, ratio bias, sunk cost). For instance, if a student is doing poorly on the tutor, and the tutor communicates this to the student, the student may use that feedback in their confidence judgment (e.g., "The way I am solving this task is wrong"; Koriat, 2012; Moore & Healy, 2008). Perhaps, though, the individual makes a personal judgment related to their overall skill (e.g., "I am bad at ratio bias problems"). Koriat (2012) claims that "confidence is modeled by analogy to the calculation of statistical level of confidence (SLC) in testing hypotheses about a population and represents the participant's assessment of the likelihood that a new sample will yield the same choice." Koriat claims that confidence monitors reliability, not validity, in choices which means that self-evaluations are sensitive to variance and number of experiences (i.e., sample size). In either theoretical case (the probabilistic mental model or the self-consistency model), self-evaluations could be based on more coherent representative understanding derived from the problem's cues via risk literacy skills. The feedback from the tutor may have enriched the experiences and knowledge used to make accurate confidence judgments. This claim assumes the tutor improved self-evaluation skill and not just decision making task performance.

5.3 Self-Evaluation is a General, Trainable Skill

One concern is that the resulting change in overconfidence is driven only by those who improve in task performance, like the effects found in the "Unskilled and Unaware" literature (Kruger & Dunning, 1999; Dunning, Johnson, Ehrlinger, & Kruger, 2003; Ehrlinger et al., 2008; Sanchez & Dunning, 2018). Looking closer at the data, individuals who took the risk literacy tutor tended to calibrate their confidence judgments across all skill levels (Figure 15). For example, confidence calibration curves for the ratio bias decision tasks showed a curvilinear relationship between confidence and accuracy in both conditions. Comparing the quadratic regression results for ratio bias showed that low-skilled individuals in the control condition were more overconfident than those in the risk tutor condition. The results imply that the lower-skilled individuals in the training condition seemed to be "Unskilled, *but aware*" (Miller & Geraci, 2011; Hartwig & Dunlosky, 2014). Calibration may be due to improved sensitivity to task performance as people generally improved in self-evaluations, but not actual performance, lending evidence to the idea of skilled self-evaluations (Cokely et al., 2018; Ericsson et al., 2007; 2018; Gigerenzer, 2015). The idea of trainable, skilled self-evaluations has many implications.

Figure 15

Control vs. Risk Tutor Quadratic Regressions Showing Confidence Calibration



Note. Two confidence calibration curves demonstrating the difference between conditions, where the blue line demonstrates perfect calibration. Overprecision is decreased for all individuals who completed the risk tutor including those lower in accuracy, as seen in the quadratic regression (F (1, 2) = 23.32, $p < .001$, $R^2 = .26$), compared to those in the control condition (F (1, 2) = 9.01, $p < .001$, $R^2 = .12$).[4]

[4] There was a significant mean difference for low ratio bias task performing individuals (Control: $M = 2.31$, $SD = 1.59$; Risk Tutor: $M = 1.74$, $SD = 1.19$), via t-test, between conditions for ratio bias overprecision $t(114) = 2.18$, $p = .03$. Results indicated that there was significant reduction in overconfidence, but not necessarily improvement in performance, meaning that training risk literacy skills may improve self-evaluations. Further, there does not seem to be signs of an interaction with individuals with pre-test graph literacy, nor statistical numeracy. This indicates that, if results are representative and can generalize, all individuals who complete the tutor, no matter their prior skill, could improve in decision making and self-evaluation skills. For analyses looking at potential interactions see appendix D.

5.4 Two Hours of Graph Literacy Training Potentially can Improve Decision Making

To the extent that the results generalize, these findings appear to have many implications for diverse issues across social, moral, ethical, organizational, and educational contexts[5]. For example, research suggests that accurate self-evaluations can be associated with the quality of ethical decision making in several ways (Feltz & Cokely, 2018; Mumford et al., 2008). Moreover, the work of Koriat et al. (2020) suggests these confidence biases and judgments may be related to many evaluative social judgments about people's personalities, social beliefs, social attitudes, and other personal preferences.

Differences in overconfidence biases are also known to weigh on organizational psychology in many ways. For example, some research finds that self-evaluations may be crucial for many decisions within organizations. More specifically, Russo and Schoemaker (1992) have discussed how overconfidence, especially in managers and other leaders, has consequences that can be associated with financial loss or even catastrophic disasters (Ben-David et al., 2013; Grant, 2021; Moore, 2020; Tetlock & Gardner, 2016). For example, there is broad agreement that overconfidence in lead engineers' decision making likely played a major role in the oilwell blowout and subsequent explosion of the Deepwater Horizon rig (Labib & Read, 2010). These bias-related effects have also been observed in simulator-based research on offshore blowouts, revealing that higher confidence judgments do not necessarily reflect

[5] One example showing the importance of overconfidence can be found in the 2019 article by Peters et al. that measured objective and subjective numeracy in patients with systematic lupus erythematosus (SLE). Results indicated that SLE patients overconfident in their numeracy skills often reported higher disease activity. Extrapolating the results of the graph literacy tutor and the effect size in reducing overconfidence to the results of Peters et al. (2019), the completion of the two-hour risk literacy tutor could lead to decreased disease activity. The reported half a standard deviation improvements in overconfidence gained from the tutor could lead to the approximate reduction of disease activity of a low numerate individual to one more in line with someone double in objective numeracy, leading to potentially better health.

overconfident errors but can be robustly related to better blowout control, potentially preventing catastrophic damage (Raza et al., 2018).

Beyond these few training examples, perhaps the largest potential opportunity for developing more advanced training applications is found within educational (K-12) and post-secondary educational contexts. While these are far beyond this paper's current scope, it bears mentioning that the state-of-the-science of self-regulated learning often emphasizes roles involving appropriate monitoring and control strategy selections. However, research on self-regulated learning often generally neglects what may be very robust roles of individual differences in metacognitive skills related to risk literacy (Dunlosky & Hertzog, 1998; Zimmerman, 2000). Relatedly, the development of superior, verifiable expertise (i.e., expert performance) often depends significantly on accurate self-evaluations to inform the development of challenging and accurate goals as required to promote continuous skill attainment (Ericsson & Lehman, 1996). Self-evaluation skills may be especially pertinent with the continuing pivot to online education, particularly as some online tutors have a “hint” function that asks users to make metacognitive judgments on a question-to-question basis if they need help (Aleven & Koedinger, 2002; Aleven et al., 2003). However, it is noteworthy that there are currently no standards and minimal research on mining, designing, and implementing feedback in the form of “hints” in intelligent online tutors.

5.5 Conclusion

“It is wise to take admissions of uncertainty seriously, but declarations of high confidence mainly tell you that an individual has constructed a coherent story in his mind, not necessarily that the story is true.” -Daniel Kahneman *Thinking Fast and Slow* (2011).

The current state-of-the-science of overconfidence often rests on untested assumptions like those expressed in the quote above. Unfortunately, it is risky to assume that people are biased without objective evidence simply because they express high confidence in their abilities or relative competencies. To illustrate some relevant concerns, consider one typical example offered as clear evidence of ubiquitous overconfidence: most individuals believe themselves to be better than the average drivers, a statistical impossibility given the assumption of a normal distribution of drivers (Svenson, 1981). In this often-cited example of the "better-than-average effect," researchers explicitly assume that this effect must reflect biases, suggesting that, logically, most people cannot be above average. Ironically, however, this assumption is not false unless distributions are perfectly normal. After all, most people will be above average drivers if the distribution of the population of drivers is skewed (e.g., if a small number of drivers are responsible for most accidents, most people will have fewer accidents than average). When looking at individual differences in objective driving skills, there is often a robust positive relationship between confidence in one's driving skills and objectively measured performance (Groeger, 2001; Victoir et al., 2005). Furthermore, research shows that educational interventions focused on driver risk perceptions could reduce biases and overconfidence (Deery, 1999).

To further illustrate related methodological concerns, consider the fact that in the U.S., most people have below-average income levels, in that the average U.S. income is almost twice as high as the median U.S. income (e.g., because the top 1% makes almost as much money as the bottom 50% most people must have below-average incomes). Ironically, however, even in cases where the population is perfectly normally distributed, about 50% of any randomly sampled group of participants who say they are "better than average" would be correct. This finding would not imply that most people are biased and overconfident because, at maximum, only half of the people could be overconfident. There are many reasons to be concerned and skeptical about the value and meaning of “better-than-average” like effects. Relatedly, many investigations of overconfidence assume that differences in group means reflect underlying cognitive processes, which often is not the case when individual data responses have been averaged or aggregated (e.g., if the average parent has 2.4 children, this does not imply that every parent has two whole children and one partial child). These and other kinds of untested or unjustifiable assumptions, often found in overconfidence research, can lead to many kinds of misinterpretations and should not be neglected in future research.

Beyond helping to clarify some limitations of methods and theoretical interpretations of previous overconfidence research, the current findings may have implications for designing effective skill training programs, including implications for developing adaptive training technologies. For example, consider opportunities for integrating information about subjective confidence judgments as part of *hint mining* procedures (See appendix D, part 4). Because intelligent tutors often adapt to individuals' needs based on the factors that have benefited other similar people, differences in (valid and biased) subjective self-evaluation judgments may inform the evaluation and selection of various training hints. To further illustrate, after completing a problem in an online tutor, a student might make a confidence judgment about that specific problem and then generate a hint to help other individuals and provide a confidence judgment about the value of that hint. The generated hints and subjective confidence judgments could then be reverse engineered to assess the extent, why, when, and how these judgments provide an essential index of the expected learning value of various hints for diverse learners. Of course, using subjective self-evaluation judgments to support hint mining operations is just one general example of the kinds of tools and insights that could follow from a more accurate assessment and theoretical understanding of the nature of individual differences in subjective self-evaluation skills and overconfidence biases.

Taken altogether, in contrast to the increasingly influential, high-profile claims suggesting that regardless of one's skills, "everyone is overconfident," the current results suggest that skills may generally protect people from many kinds of overconfidence biases across many kinds of cognitive tasks (Cokely et al., 2018; Feltz & Cokely, 2018; Ghazal et al., 2014; Okan et al., 2015; Peters, 2020; Petrova, Van de Pligt, Garcia-Retamero, 2014; Reyna et al., 2009; 2018; 2020; Szaszi et al., 2017). Consistent with Skilled Decision Theory, it appears that general risk literacy skills often causally give rise to general, trainable decision making skills and general, trainable metacognitive and self-evaluation skills. Notably, the discovery of the direct causal link between risk literacy skills and metacognitive skills appears to be the first such finding reported in the scientific literature. While more research is needed, results suggest we should be relatively confident that people can improve their general self-evaluation skills by specifically improving their risk literacy skills.

References

Adams, J. K., & Adams, P. A. (1961). Realism of confidence judgments. *Psychological review, 68*(1), 33.

Aleven, V. A., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. *Cognitive science, 26*(2), 147-179.

Aleven, V., & Koedinger, K. R. (2000, June). Limitations of student control: Do students know when they need help?. In *International conference on intelligent tutoring systems* (pp. 292-303). Springer, Berlin, Heidelberg.

Aleven, V., McLaren, B. M., Sewall, J., & Koedinger, K. R. (2006, June). The cognitive tutor authoring tools (CTAT): Preliminary evaluation of efficiency gains. In *International Conference on Intelligent Tutoring Systems* (pp. 61-70). Springer, Berlin, Heidelberg.

Aleven, V., McLaren, B., Roll, I., & Koedinger, K. (2006). Toward meta-cognitive tutoring: A model of help seeking with a Cognitive Tutor. *International Journal of Artificial Intelligence in Education*, *16*(2), 101-128.

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.

Anderson, C., Brion, S., Moore, D. A., & Kennedy, J. A. (2012). A status-enhancement account of overconfidence. *Journal of personality and social psychology*, *103*(4), 718.

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological review*, *89*(4), 369.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The journal of the learning sciences*, *4*(2), 167-207.

Anzai, Y., & Simon, H. A. (1979). The theory of learning by doing. *Psychological review*, *86*(2), 124.

Arkes, H. R., Christensen, C., Lai, C., & Blumer, C. (1987). Two methods of reducing overconfidence. *Organizational behavior and human decision processes*, *39*(1), 133-144.

Ashraf Labib and Martin Read, 'Not Just Rearranging the Deckchairs on the Titanic: Learning from Failures through Risk and Reliability Analysis', *Safety Science*, 51.1 (2013), 397–413; Don A Moore and Samuel A Swift, 'The Three Faces of Overconfidence in Organizations', in *Social Psychology of Organizations*, ed. by Rolf Van Dick and J Keith Murnighan (Oxford: Taylor & Francis, 2010), pp. 147–84.

Bandura, A. (1986). The explanatory and predictive scope of self-efficacy theory. *Journal of social and clinical psychology*, *4*(3), 359-373.

Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception & psychophysics*, *55*(4), 412-428.

Benoît, J. P., & Dubra, J. (2011). Apparent overconfidence. *Econometrica*, *79*(5), 1591-1625.

Bessarabova, E., Piercy, C. W., King, S., Vincent, C., Dunbar, N. E., Burgoon, J. K., ... & Lee, Y. H. (2016). Mitigating bias blind spot via a serious video game. *Computers in Human Behavior*, *62*, 452-466.

Biswas, G., Jeong, H., Kinnebrew, J. S., Sulcer, B., & Roscoe, R. O. D. (2010). Measuring self-regulated learning skills through social interactions in a teachable agent environment. *Research and Practice in Technology Enhanced Learning*, *5*(02), 123-152.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual review of psychology*, *64*, 417-444.

Björkman, M., Juslin, P., & Winman, A. (1993). Realism of confidence in sensory discrimination: The underconfidence phenomenon. *Perception & psychophysics*, *54*(1), 75-81.

Boekaerts, M., Pintrich, P. R., & Zeidner, M. (2000). Self-regulation: An introductory overview. *Handbook of self-regulation*, 1-9.

Bonnette, A. L., & Bruell, C. J. (2014). *Memorabilia*. Cornell University Press.

Brachtendof, J. (2017). Augustine on Self-Knowledge and Human Subjectivity. In Renz, U. (Eds.), *Self-knowledge: A history*. Oxford University Press.

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.

Buratti, S., Allwood, C. M., & Kleitman, S. (2013). First-and second-order metacognitive judgments of semantic memory reports: The influence of personality traits and cognitive styles. *Metacognition and learning*, *8*(1), 79-102.

Butler, A. C., Karpicke, J. D., & Roediger III, H. L. (2008). Correcting a metacognitive error: feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(4), 918.

Cokely, E. T., Allan, J., Ghazal, S., Feltz, A., & Garcia-Retamero, R. (Forthcoming). General decision making skill and numeracy: Component abilities.

Cokely, E. T., Feltz, A., Ghazal, S., Allan, J. N., Petrova, D., & Garcia-Retamero, R. (2018). Decision making skill: 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.

Cokely, E. T., Kelley, C. M., & Gilchrist, A. L. (2006). Sources of individual differences in working memory: Contributions of strategy to capacity. Psychonomic Bulletin & Review, 13(6), 991-997.

Cory, T. S. (2014). *Aquinas on human self-knowledge*. Cambridge University Press.

Crawford, J. D., & Stankov, L. (1996). Age differences in the realism of confidence judgements: A calibration study using tests of fluid and crystallized intelligence. *Learning and Individual Differences*, *8*(2), 83-103.

d Baker, R. S., Corbett, A. T., Koedinger, K. R., Evenson, S., Roll, I., Wagner, A. Z., ... & Beck, J. E. (2006, June). Adapting to when students game an intelligent tutoring system. In *International conference on intelligent tutoring systems* (pp. 392-401). Springer, Berlin, Heidelberg.

