2015

DEVELOPMENT OF THE INTELLIGENT GRAPHS FOR EVERYDAY RISKY DECISIONS TUTOR

Margo Woller-Carter
Michigan Technological University, mwoller@mtu.edu

DEVELOPMENT OF THE INTELLIGENT GRAPHS FOR EVERYDAY RISKY DECISIONS TUTOR

By

Margo M. Woller-Carter

A DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

In Applied Cognitive Science and Human Factors

MICHIGAN TECHNOLOGICAL UNIVERSITY

2015

This dissertation has been approved in partial fulfillment of the requirements for the Degree of DOCTOR OF PHILOSOPHY in Applied Cognitive Science and Human Factors.

Department of Cognitive and Learning Sciences

Dissertation Co-Advisor: *Edward T. Cokely*

Dissertation Co-Advisor: *Rocio Garcia-Retamero*

Committee Member: *Scott Kuhl*

Committee Member: *Robert Pastel*

Committee Member: *Kelly Steelman*

Department Chair: *Susan Amato-Henderson*

To my mom

Table of Contents

Abstract

Simple graphical visual aids have now been shown to be among the most effective means of quickly improving people's ability to evaluate and understand risks (i.e., risk literacy), particularly for diverse and vulnerable groups (e.g., older adults, less educated, less numerate, minority and immigrant samples). Although well-developed theory and standards for user-friendly graph design exist, guidelines are often violated by designers faced with constraints like conflicts of interest (e.g., persuasion and marketing vs. informed decision making). Even when information is presented in well-designed graphs, many people struggle with appropriate data interpretation. Can basic computerized graph literacy training improve essential graph and risk evaluation skills? To begin to answer this question, I conducted three studies that developed and validated psychometric tests of three component graph literacy skills, namely (1) graph type knowledge, (2) selecting appropriate graphs, and (3) knowledge of graph distortions. I then developed a computerized graph literacy training platform and conducted a mixed-factorial experiment investigating a wide-range of training effects. Results indicate that even in a sample of tech savvy college students one hour of basic computerized training can dramatically improve graph literacy (*Cohen's d* = 1.10). Results also provide some of the first evidence that graph literacy training can improve general decision making skills that involve spatial or visualization-relevant processing, such as resistance to sunk costs, framing effects, and class-inclusion illusions. Discussion focuses on practical and theoretical implications, including usability modeling that should inform continuing development of the RiskLiteracy.org Decision Making Skills Training Program.

Chapter 1: Introduction

In our modern and complex world, informed decision-making often requires the ability to evaluate and understand risk—i.e., risk literacy (Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012; Gigerenzer, 2012). Unfortunately, recent estimates suggest that nearly 35% of Americans are unable to understand the information about risk they encounter on a regular basis (Galesic & Garcia-Retamero, 2011). Graphs are simple yet powerful technologies that are widely utilized for risk communication and promotion of informed decision making around the world (Galesic, Garcia-Retamero, & Gigerenzer, 2009; Garcia-Retamero & Cokely, 2013, 2014a, 2014b; Garcia-Retamero & Galesic, 2013; Larkin & Simon, 1987; Okan, Garcia-Retamero, Cokely, & Maldonado, 2015; Pinker, 1990; Spiegelhalter, Pearson, & Short, 2011). But even if "A picture is worth a thousand words" (Brisbane, 1911), well-designed graphs only tend to improve decision making specifically when they clarify complex data structures by depicting relations in bars, lines, pies, icon arrays, and decision trees (Larkin & Simon, 1987; Mt-Isa et al., 2013; Pinker, 1990; Spiegelhalter et al., 2011). After all, not all graphs are created equally. Failures to follow well-established graph design standards can complicate and bias informed decision-making (Larkin & Simon, 1987; Okan, Woller-Carter, Garcia-Retamero, & Cokely, 2013; Woller-Carter, Okan, Cokely, & Garcia-Retamero, 2012). People also differ in their ability to understand graphs (i.e., graph literacy) meaning that some people still struggle even when presented with validated, well-designed graphs (Garcia-Retamero & Galesic, 2010; Garcia-Retamero, Okan, & Cokely, 2012).