Daniel Kahneman, *Thinking Fast and Slow* (New York: Farrar, Straus and Giroux, 2011).

Davies, C. M. (1965). Development of the probability concept in children. *Child Development*, 779-788.

Deery, H. A. (1999). Hazard and risk perception among young novice drivers. *Journal of safety research*, *30*(4), 225-236.

Dunbar, N. E., Miller, C. H., Adame, B. J., Elizondo, J., Wilson, S. N., Schartel, S. G., ... & Morrison, D. (2013). Mitigation of cognitive bias through the use of a serious game. In *Proceedings of the games learning society annual conference*.

Dunlosky, J., & Hertzog, C. (1998). Training programs to improve learning in later adulthood: Helping older adults educate themselves. *Metacognition in educational theory and practice*, *249*, 276.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4-58.

Dunning D, Heath C, Suls JM. Flawed Self-Assessment: Implications for Health, Education, and the Workplace. *Psychological Science in the Public Interest*. 2004;5(3):69-106. doi:10.1111/j.1529-1006.2004.00018.x

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. Current directions in psychological science, 12(3), 83-87.

E. Stanovich, K., & West, R. F. (1998). Individual differences in framing and conjunction effects. *Thinking & Reasoning*, *4*(4), 289-317.

Ehrlinger, J., Gilovich, T., & Ross, L. (2005). Peering into the bias blind spot: People's assessments of bias in themselves and others. *Personality and Social Psychology Bulletin*, *31*(5), 680-692.

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational behavior and human decision processes*, *105*(1), 98-121.

Einhorn, H. J., & Hogarth, R. M. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological review*, *85*(5), 395.

Emundts, D. (2017). Kant's Ideal of Self-Knowledge. In Renz, U. (Eds.), *Self-knowledge: A history*. Oxford University Press.

Epley, N., & Dunning, D. (2006). The mixed blessings of self-knowledge in behavioral prediction: Enhanced discrimination but exacerbated bias. *Personality and Social Psychology Bulletin*, *32*(5), 641-655.

Ericsson, K. A. (2006). Protocol analysis and expert thought: Concurrent verbalizations of thinking during experts' performance on representative tasks. *The Cambridge handbook of expertise and expert performance*, 223-241.

Ericsson, K. A., Hoffman, R. R., Kozbelt, A., & Williams, A. M. (Eds.). (2018). The Cambridge handbook of expertise and expert performance. Cambridge University Press.

Ericsson, K. A., & Moxley, J. H. (2019). Thinking aloud during superior performance on tasks involving decision making. *A Handbook of Process Tracing Methods*, 286.

Ericsson, K. A., Prietula, M. J., & Cokely, E. T. (2007). The making of an expert. Harvard business review, 85(7/8), 114.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. the MIT Press.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological review*, *100*(3), 363.

Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring numeracy without a math test: development of the Subjective Numeracy Scale. *Medical Decision Making*, *27*(5), 672-680.

Falchikov, N., & Boud, D. (1989). Student Self-Assessment in Higher Education: A Meta-Analysis. *Review of Educational Research*, *59*(4), 395–430.Provided everyone agrees on what how to assess driving; Eric van den Steen, 'Rational Overoptimism (and Other Biases)', *American Economic Review*, 94.4 (2004), 1141–51.

Feltz, A., & Cokely, E. T. (2018). Informing ethical decision making. *The Routledge Handbook of Neuroethics*, 304.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human perception and performance*, *3*(4), 552.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, *34*(10), 906.

Fullerton, G. S., & Cattell, J. M. (1892). On the perception of small differences. Publications of the University of Pennsylvania (No.2).

Garcia-Retamero, R., & Cokely, E. T. (2013). Communicating health risks with visual aids. *Current Directions in Psychological Science*, *22*(5), 392-399.

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.

Garcia-Retamero, R., Cokely, E. T., & Hoffrage, U. (2015). Visual aids improve diagnostic inferences and metacognitive judgment calibration. *Frontiers in Psychology*, *6*, 932.

Garcia-Retamero, R., Cokely, E. T., Ghazal, S., & Joeris, A. (2016). Measuring graph literacy without a test: A brief subjective assessment. *Medical Decision Making*, *36*(7), 854-867.

Garrett, A. (2017). Self-Knowledge and Self-Deception in Modern Moral Philosophy. In Renz, U. (Eds.), *Self-knowledge: A history*. Oxford University Press.

Ghazal, S. (2014). Component numeracy skills and decision making.

Ghazal, S., Cokely, E. T., & Garcia-Retamero, R. (2014). Predicting biases in very highly educated samples: Numeracy and metacognition. *Judgment and decision making*.

Gigerenzer, G. (2015). Risk savvy: How to make good decisions. Penguin.

Gigerenzer, G. (1993). The bounded rationality of probabilistic mental models. In *Rationality: Psychological and philosophical perspectives* (pp. 284-313). Routledge.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological review*, *103*(4), 650.

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. Oxford University Press, USA.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: a Brunswikian theory of confidence. *Psychological review*, *98*(4), 506.

Grant, A. (2021). *Think Again: The Power of Knowing What You Don’t Know*. Penguin Publishing Group

Groeger, J. A. (2001). Understanding Driving. Applying cognitive psychology to a complex everyday task. Philadelphia: Taylor & Francis Inc.

Griffin, D., & Brenner, L. (2004). Perspectives on probability judgment calibration. *Blackwell handbook of judgment and decision making*, 177-199.

Grimes, P. W. (2002). The overconfident principles of economics student: An examination of a metacognitive skill. *The Journal of Economic Education*, *33*(1), 15-30.

Hammond, K. R., & Stewart, T. R. (Eds.). (2001). *The essential Brunswik: Beginnings, explications, applications*. Oxford University Press.

Hartwig, M. K., & Dunlosky, J. (2014). The contribution of judgment scale to the unskilled-and-unaware phenomenon: How evaluating others can exaggerate over-(and under-) confidence. Memory & cognition, 42(1), 164-173.

Hermes, H., Huschens, M., Rothlauf, F., & Schunk, D. (2019). Motivating Low-Achievers—Relative Performance Feedback in Primary Schools. *NHH Dept. of Economics Discussion Paper*, (14).

Hoffman, R. R., Ward, P., Feltovich, P. J., DiBello, L., Fiore, S. M., & Andrews, D. H. (2013). *Accelerated expertise: Training for high proficiency in a complex world*. Psychology Press.

Hoffrage, U. (2004). 13 Overconfidence. *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*, 235.

Hogarth, R. M., & Karelaia, N. (2012). Entrepreneurial success and failure: Confidence and fallible judgment. *Organization Science*, *23*(6), 1733-1747.

Hogarth, R. M., Mukherjee, K., & Soyer, E. (2013). Assessing the chances of success: Naïve statistics versus kind experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(1), 14.

Huang, Yun & Aleven, Vincent & McLaughlin, Elizabeth & Koedinger, Kenneth. (2020). A General Multi-method Approach to Design-Loop Adaptivity in Intelligent Tutoring Systems. 10.1007/978-3-030-52240-7_23.

Hultsch, D. F., Hertzog, C., Small, B. J., & Dixon, R. A. (1999). Use it or lose it: engaged lifestyle as a buffer of cognitive decline in aging?. *Psychology and aging*, *14*(2), 245.

Itzhak Ben-David, John R. Graham, Campbell R. Harvey, Managerial Miscalibration*, *The Quarterly Journal of Economics*, Volume 128, Issue 4, November 2013, Pages 1547–1584, https://doi.org/10.1093/qje/qjt023

Jackson, S. A., & Kleitman, S. (2014). Individual differences in decision-making and confidence: capturing decision tendencies in a fictitious medical test. *Metacognition and learning*, *9*(1), 25-49.

Jackson, S. A., Kleitman, S., Howie, P., & Stankov, L. (2016). Cognitive abilities, monitoring confidence, and control thresholds explain individual differences in heuristics and biases. *Frontiers in Psychology*, *7*, 1559.

Juslin, P. (1993). An explanation of the hard-easy effect in studies of realism of confidence in one's general knowledge. *European Journal of Cognitive Psychology*, *5*(1), 55-71.

Juslin, P., & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological review*, *104*(2), 344.

Juslin, P., Olsson, H., & Björkman, M. (1997). Brunswikian and Thurstonian origins of bias in probability assessment: On the interpretation of stochastic components of judgment. *Journal of Behavioral Decision Making*, *10*(3), 189-209.

Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence–accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(5), 1304.

Juslin, P., Wennerholm, P., & Olsson, H. (1999). Format dependence in subjective probability calibration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(4), 1038.

Juslin, P., Winman, A., & Hansson, P. (2007). The naïve intuitive statistician: A naïve sampling model of intuitive confidence intervals. *Psychological review*, *114*(3), 678.

Juslin, P., Winman, A., & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hard–easy effect. *Psychological review*, *107*(2), 384.

Juslin, P., Winman, A., & Olsson, H. (2003). Calibration, additivity, and source independence of probability judgments in general knowledge and sensory discrimination tasks. *Organizational Behavior and Human Decision Processes*, *92*(1-2), 34-51.

Kamtekar, R. (2016). Becoming Good. *Developing the Virtues: Integrating Perspectives*, 184.

Kamtekar, R. (2017). Self-Knowledge in Plato. In Renz, U. (Eds.), *Self-knowledge: A history*. Oxford University Press.

Karelaia, N., & Hogarth, R. M. (2010). The attraction of uncertainty: Interactions between skill and levels of uncertainty in market-entry games. *Journal of Risk and Uncertainty*, *41*(2), 141-166.

Kitsantas, A., Zimmerman, B.J. Enhancing self-regulation of practice: the influence of graphing and self-evaluative standards. *Metacognition Learning* **1,** 201–212 (2006). https://doi.org/10.1007/s11409-006-9000-7

Kleitman, S., & Costa, D. S. (2014). The role of a novel formative assessment tool (Stats-mIQ) and individual differences in real-life academic performance. *Learning and Individual Differences*, *29*, 150-161.

Kleitman, S., & Stankov, L. (2007). Self-confidence and metacognitive processes. *Learning and individual differences*, *17*(2), 161-173.

Koedinger, K. R., D'Mello, S., McLaughlin, E. A., Pardos, Z. A., & Rose, C. P. (2015). Data mining and education. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(4), 333-353.

Koriat, A. (2000). The feeling of knowing: Some metatheoretical implications for consciousness and control. *Consciousness and cognition*, *9*(2), 149-171.

Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological review*, *119*(1), 80.

Koriat, A. (2017). Can people identify “deceptive” or “misleading” items that tend to produce mostly wrong answers?. *Journal of Behavioral Decision Making*, *30*(5), 1066-1077.

Koriat, A. (2018). When reality is out of focus: Can people tell whether their beliefs and judgments are correct or wrong?. *Journal of Experimental Psychology: General*, *147*(5), 613.

Koriat, A. (2019). Confidence judgments: The monitoring of object-level and same-level performance. *Metacognition and Learning*, *14*(3), 463-478.

Koriat, A., Adiv, S., & Schwarz, N. (2016). Views that are shared with others are expressed with greater confidence and greater fluency independent of any social influence. *Personality and Social Psychology Review*, *20*(2), 176-193.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human learning and memory*, *6*(2), 107.

Koriat, A., Undorf, M., Newman, E., & Schwarz, N. (2020). Subjective confidence in the response to personality questions: Some insight into the construction of people’s responses to test items. *Frontiers in Psychology*, *11*.

Kramarski, B., & Mevarech, Z. R. (2003). Enhancing Mathematical Reasoning in the Classroom: The Effects of Cooperative Learning and Metacognitive Training. *American Educational Research Journal*, *40*(1), 281–310.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, *77*(6), 1121.

Lang, C., Heffernan, N., Ostrow, K., & Wang, Y. (2015). The Impact of Incorporating Student Confidence Items into an Intelligent Tutor: A Randomized Controlled Trial. *International Educational Data Mining Society*.

Larrick, R. P., Burson, K. A., & Soll, J. B. (2007). Social comparison and confidence: When thinking you’re better than average predicts overconfidence (and when it does not). *Organizational Behavior and Human Decision Processes*, *102*(1), 76-94.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know?. *Organizational behavior and human performance, 20*(2), 159-183.

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational behavior and human performance*, *26*(2), 149-171.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1977). Calibration of probabilities: The state of the art. *Decision making and change in human affairs*, 275-324.
Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1977). Calibration of probabilities: The state of the art. *Decision making and change in human affairs*, 275-324.
Lipkus, I. M., & Peters, E. (2009). Understanding the role of numeracy in health: proposed theoretical framework and practical insights. *Health Education & Behavior*, *36*(6), 1065-1081.

Markopoulos, E., & Vanharanta, H. (2017). Delphic Maxims Based Applied Philosophy for Business and Governance Management. In *Advances in Human Factors, Business Management, Training and Education* (pp. 33-45). Springer, Cham.

Mata, R., Schooler, L. J., & Rieskamp, J. (2007). The aging decision maker: cognitive aging and the adaptive selection of decision strategies. *Psychology and aging*, *22*(4), 796.

Miller, T. M., & Geraci, L. (2011). Unskilled but aware: Reinterpreting overconfidence in low-performing students. Journal of experimental psychology: learning, memory, and cognition, 37(2), 502.

Mitchum, A. L., & Kelley, C. M. (2010). Solve the problem first: Constructive solution strategies can influence the accuracy of retrospective confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(3), 699.

Moore, C. (2015). *Socrates and self-knowledge*. Cambridge University Press.

Moore, D. A. (2020). *Perfectly Confident: How to Calibrate Your Decisions Wisely*. HarperCollins.

Moore, D. A., & Cain, D. M. (2007). Overconfidence and underconfidence: When and why people underestimate (and overestimate) the competition. *Organizational Behavior and Human Decision Processes*, *103*(2), 197-213.

Moore, D. A., & Dev, A. S. (2017). Individual differences in overconfidence. *Encyclopedia of Personality and Individual Differences. Springer. Retrieved from http://osf. io/hzk6q*.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological review*, *115*(2), 502.

Moore, D. A., & Schatz, D. (2017). The three faces of overconfidence. *Social and Personality Psychology Compass*, *11*(8), e12331.

Moore, D. A., & Small, D. A. (2008). When it is rational for the majority to believe that they are better than average. *Rationality and social responsibility: Essays in honor of Robyn Mason Dawes*, 141-174.

Moore, D. A., Swift, S. A., Minster, A., Mellers, B., Ungar, L., Tetlock, P., ... & Tenney, E. R. (2017). Confidence calibration in a multiyear geopolitical forecasting competition. *Management Science*, *63*(11), 3552-3565.

Morewedge, C. K., Yoon, H., Scopelliti, I., Symborski, C. W., Korris, J. H., & Kassam, K. S. (2015). Debiasing decisions: Improved decision making with a single training intervention. Policy Insights from the Behavioral and Brain Sciences, 2(1), 129-140.

Mousavi, S., & Gigerenzer, G. (2011). Revisiting the "error" in studies of cognitive errors.

Mumford MD, Connelly S, Brown RP, et al. A Sensemaking Approach to Ethics Training for Scientists: Preliminary Evidence of Training Effectiveness. *Ethics Behav*. 2008;18(4):315-339. doi:10.1080/10508420802487815

Nelson, T. O. (1990). Metamemory: A theoretical framework and new findings. In *Psychology of learning and motivation* (Vol. 26, pp. 125-173). Academic Press.

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.

Okan, Y., Garcia-Retamero, R., Cokely, E. T., & Maldonado, A. (2015). Improving risk understanding across ability levels: Encouraging active processing with dynamic icon arrays. Journal of Experimental Psychology: Applied, 21(2), 178.

Okan, Y., Stone, E. R., & Bruine de Bruin, W. (2018). Designing graphs that promote both risk understanding and behavior change. *Risk Analysis*, *38*(5), 929-946.

Olsson, H. (2014). Measuring overconfidence: Methodological problems and statistical artifacts. *Journal of Business Research*, *67*(8), 1766-1770.

Olsson, H., & Winman, A. (1996). Underconfidence in sensory discrimination: The interaction between experimental setting and response strategies. *Perception & Psychophysics*, *58*(3), 374-382.

Pallier, G., Wilkinson, R., Danthiir, V., Kleitman, S., Knezevic, G., Stankov, L., & Roberts, R. D. (2002). The role of individual differences in the accuracy of confidence judgments. *The Journal of general psychology*, *129*(3), 257-299.

Peirce, C. S., & Jastrow, J. (1884). On small differences of sensation. Memoirs National Academy of Sciences, 3, 73-83.

Perler, D. (2017). Self-Knowledge in Scholasticism. In Renz, U. (Eds.), *Self-knowledge: A history*. Oxford University Press.

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

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

Peters, E., Tompkins, M. K., Knoll, M. A., Ardoin, S. P., Shoots-Reinhard, B., & Meara, A. S. (2019). Despite high objective numeracy, lower numeric confidence relates to worse financial and medical outcomes. Proceedings of the National Academy of Sciences, 116(39), 19386-19391.

Petrova, D. G., Van der Pligt, J., & Garcia-Retamero, R. (2014). Feeling the numbers: On the interplay between risk, affect, and numeracy. *Journal of Behavioral Decision Making*, *27*(3), 191-199.

Petrova, D., Garcia-Retamero, R., & Cokely, E. T. (2015). Understanding the harms and benefits of cancer screening: A model of factors that shape informed decision making. Medical Decision Making, 35, 847 858.

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.

Philip E Tetlock and Dan Gardner, *Superforecasting: The Art and Science of Prediction* (New York: Signal, 2015).

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychological review*, *117*(3), 864.

Prims, J. P., & Moore, D. A. (2017). Overconfidence over the lifespan. *Judgment and decision making*, *12*(1), 29.

Pronin, E., Daniel Y Lin, and Lee Ross, (2002). The Bias Blind Spot: Perceptions of Bias in Self versus Others, *Personality and Social Psychology Bulletin*, 28 (3). 369–81.

Pronin, E., & Kugler, M. B. (2007). Valuing thoughts, ignoring behavior: The introspection illusion as a source of the bias blind spot. *Journal of Experimental Social Psychology*, *43*(4), 565-578.

Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin*, *28*(3), 369-381.

Raza, M. A., Salehi, S., Ghazal, S., Ybarra, V. T., Naqvi, S. A. M., Cokely, E. T., & Teodoriu, C. (2019). Situational awareness measurement in a simulation-based training framework for offshore well control operations. *Journal of Loss Prevention in the Process Industries*, *62*, 103921.

Reginster, B. (2017). Self-Knowledge as Freedom in Schopenhauer and Freud. In Renz, U. (Eds.), *Self-knowledge: A history*. Oxford University Press.

Renz, U. (2018). Self-Knowledge and Knowledge of Mankind in Hobbes' Leviathan. *European Journal of Philosophy*, *26*(1), 4-29.

Renz, U. (Ed.). (2017). *Self-knowledge: A history*. Oxford University Press.

Renz, U., Moyal-Sharrock, D., Munz, V., & Coliva, A. (2015). Becoming aware of one's thoughts: Kant on self-knowledge and reflective experience.

Reyna, V. (2018). When irrational biases are smart: A fuzzy-trace theory of complex decision making. Journal of Intelligence, 6(2), 29.

Reyna, V. F., & Brust-Renck, P. G. (2020). How representations of number and numeracy predict decision paradoxes: A fuzzy-trace theory approach. Journal of Behavioral Decision Making, 22(6), 606-628. https://doi.org/10.1002/bdm.2179

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.

Rivers, K., & Koedinger, K. R. (2017). Data-driven hint generation in vast solution spaces: a self-improving python programming tutor. *International Journal of Artificial Intelligence in Education*, *27*(1), 37-64.

Roll, I., Aleven, V., McLaren, B. M., & Koedinger, K. R. (2007, June). Can Help Seeking Be Tutored? Searching for the Secret Sauce of Metacognitive Tutoring. In *AIED* (Vol. 2007, pp. 203-10).

Russo, J. E., & Schoemaker, P. J. (1992). Managing overconfidence. *Sloan management review*, *33*(2), 7-17.

Salehi, S., Kiran, R., Jeon, J., Kang, Z., Cokely, E. T., & Ybarra, V. (2018). Developing a cross-disciplinary, scenario-based training approach integrated with eye tracking data collection to enhance situational awareness in offshore oil and gas operations. *Journal of Loss Prevention in the Process Industries, 56,* 78-94.

Sanchez, C., & Dunning, D. (2018). Overconfidence among beginners: Is a little learning a dangerous thing?. Journal of Personality and Social Psychology, 114(1), 10.

Szaszi, B., Szollosi, A., Palfi, B., & Aczel, B. (2017). The cognitive reflection test revisited: exploring the ways individuals solve the test. Thinking & Reasoning, 23(3), 207-234.

Scopelliti, I., Morewedge, C. K., McCormick, E., Min, H. L., Lebrecht, S., & Kassam, K. S. (2015). Bias blind spot: Structure, measurement, and consequences. *Management Science*, *61*(10), 2468-2486.

Sitzman, D. M., Rhodes, M. G., & Tauber, S. K. (2014). Prior knowledge is more predictive of error correction than subjective confidence. Memory & Cognition, 42(1), 84-96.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 204.

Sperling, R. A., Howard, B. C., Staley, R., & DuBois, N. (2004). Metacognition and self-regulated learning constructs. *Educational research and evaluation*, *10*(2), 117-139.

Stankov, L. (2013). Noncognitive predictors of intelligence and academic achievement: An important role of confidence. *Personality and Individual Differences*, *55*(7), 727-732.

Stankov, L., & Kleitman, S. (2008). Processes on the borderline between cognitive abilities and personality: Confidence and its realism. *The handbook of personality theory and testing*, 541-555.

Stankov, L., & Kleitman, S. (2014). Whither metacognition. *Learning and individual differences*, *29*, 120-122.

Stankov, L., & Lee, J. (2008). Confidence and cognitive test performance. *Journal of educational psychology*, *100*(4), 961.

Stankov, L., & Lee, J. (2014). Overconfidence across world regions. *Journal of Cross-Cultural Psychology*, *45*(5), 821-837.

Stankov, L., Kleitman, S., & Jackson, S. A. (2015). Measures of the trait of confidence. *Measures of personality and social psychological constructs*, 158-189.

Stankov, L., Lee, J., & Paek, I. (2009). Realism of confidence judgments. *European Journal of Psychological Assessment*, *25*(2), 123.

Stankov, L., Lee, J., Luo, W., & Hogan, D. J. (2012). Confidence: A better predictor of academic achievement than self-efficacy, self-concept and anxiety?. *Learning and Individual Differences*, *22*(6), 747-758.

Stankov, L., Morony, S., & Lee, Y. P. (2014). Confidence: the best non-cognitive predictor of academic achievement?. *Educational psychology*, *34*(1), 9-28.

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of experimental psychology: general*, *127*(2), 161.

Symborski, C., Barton, M., Quinn, M., Morewedge, C., Kassam, K., Korris, J. H., & Hollywood, C. A. (2014). Missing: A serious game for the mitigation of cognitive biases. In Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) (No. 14295, pp. 1-13).

Symborski, C., Barton, M., Quinn, M. M., Korris, J. H., Kassam, K. S., & Morewedge, C. K. (2016). The Design and Development of Serious Games Using Iterative Evaluation. Games and Culture, 1, 17.

Szaszi, B., Szollosi, A., Palfi, B., & Aczel, B. (2017). The cognitive reflection test revisited: exploring the ways individuals solve the test. *Thinking & Reasoning*, *23*(3), 207-234.

Terry R. & Ybarra, V.T. (Ed.). (2020). The Eyes have it: Emphasizing Data Visualization When Teaching Students Meeting a Quantitative Literacy Requirement. *Teaching Statistics and Quantitative Methods in the 21st Century*. Chapter 10. Routledge.

Tetlock, P. E. (2000). Cognitive biases and organizational correctives: Do both disease and cure depend on the politics of the beholder?. *Administrative Science Quarterly*, *45*(2), 293-326.

Tetlock, P. E., & Levi, A. (1982). Attribution bias: On the inconclusiveness of the cognition-motivation debate. *Journal of Experimental Social Psychology*, *18*(1), 68-88.

Tetlock, P. E., Mitchell, G., & Murray, T. L. (2008). The challenge of debiasing personnel decisions: Avoiding both under-and overcorrection. *Industrial and Organizational Psychology*, *1*(4), 439.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & cognition*, *39*(7), 1275-1289.

Van Zant AB, Moore DA. Avoiding the Pitfalls of Overconfidence While Benefiting from the Advantages of Confidence. *California Management Review*. 2013;55(2):5-23.

Victoir, A., Eertmans, A., van den Bergh, O., & van den Broucke, S. (2005). Learning to drive safely: Social-cognitive responses are predictive of performance rated by novice drivers and their instructors. Transportation Research F: Traffic psychology and behaviour, 8(1), 59–74.

Watts, L. L., Mulhearn, T. J., Medeiros, K. E., Steele, L. M., Connelly, S., & Mumford, M. D. (2017). Modeling the instructional effectiveness of responsible conduct of research education: A meta-analytic path-analysis. *Ethics & Behavior*, *27*(8), 632-650.

West, R. F., Meserve, R. J., & Stanovich, K. E. (2012). Cognitive sophistication does not attenuate the bias blind spot. *Journal of personality and social psychology*, *103*(3), 506.

William, J. (1950). *Principles of Psychology Vol. I and II Complete*. Dover Publications.

Wilson, T. D. (2009). Know thyself. *Perspectives on Psychological Science*, *4*(4), 384-389.

Winman, A., Hansson, P., & Juslin, P. (2004). Subjective probability intervals: how to reduce overconfidence by interval evaluation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(6), 1167.

Woller-Carter, M. (2015). Development of the intelligent graphs for everyday risky decisions tutor.

Woller-Carter, M. M., Okan, Y., Cokely, E. T., & Garcia-Retamero, R. (2012, September). Communicating and distorting risks with graphs: An eye-tracking study. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 56, No. 1, pp. 1723-1727). Sage CA: Los Angeles, CA: SAGE Publications.

Ybarra, V. T., Allan, J. N., Cokely, E. T., Feltz, A, & Garcia-Retamero, R. (in-prep) Numerate Individuals are Less Biased and Know It

Ybarra, V. (2018). Self-Evaluation of Skills and Overconfidence Vulnerability: Are Most People Blind to their Own Decision Making Biases?.

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*.

Yudelson, M. V., Koedinger, K. R., & Gordon, G. J. (2013, July). Individualized bayesian knowledge tracing models. In *International conference on artificial intelligence in education* (pp. 171-180). Springer, Berlin, Heidelberg.

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In *Handbook of self-regulation* (pp. 13-39). Academic Press.

Zimmerman, B. J., & Schunk, D. H. (Eds.). (2011). *Educational psychology handbook series.Handbook of self-regulation of learning and performance.* Routledge/Taylor & Francis Group.

Appendix A

Pre-Test Measures and Tasks

Subjective Numeracy Scale (Fagerlin et al., 2007).

Item 7 is reverse scored, and the average of all items is the overall subjective numeracy. The average of items 1-4 comprises the abilities subscale. The average of items 5-8 comprises the preferences subscale. Higher scores mean higher skill levels.

For each of the following questions, please check the box that best reflects how good you are at doing the following things:

	1	2	3	4	5
How good are you at working with fractions? (1)	○	○	○	○	○
How good are you at working with percentages? (2)	○	○	○	○	○
How good are you at calculating a 15% tip? (3)	○	○	○	○	○
How good are you at figuring out how much a shirt will cost if it is 25% off? (4)	○	○	○	○	○

For each of the following questions, please check the box that best reflects your answer:

When reading the newspaper, how helpful do you find tables and graphs that are parts of a story?

	1	2	3	4	5
(1)	○	○	○	○	○

When people tell you the chance of something happening, do you prefer that they use words ("it rarely happens") or numbers ("there's a 1% chance")?

	1	2	3	4	5
(1)	○	○	○	○	○

When you hear a weather forecast, do you prefer predictions using percentages (e.g., "there will be a 20% chance of rain today") or predictions using only words (e.g., "there is a small chance of rain today")?

	1	2	3	4	5
(1)	○	○	○	○	○

How often do you find numerical information to be useful?

	1	2	3	4	5
(1)	○	○	○	○	○

Subjective Graph Literacy

Below is the used 5-item subjective graph literacy scale. There is a full 10 item subjective graph literacy scale as detailed in Garcia-Retamero et al. (2016). This scale was also asked again in the post test to see if subjective graph proficiency changed after the intervention.

Please respond to the following questions.

	1 - not at all good	2	3	4	5	6 - extremely good
How good are you at working with bar charts? (1)	○	○	○	○	○	○
How good are you at working with line plots? (2)	○	○	○	○	○	○
How good are you at working with pie charts? (3)	○	○	○	○	○	○
How good are you at inferring the size of a bar in a bar chart? (4)	○	○	○	○	○	○
How good are you at determining the difference between 2 bars in a bar chart? (5)	○	○	○	○	○	○

The additional 5 Subjective Graph Literacy Questions are as follows:

6. How good are you at projecting a future trend from a line chart? (1 = not at all good; 6 = extremely good)

7. Are graphs easier to understand than numbers? (1 = not at all; 6 = much easier)

8. How often do you find graphical information to be useful? (1 = never; 6 = very often)

9. To what extent do you believe in the saying ‘‘a picture is worth one thousand words’’? (1 = not at all; 6 = extremely

10. When reading books or newspapers, how helpful do you find graphs that are part of a story? (1 = not at all; 6 = extremely)

Graph Construction Knowledge.

The created scale was used as a pseudo pre-test on graphical information.

When using a bar graph which situation/goal is most appropriate for its use?

○ a. Communicate Information about Proportions

○ b. Compare Several Data Points

○ c. Depiction of Large Very Numbers

○ d. Communicate Treatment Risk Reduction or Risk of Side Effects

○ e. Trends Over Time

When using a line graph which situation/goal is most appropriate for its use?

○ a. Communicate Information about Proportions

○ b. Compare Several Data Points

○ c. Depiction of Large Very Numbers

○ d. Communicate Treatment Risk Reduction or Risk of Side Effects

○ e. Trends Over Time

When using a pie graph which situation/goal is most appropriate for its use?

○ a. Communicate Information about Proportions

○ b. Compare Several Data Points

○ c. Depiction of Large Very Numbers

○ d. Communicate Treatment Risk Reduction or Risk of Side Effects

○ e. Trends Over Time

When using a grid which situation/goal is most appropriate for its use?

- ◯ a. Communicate Information about Proportions
- ◯ b. Compare Several Data Points
- ◯ c. Depiction of Large Very Numbers
- ◯ d. Communicate Treatment Risk Reduction or Risk of Side Effects
- ◯ e. Trends Over Time

When using an icon array which situation/goal is most appropriate for its use?

- ◯ a. Communicate Information about Proportions
- ◯ b. Compare Several Data Points
- ◯ c. Depiction of Large Very Numbers
- ◯ d. Communicate Treatment Risk Reduction or Risk of Side Effects
- ◯ e. Trends Over Time

When should shadows and truncated scales be used?

- ◯ a. When the presented information is highly complex
- ◯ b. When the presented information is very simple
- ◯ c. When there are many data points being depicted
- ◯ d. When there are few data points being depicted
- ◯ e. Never

An Icon Array should depict both the _____ representing the numerator and the ____ representing the denominator.

- a. Population at Risk, Affected Individuals
- b. Affected individuals, Population at risk
- c. Nontreated individuals, Treated individuals
- d. Treated individuals, Nontreated individuals
- e. Both b and d

Visual aids are especially useful for people with…

- a. People of low numeracy
- b. People with high levels of education
- c. Limited medical knowledge
- d. Both a and c
- e. All of the above

Numerical information in visual aids should be framed using…

- a. Decimals
- b. Fractions
- c. Frequencies
- d. Percentages
- e. All of the above

How should icon arrays be arranged?

- a. In a Block
- b. Random Scattering
- c. From smallest to largest
- d. From largest to smallest
- e. Does not matter how the icons are arranged

When using a graph with an axis (e.g bar/line graph) the y axis represents the _____ variable and the x axis represents the _____

- a. Vertical, Horizontal
- b. Horizontal, Vertical
- c. Independent, Dependent
- d. Dependent, Independent
- e. None of the above

A visual aid should always strive to be what above all else?

- a. Attractive
- b. Attention grabbing
- c. Transparent
- d. Complex
- e. Dynamic

Berlin Numeracy Test – S.

The Berlin Numeracy Test (Cokely et al., 2012) consists of 4 numeracy problems of increasing difficulty. An additional 3 items were included from the Schwartz (1997) numeracy scale to add a set of validated questions that individuals often find easier.

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.

1. 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? (Schwartz, 1997)

2. 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? (Schwartz, 1997)

3. 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? (Schwartz, 1997)

4. 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. (Cokely et al., 2012)

5. 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)? (Cokely et al., 2012)

6. 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? (Cokely et al., 2012)

7. 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. (Cokely et al., 2012)

Appendix B

Post-Test Measures and Tasks

Graph Literacy Scale.

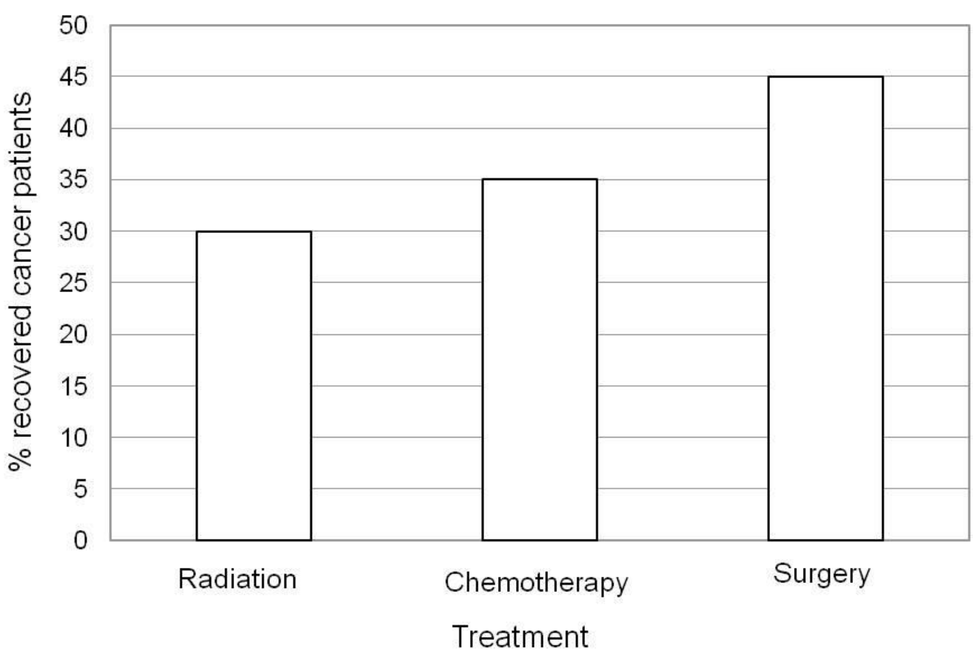
Questions 1-13 are the original graph literacy scale (Galesic & Garcia-Retamero, 2011) and 14-17 are the additional difficult items from the sources cited for each question. Correct responses are indicated in bold. The total score is the sum of correct responses. Higher scores mean higher skill levels.

1. Here is some information about cancer therapies.

What percentage of patients recovered after chemotherapy? __**35**__%

2. Here is some information about cancer therapies.

What is the difference between the percentage of patients who recovered after a surgery and the percentage of patients who recovered after radiation therapy? __**15**__%

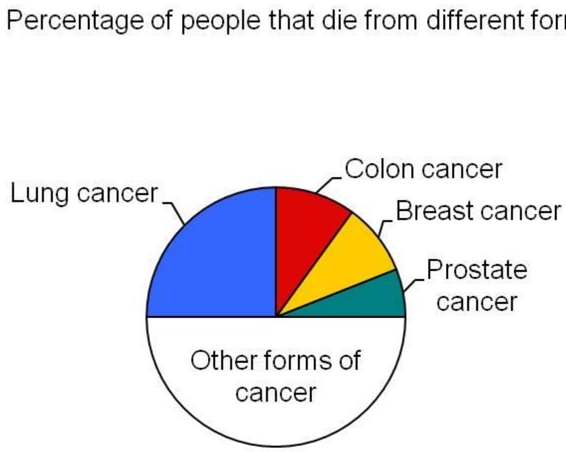


3. Here is some information about different forms of cancer.

Of all the people who die from cancer, approximately what percentage dies from lung cancer? __**25**__%

4. Here is some information about different forms of cancer.

Approximately what percentage of people who die from cancer die from colon cancer, breast cancer, and prostate cancer taken together? __**25**__%



5. Here is some information about an imaginary disease called Adeolitis.

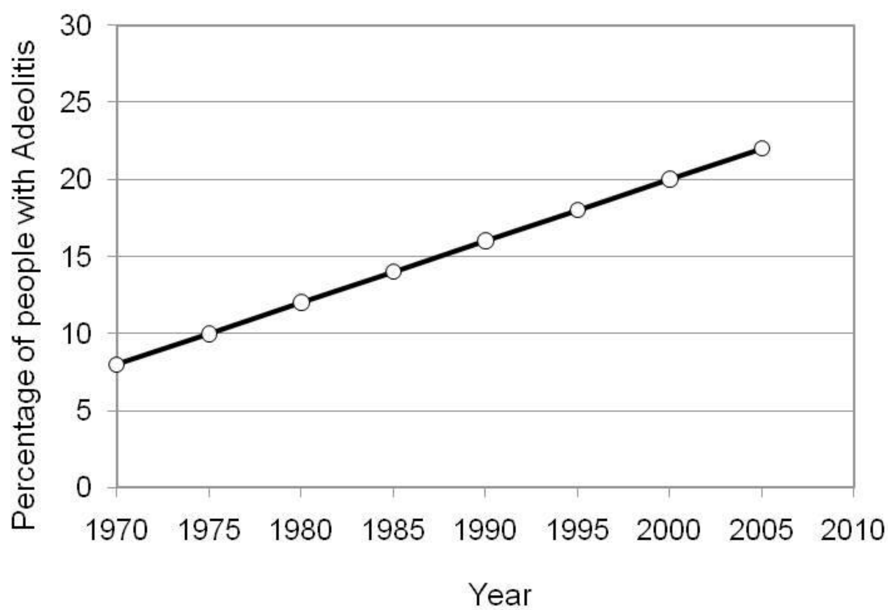
Approximately what percentage of people had Adeolitis in the year 2000? __**20**__%

6. Here is some information about an imaginary disease called Adeolitis.

When was the increase in the percentage of people with Adeolitis higher? [From 1975 to 1980, From 2000 to 2005**, Increase was the same in both intervals**, Don't know]

7. Here is some information about an imaginary disease called Adeolitis.

According to your best guess, what will the percentage of people with Adeolitis be in the year 2010? _**any value between 23 and 25**_

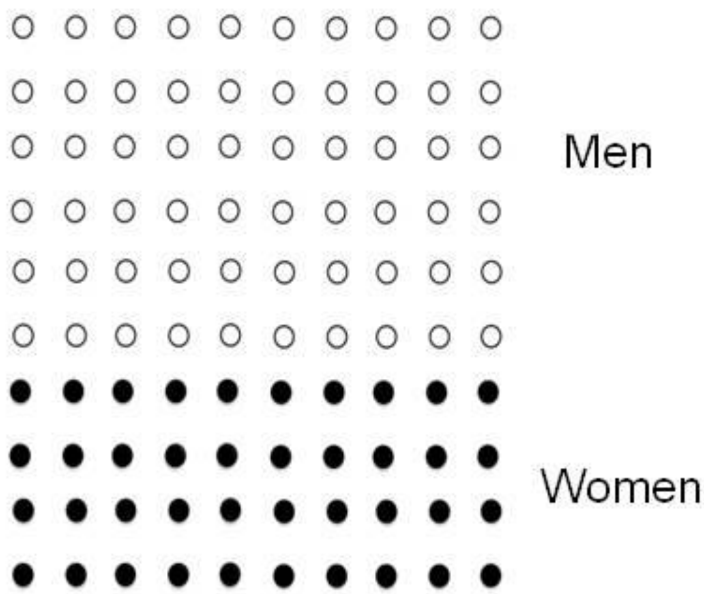


8. The following figure shows the number of men and women among patients with disease X. The total number of circles is 100.

Of 100 patients with disease X, how many are women? __**40**__

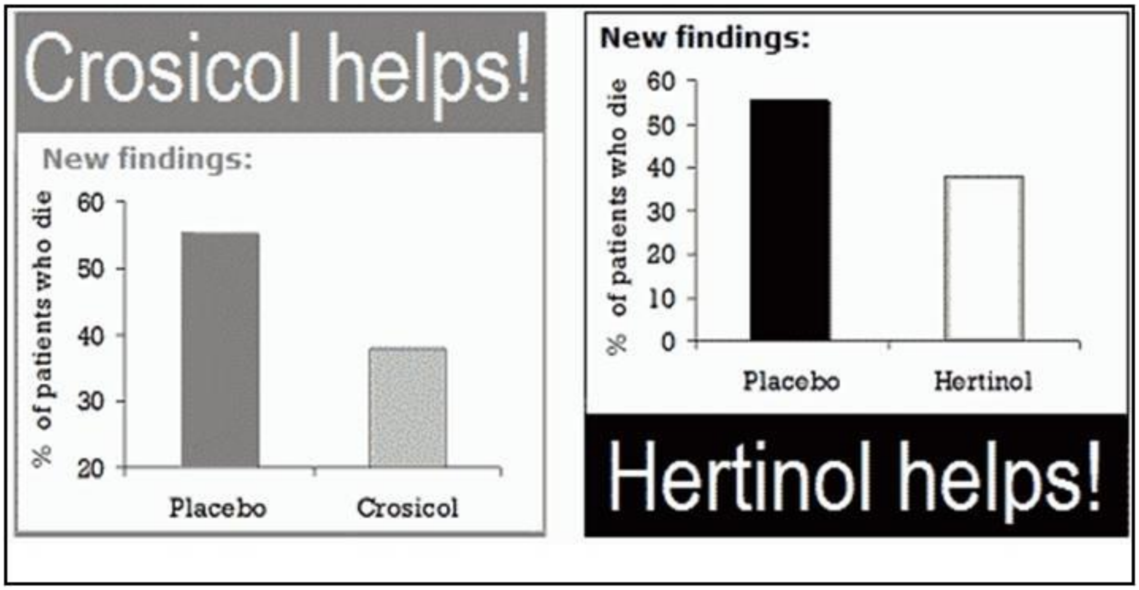
9. The following figure shows the number of men and women among patients with disease X. The total number of circles is 100.

How many more men than women are there among 100 patients with disease X? __**20**__



10. In a magazine you see two advertisements, one on page 5 and another on page 12. Each is for a different drug for treating heart disease, and each includes a graph showing the effectiveness of the drug compared to a placebo (sugar pill).

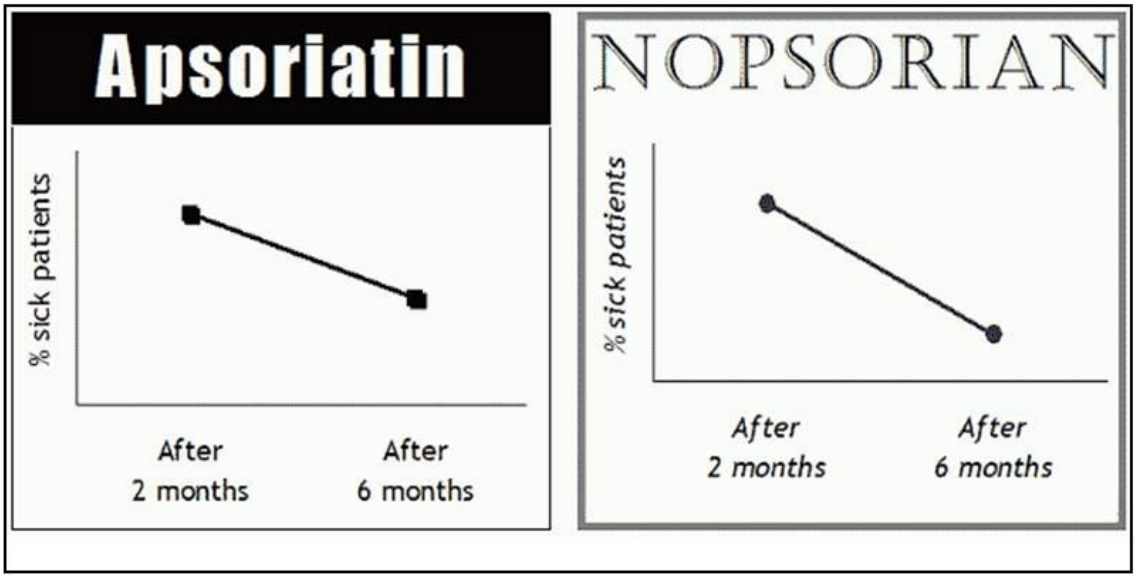
Compared to the placebo, which treatment leads to a larger decrease in the percentage of patients who die? [Crosicol, Hertinol, **They are equal**, Can’t say]



11. In the newspaper you see two advertisements, one on page 15 and another on page Each is for a different treatment of psoriasis, and each includes a graph showing the effectiveness of the treatment over time.

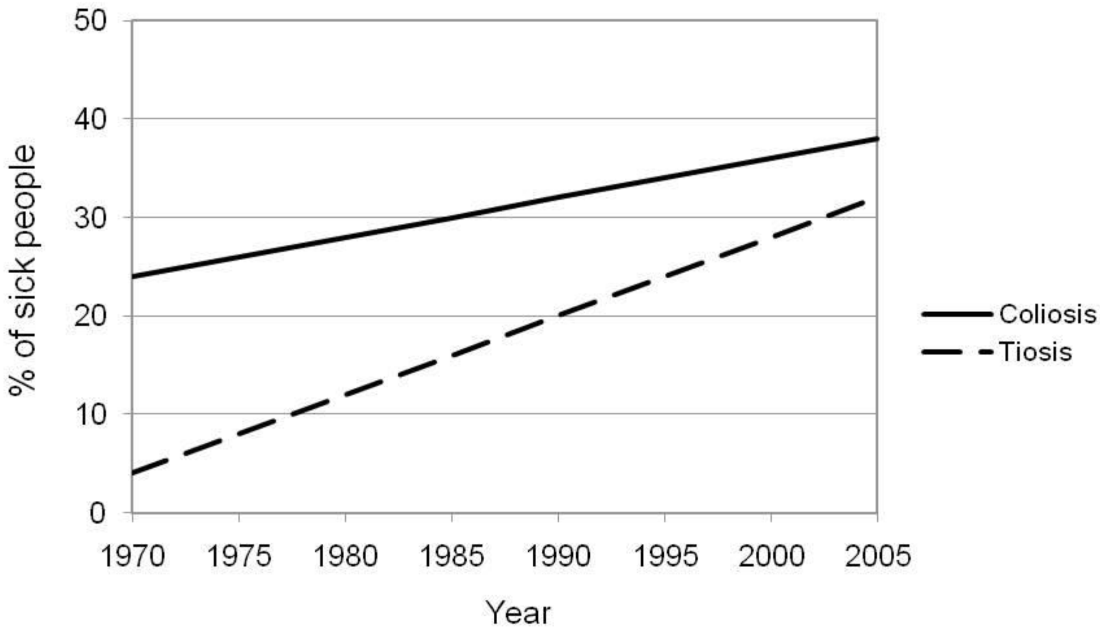
Which of the treatments contributes to a larger decrease in the percentage of sick patients?

[Apsoriatin, Nopsorian, They are equal, **Can't say**]



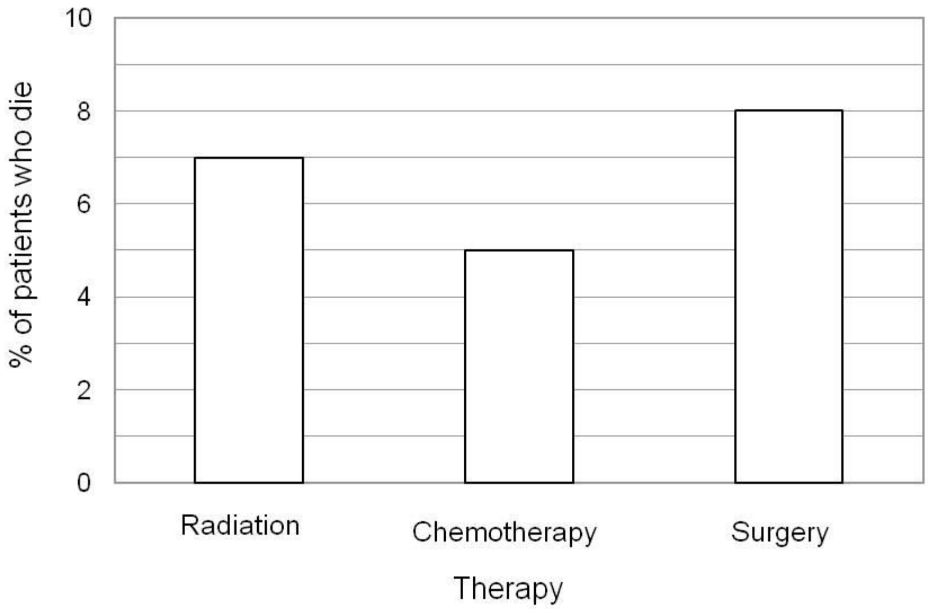
12. Here is some information about the imaginary diseases Coliosis and Tiosis.

Between 1980 and 1990, which disease had a higher increase in the percentage of people affected? [Coliosis, **Tiosis**, The increase was equal, Can't say]



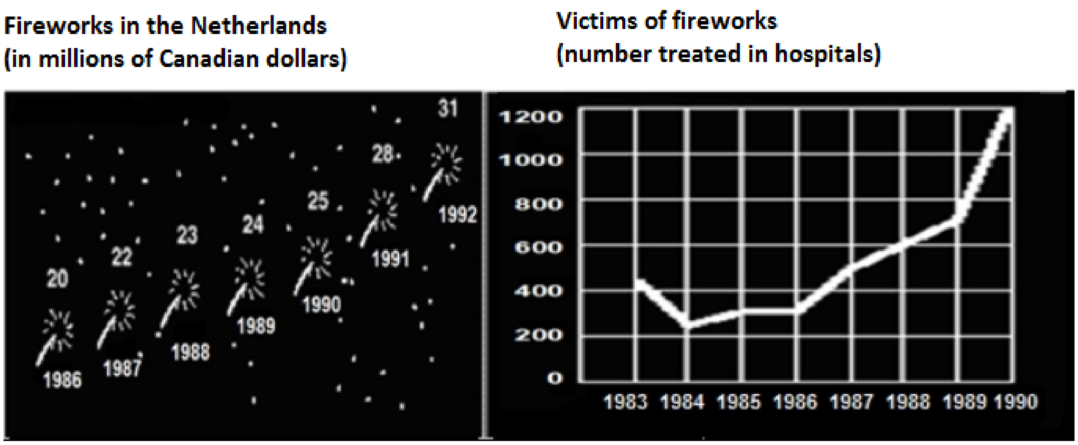
13. Here is some information about cancer therapies.

What is the percentage of cancer patients who die after chemotherapy? __**5**__%



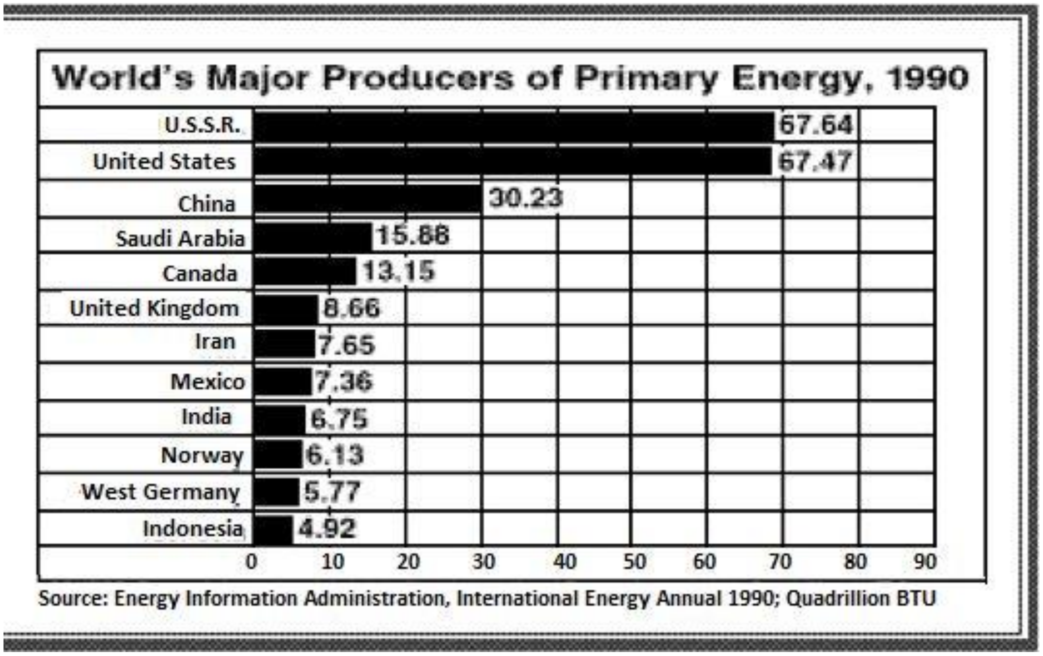
14. Below you can see information about Fireworks in the Netherlands. Graphs for question 14 of the graph literacy scale (Kirsch, 2001).

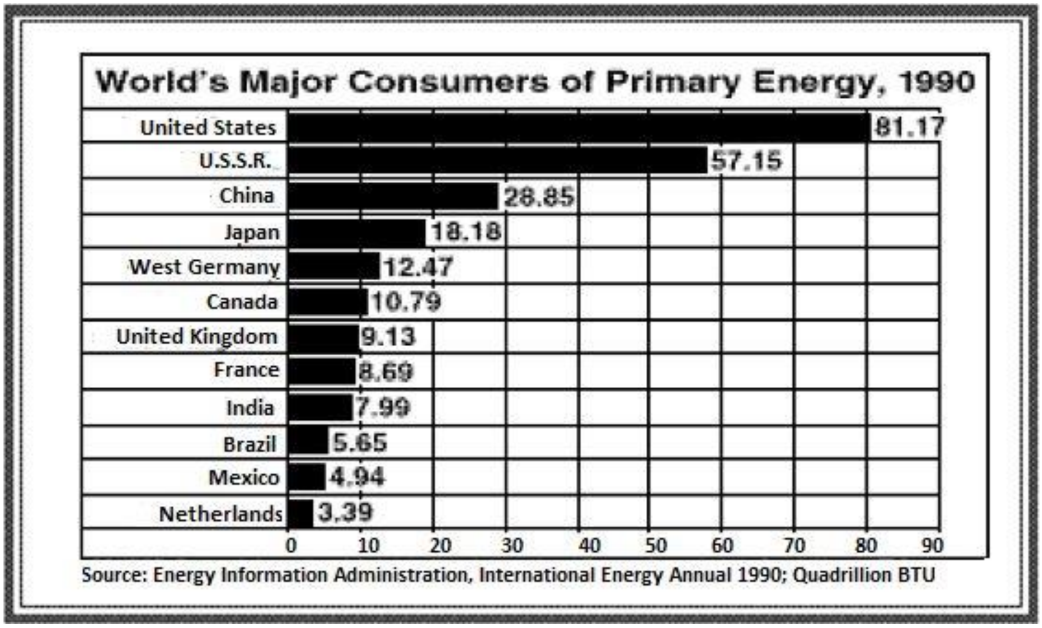
How many more people were injured in 1989 than in 1988? _**100**_ People



15. Below you can see information concerning the World's major producers and consumers of primary energy. Graphs for question 15 of the graph literacy scale (Evetts, 2002).

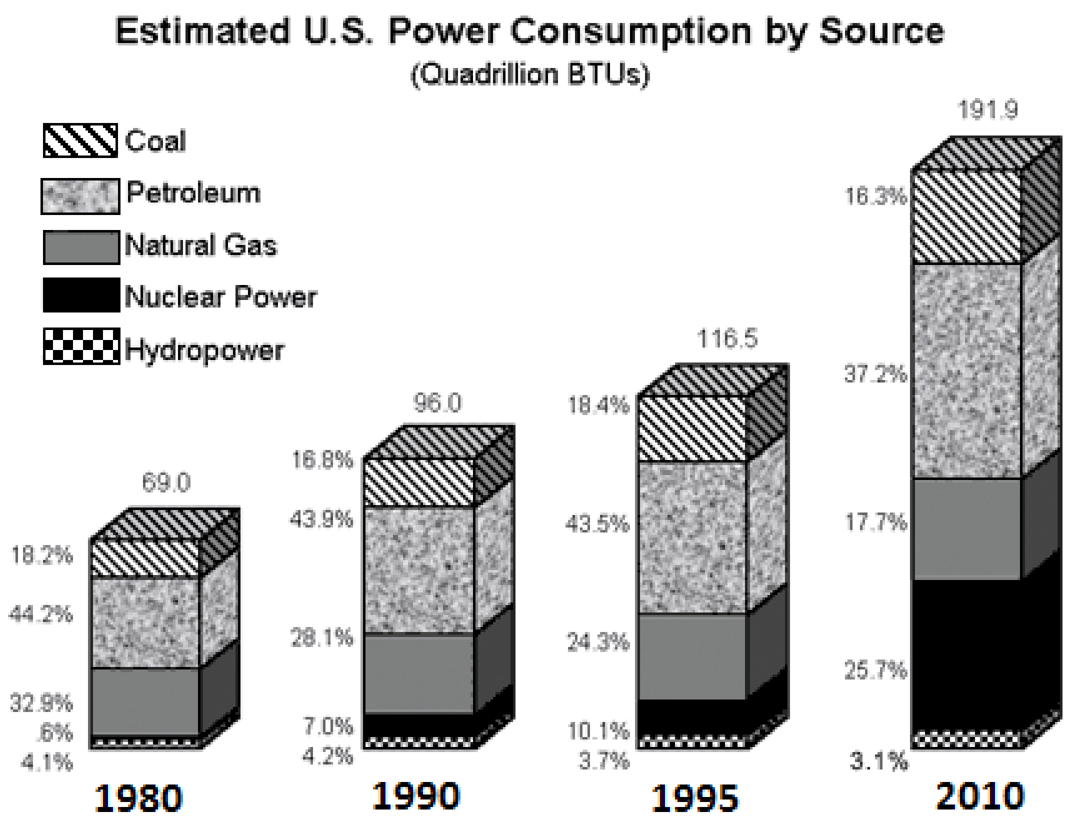
How much more energy is produced than consumed in Canada? _**2.36**_ Quadrillion BTU





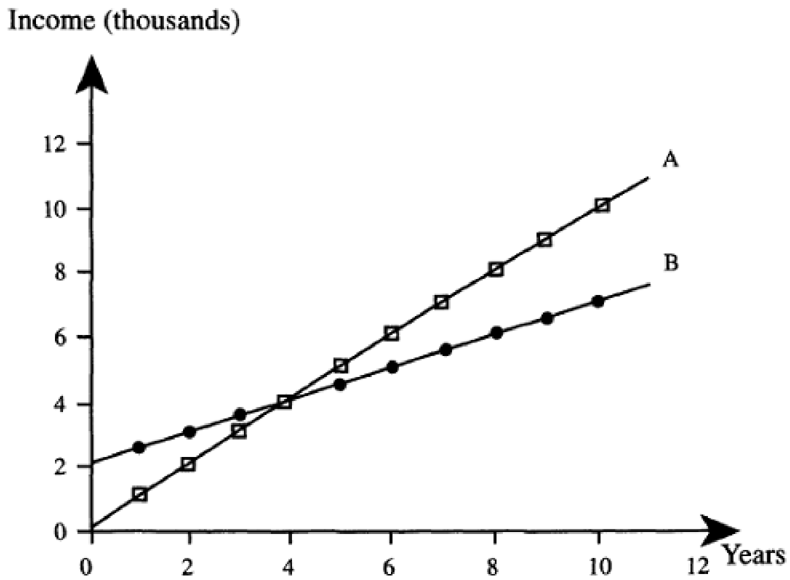
16. The graph below shows predictions of United States energy consumption through the year 2010. Use the graph to answer the question that follows. Graph for question 16 of the graph literacy scale (National Center for Education Statisics).

In the year 2010, which energy source is predicted to supply a larger percentage of the total power than it did in 1990? [Coal, Petroleum, Natural gas, **Nuclear power**, Hydro-power, I don't know]



17. Until the year 1994, the change rate in the income of Company B was... Graph for question 17 of the graph literacy scale (Kramarski & Mevarech, 2003).

[Greater than the change rate in the income of Company A, **Smaller than the change rate in the income of Company A**, Equal to the change rate in the income of Company A]



Visualizable Decision Making Tasks (Ybarra et al., 2018).

Tasks were created and psychometrically assessed to give a measure of an individual's skill at "visualizable" decision tasks. Each section is scored by adding up correct answers with higher numbers meaning more skill in that decision task. Bolded Answers are the correct answers

1. Ratio Bias

A. Imagine that you are presented with two bowls of folded tickets. One bowl contains 1 ticket marked "winner" and 9 blank tickets. The other bowl contains 10 tickets marked "winner" and 90 blank tickets. You must draw one ticket (without peeking, of course) from either bowl: If you draw a ticket marked "winner", you win $10.00; otherwise, you win nothing and the game is over. If you were given the choice, which bowl would you choose from?

- ◯ Bowl with 1 winning ticket
- ◯ Bowl with 10 winning tickets
- ◯ **No preference**

B. Imagine a situation in which you are watching TV late one night when you are startled by the telephone ringing. The caller says he is a policeman with unfortunate news to report. You learn that the person you love most in your life has been in an automobile accident and is in critical in a hospital. In desperation, you rush to the hospital. When you arrive there, the physician in charge of the case informs you that your loved one is in a coma and needs an immediate blood transfusion. You are asked to choose between samples from two blood banks, both of which have a rate of HIV transmission that is barely acceptable, but they are the only samples available at the moment. One has a record of one HIV positive case out of a thousand transfusions, and the other has a record of ten HIV positive cases out of ten thousand transfusions. You know that the proportions are identical, but somehow you may feel that the chance of transmission is greater in one sample than in the other.
In which sample, if either, do you think there was a greater likelihood of HIV contamination?

- ◯ Greater probability in 1,000 sample
- ◯ Slightly greater probability in 1,000 sample
- ◯ **Same probability in both samples**
- ◯ Slightly greater probability in 10,000 sample
- ◯ Greater probability in 10,000 sample

C. Suppose you are faced with two trays of jellybeans, one of which contained 10 beans and one with 100. In each , you must draw one jellybean without peeking, of course, from either tray. Should you pick a black jelly bean, you will win $50, Should you pick a white jellybean, you will win nothing.

Now, what makes your choice interesting is that for each time you choose, both trays will offer identical odds of selecting a black jellybean. Still, many (but not all) people have a preference for picking one of the trays. Your job is to indicate whether you would prefer to pick from the large tray (100 beans), the small tray (10 beans) or if you have no preference.

○ Tray with 10 jelly beans

○ **No preference**

○ Tray with 100 jelly beans

D. You are informed by a new study of the WHO (World Health Organization) that, in Peru, the risk for a woman to be a victim of domestic violence by her partner is 40 in 100. Women are also likely to experience verbal abuse, which is a form of abuse in which a partner says harsh words to bully and belittle a person. Another form of verbal abuse is silent, in which a person acts as if their partner does not exist. Women have a 2 in 5 chance of experiencing verbal abuse. Is domestic violence or verbal abuse a bigger problem in Peru?

○ Domestic abuse is significantly bigger problem

○ Domestic abuse is a slightly bigger problem

○ **Problems have same magnitude**

○ Verbal abuse is a slightly bigger problem

○ Verbal abuse is a significantly bigger problem

E. You have a coupon for two different pizza places. In both locations, a large pizza is $10. For Norman Pizzeria, you can save 20% on one pizza. At Paul’s Pizza, you receiving $2 off each pizza when you order two pizzas. At which pizza place will you save the largest percent of money?

○ Norman Pizzeria

○ **You will save the same amount at both pizza places**

○ Paul's Pizza

Out of the above questions, how many do you think you got right?

- ○ 1
- ○ 2
- ○ 3
- ○ 4
- ○ 5

How confident are you that your previous answers are correct?

- ○ 1 - Not confident at all
- ○ 2
- ○ 3
- ○ 4
- ○ 5 - Half confident the answers are correct
- ○ 6
- ○ 7
- ○ 8
- ○ 9
- ○ 10 - Fully confident

2. Sunk Cost

A. Imagine on your way home you buy a tv dinner on sale for $3 at the local grocery store. A few hours later you decide it is time for dinner, so you get ready to put the tv dinner in the oven. Then you get an idea. You call up your friend to ask if he would like to come over for a quick tv dinner and then watch a good movie on tv. Your friend says “Sure.” So you go out to buy a second tv dinner. However, all the on-sale tv dinners are gone. You therefore have to spend $5 (the regular price) for the tv dinner identical to the one you just bought for $3. You go home and put both dinners in the oven. When the two dinners are fully cooked, you get a phone call. Your friend is ill and cannot come. You are not hungry enough to eat both dinners. You cannot freeze one. You must eat one and discard the other. Which one do you eat?

- ◯ $3 dinner
- ◯ $5 dinner
- ◯ **no preference**

B. Agatha decides after careful consideration that she wants to take cello lessons. Agatha spends $1,000 on a beginner cello and an additional $200 on the first 3 months of cello lessons. After 3 months of lessons, Agatha realizes that she no longer enjoys the cello and wants to stop taking lessons. It is almost certain that if she signs up for more lessons, she will not enjoy them and will never enjoy playing the cello. What should Agatha do?

- ◯ **Agatha should stop taking cello lessons because it would be a waste of time and money to take more lessons that she won't enjoy**
- ◯ Agatha should continue with the lessons because otherwise she will have wasted the money and time already spent
- ◯ Agatha should continue lessons to teach herself that next time she should be more careful about what hobbies she selects for herself
- ◯ Agatha should continue with the lessons because if she was foolish enough to select a hobby that she doesn't enjoy, she deserves to suffer by continuing with her lessons
- ◯ Agatha should continue with the lessons. If she stops taking lessons, that would mean she made a bad decision in deciding to take cello lessons.... If it was the right decision then, it is still the right decision

C. Imagine that your credit card company selects you to win a free weekend cruise to Macau in exchange for some market research that you must participate in. The research involves participation in focus groups and questionnaires- it takes a total of about 12 hours spread over two weeks. You participated in the research and collected your free ticket to the cruise. However, a couple of days before the cruise you get invited to an all-expenses paid weekend trip to Bangkok for the same weekend as the cruise. You would love to visit Bangkok, but have also really been looking forward to this cruise. How likely are you to go on the cruise?

○ Extremely likely

○ Moderately likely

○ Slightly likely

○ **Neither likely nor unlikely**

○ **Slightly unlikely**

○ **Moderately unlikely**

○ **Extremely unlikely**

D. You go to the bar, on the spur of the moment, and buy a pitcher of beer for $6.95. You finish the beer and then decide to buy a second pitcher of beer for the same price. But after a few sips you feel you had too much to drink. No one else wants to drink it. How likely are you to keep drinking the rest of pitcher?

- Extremely likely
- Moderately likely
- Slightly likely
- **Neither likely nor unlikely**
- **Slightly unlikely**
- **Moderately unlikely**
- **Extremely unlikely**

E. You purchase a 3-month supply of weight loss products for $200. When you try the product, you discover that it tastes horribly, but you still take use it everyday because you really want to lose weight. The box says you should see results within 6 weeks, but you have used the products as instructed for 10 weeks and have seen no results. The purchase is non-refundable. How likely are you to continue using the weight loss products?

- Extremely likely
- Moderately likely
- Slightly likely
- **Neither likely nor unlikely**
- **Slightly unlikely**
- **Moderately unlikely**
- **Extremely unlikely**

F. Imagine that you paid $20 two weeks ago for a ticket to attend a concert. You have really been looking forward to going to this concert. The day before the concert you find that you have lost the ticket. Later that day you meet someone who has a ticket which you could buy for $10. How likely are you to purchase another ticket?

- ○ Extremely likely
- ○ Moderately likely
- ○ Slightly likely
- ○ **Neither likely nor unlikely**
- ○ **Slightly unlikely**
- ○ **Moderately unlikely**
- ○ **Extremely unlikely**

How many of the above questions did you answer correctly?

- ○ 1
- ○ 2
- ○ 3
- ○ 4
- ○ 5
- ○ 6

How confident are you that your previous answers are correct?

- ◯ 1 - Not confident in your answers
- ◯ 2
- ◯ 3
- ◯ 4
- ◯ 5 - Half confident in your answers
- ◯ 6
- ◯ 7
- ◯ 8
- ◯ 9
- ◯ 10 - Fully confident in your answers

3. Framing.

This section has two near identical questions that are altered to assess the framing effect. In order to avoid recency, it is suggested to not have both questions near one another.

a-1.

In this brief survey we want to know what associations or thoughts come to mind when making consumer purchases. We will present you with pairs of possible associates. In each pair we want you to indicate by filling in one of the squares which item in the pair you are most apt to associate with a purchase of 75%

lean ground beef and the extent to which you associate the purchase with that item rather than the other item in the pair.

	1	2	3	4	5	
Good-tasting	○	○	○	○	○	Bad-tasting
Greasy	○	○	○	○	○	Greaseless
High quality	○	○	○	○	○	low quality
fat	○	○	○	○	○	lean

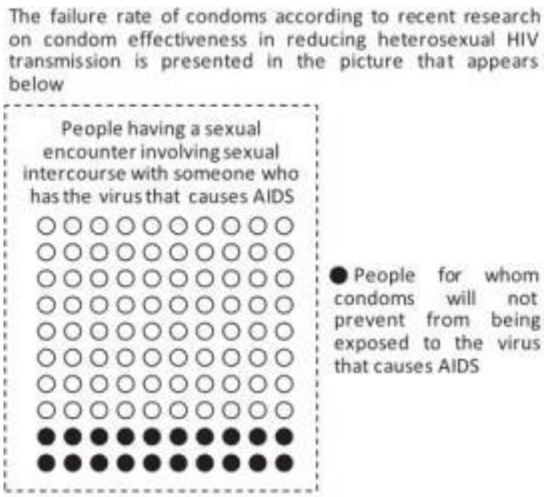
a-2.

In this brief survey we want to know what associations or thoughts come to mind when making consumer purchases. We will present you with pairs of possible associates. In each pair we want you to indicate by filling in one of the squares which item in the pair you are most apt to associate with a purchase of 25% fat ground beef and the extent to which you associate the purchase with that item rather than the other item in the pair.

	1	2	3	4	5	
Good-tasting	○	○	○	○	○	Bad-tasting
Greasy	○	○	○	○	○	Greaseless
High quality	○	○	○	○	○	low quality
fat	○	○	○	○	○	lean

b-1.

Please read the following infographic

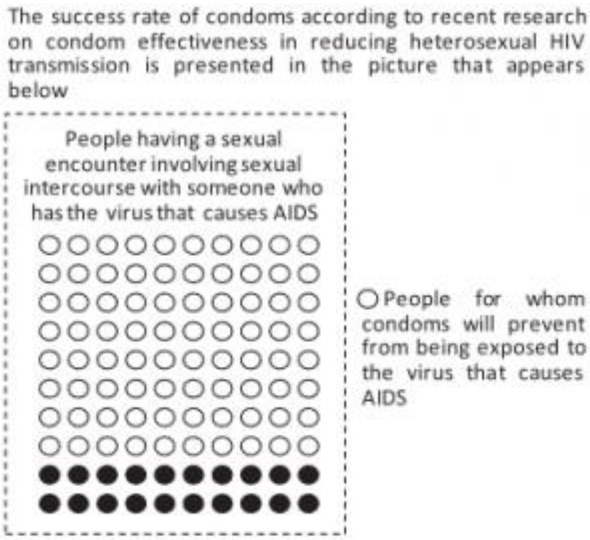


After reading the infographic above, what is the likelihood you would use a condom during your next sexual encounter? (please assume this as a hypothetical situation)

- ◯ 1 - I have no intention of doing this
- ◯ 2
- ◯ 3
- ◯ 4
- ◯ 5
- ◯ 6
- ◯ 7
- ◯ 8
- ◯ 9 - I am certain that I will do this

b-2.

Please read the following infographic



After reading the infographic above, what is the likelihood you would use a condom during your next sexual encounter? (please assume this as a hypothetical situation)

- 1 - I have no intention of doing this
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9 - I am certain that I will do this

c-1.

Imagine that you are approached by an individual to gamble. You are given $20. You can either keep the $20 and NOT gamble or you can pay the $20 to take the gamble.

If you gamble you have the chance to win $150. The probability of losing is 85%.

Please rate on a scale of 1-10 on how likely you are to take the gamble. (1 = would definitely not take the gamble, 10 = definitely would take the gamble).

- ○ 1 - would definitely not take the gamble
- ○ 2
- ○ 3
- ○ 4
- ○ 5
- ○ 6
- ○ 7
- ○ 8
- ○ 9
- ○ 10 - definitely would take the gamble

c-2.

Imagine that you are approached by an individual to gamble. You are given $20. You can either keep the $20 and NOT gamble or you can pay the $20 to take the gamble.

If you gamble you have the chance to win $150. The probability of winning is 15%.

Please rate on a scale of 1-10 on how likely you are to take the gamble. (1 = would definitely not take the gamble, 10 = definitely would take the gamble).

- ○ 1 - would definitely not take the gamble
- ○ 2
- ○ 3
- ○ 4
- ○ 5
- ○ 6
- ○ 7
- ○ 8
- ○ 9
- ○ 10 - definitely would take the gamble

How confident are you that your above answers are correct?

- ○ 1 - Not confident at all
- ○ 2
- ○ 3
- ○ 4
- ○ 5 - Half confident the answers are correct
- ○ 6
- ○ 7
- ○ 8
- ○ 9
- ○ 10 - Fully confident

Probabilistic Numeracy Questions.

Nine probabilistic numeracy questions from Ghazal (2014) were used to asses pre-post test measures of numeracy.

1. 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?

- ○ 1 time out of 6 rolls of the dice
- ○ 5 times out of 6 rolls of the dice
- ○ 1 time out of 2 rolls of the dice
- ○ 1 out of 5 rolls of the dice
- ○ 6 out of 1 roll of the dice

2. 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 out of 36 rolls of the dice
- ○ 3 out of 6 rolls of the dice
- ○ 1 out of 4 rolls of the dice
- ○ 2 out of 6 rolls of the dice
- ○ 2 out of 36 rolls of the dice

3. 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?

- ◯ About 195
- ◯ About 150
- ◯ About 135
- ◯ About 80
- ◯ About 64

4. Suppose you are taking an 8 question multiple choice test and each question has 4 options. Imagine that you don't know anything about the test and so you guess without reading the questions. What's the probability that you would get 100% correct on this test just by chance alone?

- ◯ 1/4
- ◯ 1/8
- ◯ 1/4096
- ◯ 1/16384
- ◯ 1/65536

5. Imagine that you are throwing 6 dice up in the air. What is the probability that all of them would land on even numbers?

- ○ 1/432
- ○ 3/216
- ○ 1/64
- ○ 3/6
- ○ 1/21

6. Imagine you are drawing a picture, and are missing 2 spots you want to color. There are 7 colors to choose from. What's the probability that both spots end up colored orange?

- ○ 1/49
- ○ 2/49
- ○ 1/7
- ○ 2/7
- ○ 6/7

7. Imagine you are throwing 8 dice up in the air. What's the probability that half will land on an even number, while the other half land on 1?

- ◯ 1/10368
- ◯ 1/20736
- ◯ 1/432
- ◯ 1/1728
- ◯ 1/6

8. Phil is holding 4 cards in his hand: 8 of clubs, 5 of hearts, king of hearts, and ace of diamonds. If he places them on a table in random order, what is the probability that the first and last cards will both be hearts?

- ◯ 1/2
- ◯ 1/3
- ◯ 1/4
- ◯ 1/6
- ◯ 1/8

9. n is an integer chosen at random from the set {5, 7, 9, 11}
p is chosen at random from the set {2, 6, 10, 14, 18}
What is the probability that n + p = 23?

- ◯ 0.1
- ◯ 0.2
- ◯ 0.25
- ◯ 0.3
- ◯ 0.4

Bias Blind Spot.

The Bias Blind Spot is a measure of one's general overplacement confidence of bias vulnerability. It is scored by asking for each question "To what extent do you believe that you are likely to commit the bias?". Then score from not at all (1) to very often (7). Then a second question asking, "To what extent do you believe the average individual would commit the bias?". Then score from not at all (1) to very often (7). A difference score is computed for each question by the following equation: "others score" – "yourself score" = Bias Blind Spot Score. Then for the Bias Blind Spot Score is aggregated and averaged for all questions. Previous studies (Ybarra, 2018) found that the original (Scopelliti et al., 2015) and new "Cognitive" items from Ybarra (2018) were correlated, but largely different. It is recommended to create an aggregate of all questions then score all questions from separate studies independently.

1. Some people show a tendency to judge a harmful action as worse than an equally harmful inaction. For example, this tendency leads to thinking it is worse to falsely testify in court that someone is guilty, than not to testify that someone is innocent. (Scopelliti et al., 2015)

2. Many psychological studies have found that people react differently to presented information depending on how it is "framed". For example, information framed positively will be interpreted differently compared to the same information framed negatively. (Scopelliti et al., 2015)

3. Psychologists have claimed that some people show a tendency to do or believe a thing only because many other people believe or do that thing, to feel safer or to avoid conflict. (Scopelliti et al., 2015)

4. Many psychological studies have shown that people react to counterevidence by strengthening their beliefs. For example, when exposed to negative evidence about their favorite political candidate, people tend to implicitly counterargue against that evidence, therefore strengthening their favorable feelings toward the candidate. (Scopelliti et al., 2015)

5. Some people show a tendency to continue an endeavor once an investment in money, effort, or time has been made. This tendency leads to thinking that it is better to invest further into an endeavor even if it is better to not. For example, sitting through a movie after you realize it was awful halfway through. (Scopelliti et al., 2015)

6. Some people show a tendency to choose probabilities that are larger ratios over probabilities that are equal or superior but are expressed with smaller ratios. For example, people would choose a drug that cures 10/100 patients over 1/10 patients, even though they cure equal amounts of individuals. (Scopelliti et al., 2015)

7. Psychologists have claimed that some people show a "disconfirmation" tendency in the way they evaluate research about potentially dangerous habits. That is, they are more critical and skeptical in evaluating evidence that an activity is dangerous when they engage in that activity than when they do not. (Scopelliti et al., 2015)

8. Psychologists have identified an effect called “diffusion of responsibility,” where people tend not to help in an emergency when other people are present. This happens because as the number of bystanders increases, a bystander who sees other people standing around is less likely to interpret the incident as a problem, and also is less likely to feel individually responsible for taking action. (Scopelliti et al., 2015)

9. Research has found that people will make irrational decisions to justify actions they have already taken. For example, when two people engage in a bidding war for an object, they can end up paying much more than the object is worth to justify the initial expenses associated with bidding. (Scopelliti et al., 2015)

10. Extensive psychological research has shown that when presented with a new situation that people will compare the likelihood of the event to an event we think is similar. For example, people think medical symptoms resemble their causes, for instance people mistakenly believe ulcers are caused by stress. (Scopelliti et al., 2015)

11. Psychologists have claimed that some people show a tendency to make “overly dispositional inferences” in the way they view victims of assault crimes. That is, they are overly inclined to view the victim’s plight as one he or she brought on by carelessness, foolishness, misbehavior, or naivetë. (Scopelliti et al., 2015)

12. Many psychological studies have found that people tend to weigh their judgments toward more recent information. That is, people may believe that an event is more likely to happen based on the amount of current news coverage rather than actual prevalence. (Scopelliti et al., 2015)

13. Psychologists have claimed that people exhibit a “gambler’s fallacy” when an event happens more frequently than normal as it is falsely thought that there is a chance that it will happen less frequently in the future and vice-versa. For example, when you are playing a slot machine and losing, someone may believe that they are “due” for a win. (Scopelliti et al., 2015)

14. Psychologists have claimed that some people show a “halo” effect in the way they form impressions of attractive people. For instance, when it comes to assessing how nice, interesting, or able someone is, people tend to judge an attractive person more positively than he or she deserves. (Scopelliti et al., 2015)

15. Extensive psychological research has shown that people possess an unconscious, automatic tendency to be less generous to people of a different race than to people of their race. This tendency has been shown to affect the behavior of everyone from doctors to taxi drivers. (Ybarra, 2018)

16. Psychologists have identified a tendency called the “ostrich effect,” an aversion to learning about potential losses. For example, people may try to avoid bad news by ignoring it. The name comes from the common (but false) legend that ostriches bury their heads in the sand to avoid danger. (Ybarra, 2018)

17. Psychologists have claimed that people believe that two events happening together is more probable than one event. For example, people believe it is more probable that Katy Perry will release a hit single and win a Grammy than if Katy Perry will just release a hit song. (Ybarra, 2018)

18. Many psychological studies have found that people have the tendency to underestimate the impact or the strength of another person's feelings. For example, people who have not been victims of discrimination do not really understand a victim's social suffering and the emotional effects of discrimination. (Ybarra, 2018)

19. Psychologists have claimed that some people show a "self-interest" effect in the way they view political candidates. That is, people's assessments of qualifications, and their judgments about the extent to which particular candidates would pursue policies good for the American people as a whole, are influenced by their feelings about whether the candidates' policies would serve their own particular interests. (Ybarra, 2018)

20. Psychological research has shown that people may display a "knew-it-all-along" effect after an event has occurred. That is, people will see an outcome as predictable, despite there being little basis to predict it beforehand. (Ybarra, 2018)

21. People often fail to notice the statistical likelihood that after an initial measurement of an extreme score, the following scores will move towards the average after additional measurements. For example, people tend to think that a rookie baseball player who breaks records in their first year will continue to excel, but the athlete's performance the next year is likely to be average compared to other players. (Ybarra, 2018)

22. Psychologists have claimed that some people show a "self-serving" tendency in the way they view their academic or job performance. That is, they tend to take credit for success but deny responsibility for failure. They see their successes as the result of personal qualities, like drive or ability, but their failures as the result of external factors, like unreasonable work requirements or inadequate instructions. (Ybarra, 2018)

23. Psychologists have argued that gender biases lead people to associate men with technology and women with housework. (Ybarra, 2018)

Financial Literacy (Lusardi & Mitchell, 2009).

1. Basic Financial Literacy

Suppose you had $100 in a savings account and the interest rate was 2% per year. After 5 years, how much do you think you would have in the account if you left the money to grow?

- ◯ **More than $102**
- ◯ Exactly $102
- ◯ Less than $102
- ◯ Don't know
- ◯ Refuse to respond

Suppose you had $100 in a savings account and the compound interest rate is 20% per year and you never withdraw money or interest payments. After 5 years, how much would you have on this account in total?

- ◯ **More than $200**
- ◯ Exactly $200
- ◯ Less than $200
- ◯ Don't know
- ◯ Refuse to respond

Imagine that the interest rate on your savings account was 1% per year and inflation was 2% per year. After 1 year, how much would you be able to buy with the money in this account?

- ○ More than today
- ○ Exactly the same amount as today
- ○ **Less than today**
- ○ Don't know
- ○ Refuse to respond

Suppose that in the year 2020, your income has doubled and prices of all goods have doubled too. In 2020, how much will you be able to buy with your income?

- ○ More than today
- ○ **The same amount as today**
- ○ Less than today
- ○ Don't know
- ○ Refuse to answer

2. Sophisticated Financial Literacy
Which of the following statements describes the main function of the stock market?

- ○ The stock market helps to predict stock earnings
- ○ The stock market results in an increase in the price of stocks
- ○ **The stock market brings those who want to buy stock together with those who want to sell stock**
- ○ None of the above
- ○ Don't know
- ○ Refuse to respond

Which of the following statements is correct?

- ○ Once one invests in a mutual fund, one cannot withdraw the money in the first year
- ○ Mutual funds can invest in several assets, for example invest in both stocks and bonds
- ○ Mutual funds pay a guaranteed rate of return which depends on their past performance
- ○ **None of the above**
- ○ Don't know
- ○ Refuse to respond

If the interest rate falls, what should happen to bond prices?

- ○ **Rise**
- ○ Fall
- ○ Stay the same
- ○ Don't know
- ○ Refuse to respond

True or false? Buying a company stock usually provides a safer return than a stock mutual fund.

- ◯ True
- ◯ **False**
- ◯ Don't know
- ◯ Refuse to respond

True or false? Stocks are normally riskier than bonds.

- ◯ **True**
- ◯ False
- ◯ Don't know
- ◯ Refuse to respond

Considering a long time period (for example 10 or 20 years), which asset normally gives the highest return?

- ◯ Savings accounts
- ◯ Bonds
- ◯ **Stocks**
- ◯ Don't know
- ◯ Refuse to respond

Normally, which asset displays the highest fluctuations over time?

- ◯ Savings accounts
- ◯ Bonds
- ◯ **Stocks**
- ◯ Don't know
- ◯ Refuse to respond

When an investor spreads his money among different assets, does the risk of LOSING money:

- ◯ Increase
- ◯ **Decrease**
- ◯ Stay the same
- ◯ Don't know
- ◯ Refuse to respond

Appendix C
Limitations

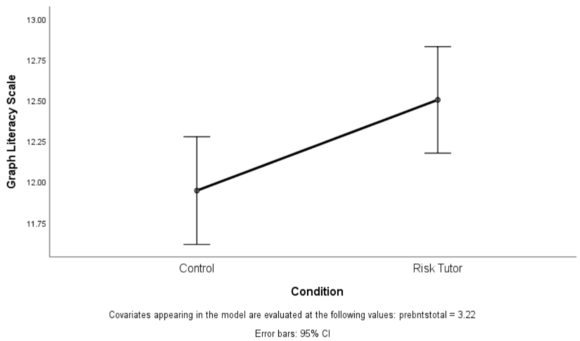
A significant amount of attrition in participants returning for part 2 in a completed third condition led to removing the condition from statistical analyses. The third condition had individuals complete the risk tutor then two weeks later complete the post-test measures at home. This condition was removed as only 88 participants returned to complete part 2 compared to the other conditions where approximately 135 participants returned. Reasons for the attrition could not be identified; thus, the third condition was removed. Unfortunately, this removal of the participants led to unequal timing in some post-test measurements (e.g., graph literacy) between conditions. The control condition completed all post-test measures at home, while the risk literacy tutor condition completed the post-test measures immediately after completing the tutor in the laboratory. There were no other differences between conditions. Due to this, there is a chance that some between-group differences may be due to differences in testing location (in-lab vs. at home).

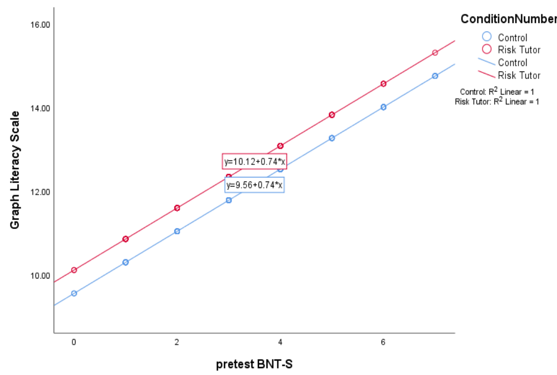
Appendix D
Supplementary Analyses

1. ANCOVA Graphs for Post-test Variable.

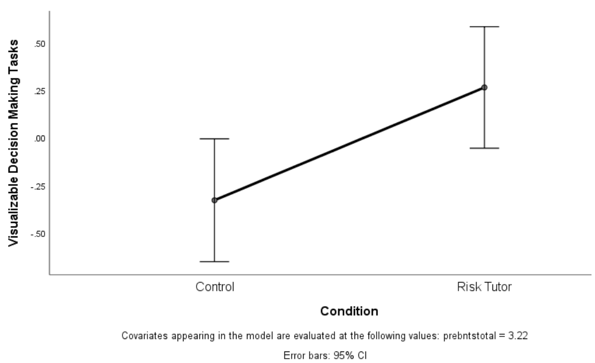
Results indicated that there were no interactions of pre-test statistical numeracy with the condition.

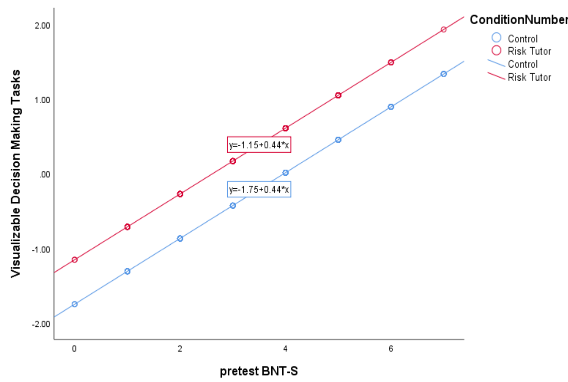
a. Graph Literacy



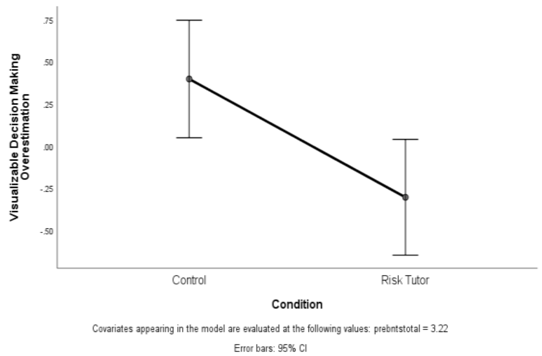


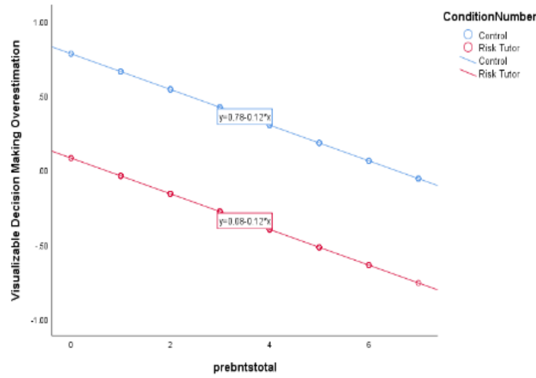
b. Visualizable Decision Making



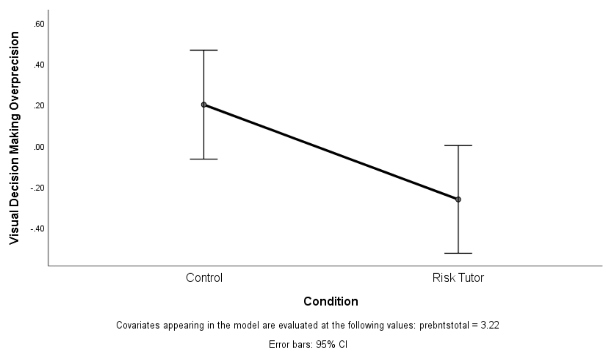


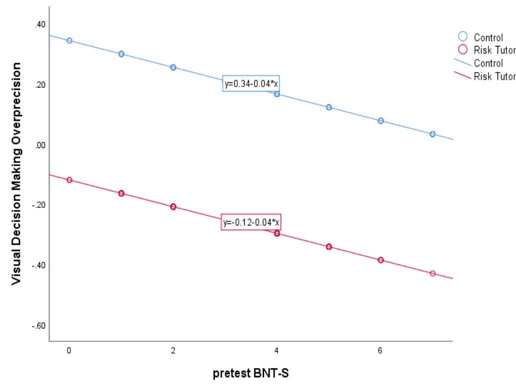
c. Overestimation



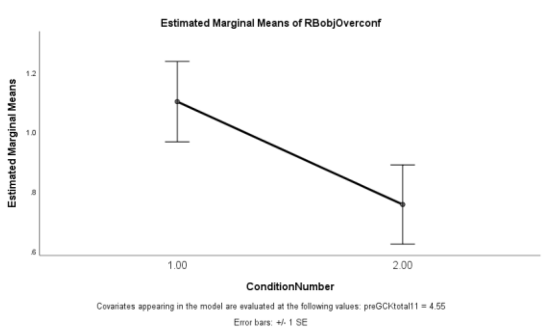


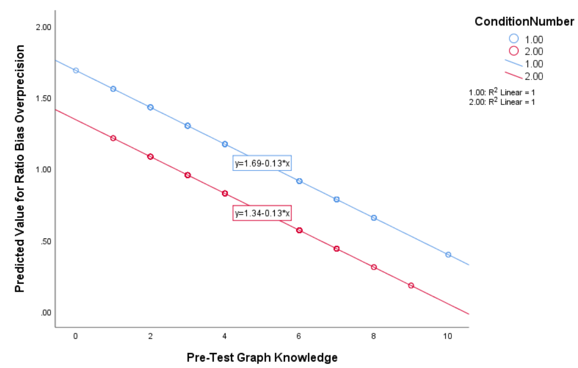
d. Overprecision



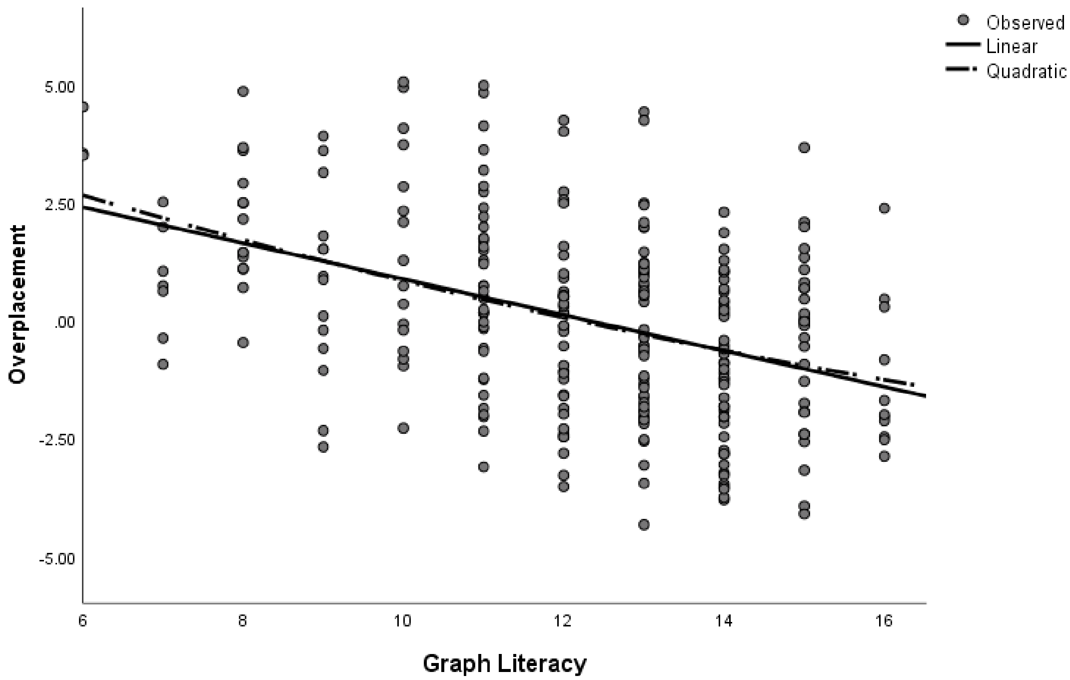


e. Pregraph construction knowledge and overestimation





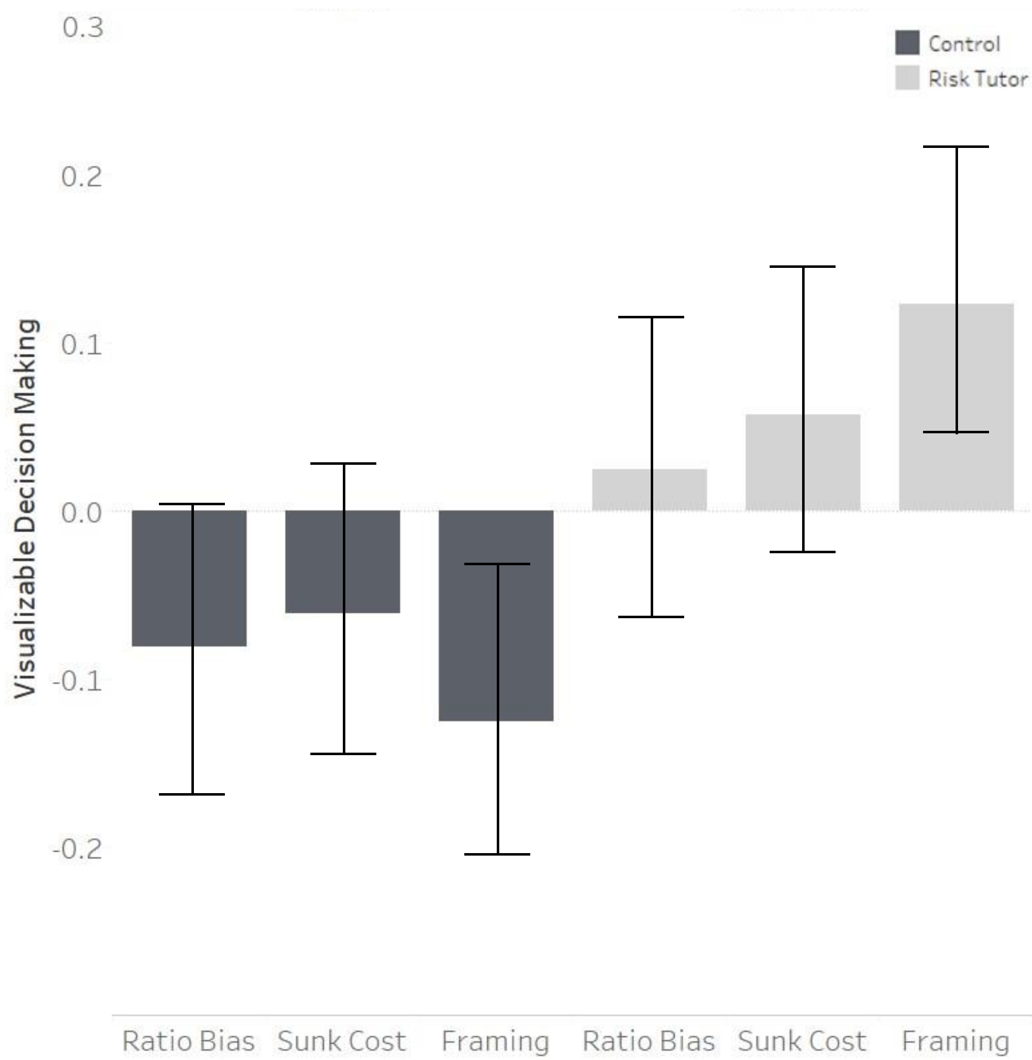
The relationship between post-test graph literacy and overplacement. Further evidence of the relationship between risk literacy skills and social comparisons.



2. Condition differences on visualizable decision making tasks.

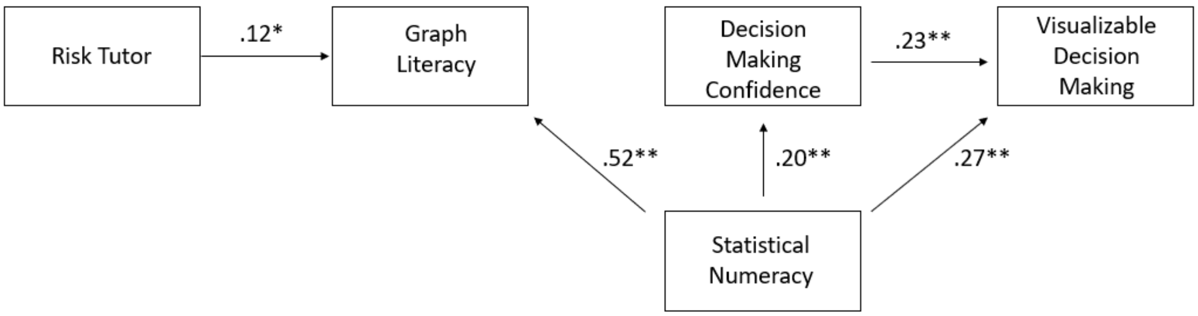
Regression analyses show trending improvements for ratio bias and sunk cost and significant improvements in framing effects for those when

completed the risk tutor. Improvements in untrained, individual visualizable decision tasks can be seen when incorporating trained risk literacy skills (e.g., graph literacy) into the regression models.



3. Competing SEM Models

The competing structural equation model shows how there is no relation between graph literacy and decision making confidence. Results lend evidence to the idea that the presented model in Figure 14 was selected. This model additionally showed poor fit : $\chi2$ (3) = 32.35, CFI = .84, TLI = .52, RMSEA = .19 (0.13 – 0.25), SRMR = 0.07.



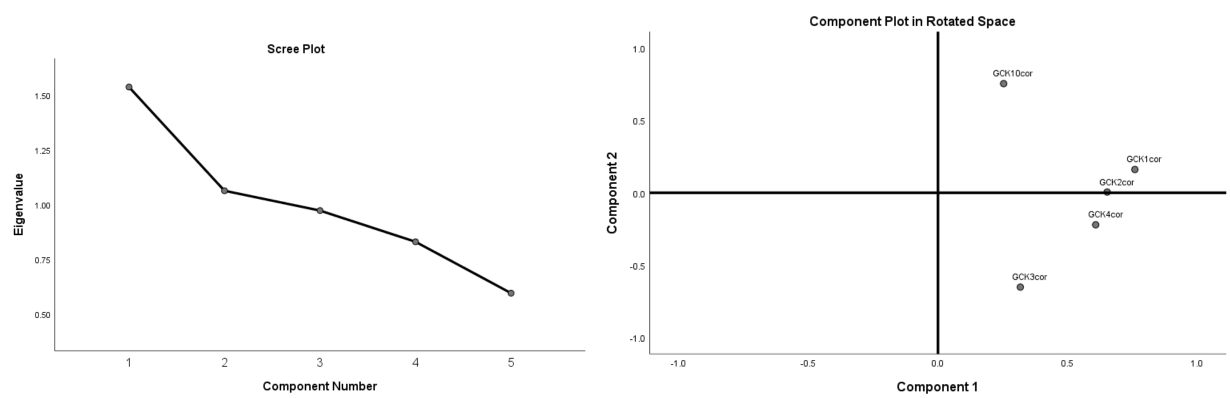
4. Correlations for All Measures

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. Condition (positive is risk tutor)	1.0																
Pretest Measures																	
2. Berlin Numeracy Test -S	-.09	1.0															
3. Subjective Numeracy	.00	.48**	1.0														
4. Graph Construction Knowledge	-.06	.43**	.17**	1.0													
5. Subjective Graph Literacy 1	-.06	.24**	.52**	-.01	1.0												
Posttest Objective Measures																	
6. Graph Literacy	.08	.51**	.36**	.33**	.14**	.1.0											
Visualizable Decision Tasks																	
7. Ratio Bias	.05	.40**	.23**	.25**	.21**	.29**	1.0	.									
8. Sunk Cost	.06	.06	-.05	.14**	.04	.22**	.23**	.1.0									
9. Framing	.12*	.21**	.06	.18**	.08	.36**	.24**	.10	1.0								
Financial Literacy																	
10. Basic Financial Literacy	-.02	.49**	.33**	.23**	.14*	.43**	.31**	.07	.26**	1.0							
11. Sophisticated Financial Literacy	-.17**	.23**	.19**	.18**	.09	.18**	.14**	.04	.11	.44**	1.0						
Confidence Measures																	
12. Subjective Graph Literacy 2	.01	.32**	.48**	.11	.59**	.27**	.22**	.00	.17**	.26**	.20**	1.0					
13. Ratio bias Overprecision	-.10	-.18**	-.05	--.13*	-.04	-.09	n.a	-.18	-.09	-.04	-.02	.01	1.0				
14. Ratio Bias Overestimation	-.09	-.19**	-.01	-.18**	.02	-.11	n.a	-.18	-.09	-.11	-.02	.01	.86**	1.0			
15. Sunk Cost Overprecision	-.13*	.12*	.06*	-.01	.12	.02	-.01	n.a.	.07	.17**	.14*	.11	.29**	.27**	1.0		
16. Sunk Cost Overestimation	-.06	.10	.12*	-.03	.18**	.01	.03	n.a	.08	.20**	.07	.22**	.26**	.35**	.84**	1.0	
17. Framing Overestimation	-.19**	-.07	.03	-.12	.05	-.27**	-.12	-.11	n.a	-.10	-.01	.01	.17**	.20**	.17**	.20**	1.0

Note: * = $p < .05$, * = $p < .01$; n.a. = measures were excluded as measures were used in the created measure (e.g., “ratio bias” was used to create the measure “ratio bias overprecision”).

5. Factor Analysis of Graph Construction Knowledge Scale

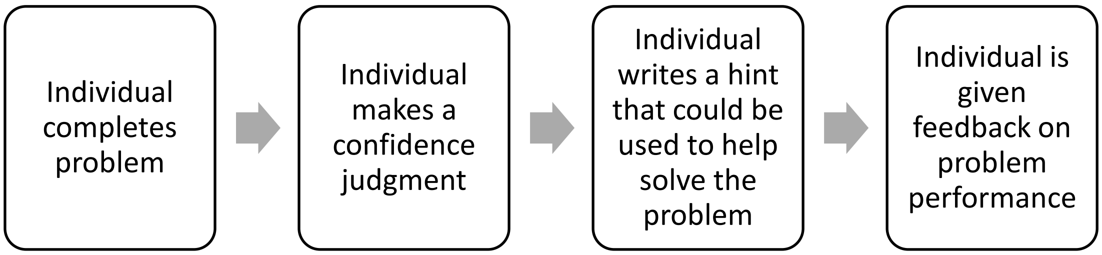
The 11 item graph knowledge scale was analyzed using a varimax rotated factor analysis. The full scale showed 59% variance explained over 5 components. After item reduction, a 5 item graph knowledge scale showed 51% variance explained over 2 components at a .39 Cronbach's Alpha (Questions include: 1, 2, 3, 4, and 10). The Graph Construction Knowledge scale needs refinement to be used for future endeavors if more domain specific graph knowledge (e.g., "pie graphs are most appropriate to communicate proportions") is needed to be assessed or to be used as parallel forms to the Graph Literacy Scale.



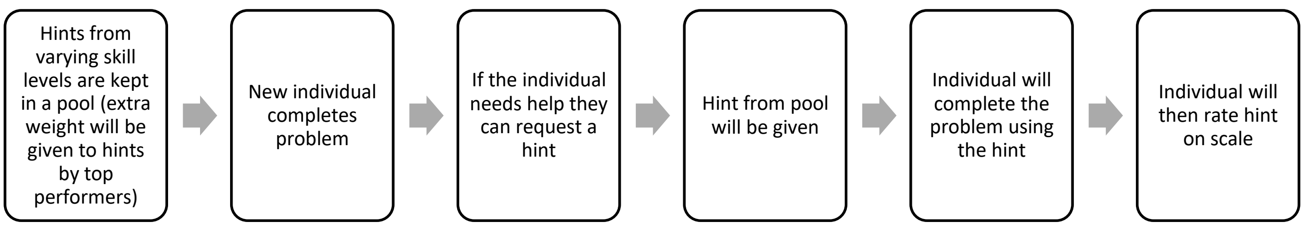
The figure above shows the finalized scree plots for the five item graph construction knowledge scale. The five items explain all but 8% of the total variance from the full eleven item scale.

6. The hint mining process in conjunction with online tutors

1. Hint Generation and Mining



2. Hint Rating and Use



3. Hint Categorization and Personalization