The manifold major social, economic, and health consequences of lower levels of risk literacy and graph literacy are well documented (Cokely et al., 2012; Garcia-Retamero & Cokely, 2012, 2015a, 2015b, Submitted; Garcia-Retamero, Cokely, & Galesic, 2013; Garcia-Retamero & Galesic, 2013; Ghazal, Cokely, & Garcia-Retamero, 2014; Peters, 2012; Reyna, Nelson, Han, & Dieckmann, 2009). For example, people who are less graph literate are much more likely to choose to pay higher prices for less effective products (Woller-Carter et al., 2012), they are much more likely to choose less effective treatment options and to fail to comply with their treatment regiments (Okan et al., 2013; Woller-Carter et al., 2012), and they also much more likely to recommend ineffective or even potentially deadly public policies (Nelson, Hesse, & Croyle, 2009; Okan et al., 2013). While not all graphs depict information related to decision making and risk, essential graph literacy sub-skills like reading points on a graph, comparing between points, and predicting trends with data presented in graphs (Galesic & Garcia-Retamero, 2011) are common tasks that make graph literacy a topic of great theoretical and practical interest to the risk communication and behavioral health and finance communities (Garcia-Retamero & Cokely, Submitted; Garcia-Retamero & Galesic, 2013; Lipkus & Hollands, 1999; McCarley et al., 2015). How can we help people more efficiently learn to evaluate and understand graphs and risks?

Graph Comprehension

Graphs are a relatively new type of technology used for information representation. Historical scholarship traces the earliest precursors of modern graphs to systems used during the 10th century for depicting planets' paths over time and to the

emergence of longitude differences for depicting differences between two cities that became more common in the 17th century. However, it wasn't until the 18th century when the first modern abstract graphs were created and used for depicting and expressing data (Friendly, 2008; Tufte, 2001). Modern scholarship on the psychological mechanisms is extensive (see Shah and Hoeffner, 2002) although much of the essential context is reflected by two of the most influential theories of graph comprehension and associated research. Consider for example the highly influential theory of graph comprehension by Pinker (1990). Pinker's (1990) theory describes the basic visual information processes required to encode graphs and the cognitive processes required to convert the basic visual information into a meaningful description of (1) the graph and (2) information relations among within the graph.

Carpenter and Shah (1998) built on Pinker's (1990) theory of graph comprehension, comparing two theoretical models of graph comprehension. The first model, *the pattern-recognition model*, accords with the processes proposed by Pinker (1990), including (1) encoding the visual pattern, (2) translating the visual pattern into the conceptual/quantitative relationship, and (3) identifying the referents from the conceptual/quantitative relationships. The pattern-recognition model predicts that most of the cognitive processing takes place in step 1 during encoding of the visual pattern, which is then followed by the less cognitively taxing processes in steps 2 and 3. The second model, *the integrative model*, assumes the same three processes as the pattern-recognition model, but predicts an iterative cycle instead of ordered steps. Rather than processing the whole display at once in the encoding step, visual "chunks" are recognized

and encoded. The "chunks" are then interpreted and added to the current cognitive representation of the display. The visual "chunks" become more complex with each cycle until a full, precise and detailed cognitive representation of the display is stored in memory.

In a competitive analysis of the two graph comprehension models, Carpenter and Shah (1998) conducted an experiment using data from eye-tracking and verbal protocol analysis. Participants were asked to describe line graphs that varied in complexity, in the form of additional lines in the graph. The pattern-recognition model's *a priori* account predicted a small amount of time and gazes on the axes and labels, with the majority of the time and gaze focused on the pattern of the graph. Additionally, the model predicted that increased complexity would only affect the amount of time and gaze on the pattern but not the axes or labels, with few notable exceptions (e.g., the additional labels requiring minimal additional time and gazes to be encoded). In contrast, the integrative model predicted that a greater portion of time and gaze should occur on the axes and labels as compared to that on the pattern. The integrative model further predicted that as complexity increased so too should the time and gazes spent on the pattern, axes and labels. Analyses revealed that data largely supported the integrative model, which better predicted the amount of both time and gazes spent on the axes and labels compared to the pattern, and the increase in time and gazes with increased complexity.

A second experiment was also by conducted by Carpenter and Shah (1998) to further investigate the interpretation process at work in graph comprehension. The integrative model predicted a monotonic relationship between gazes and the number of

distinct functions presented in the graph. Instead of asking participant to describe the graph, the researchers read the title of the graph and asked participants a question about a relationship in the graph. The time and gazes by area of interest were again consistent with the integrative model. A more precise mathematical model of the gaze pattern was derived from the data. The mathematical model first scans each of the areas of interest once to get an overview of the graph. Then a single function is identified in the pattern, which is followed by one of three, 1) describe the change in the x-axis, including direction, scale, units and/or referent, 2) describe the change in the y-axis, including direction, scale, units and/or referent, and 3) describe the change in the z-labels, including direction, scale, units and/or referent.

Results from these and other validation studies suggest that the integrative model of graph comprehension can provide a useful framework for the examination of individual differences in graph skills. Because the integrative model focuses on specific cognitive operations and their relations to specific aspects of the visual stimuli, it provides reasonable starting points for examining potential individual differences in skills related to (1) production deficiencies—i.e., failures to use appropriate strategies as compared to (2) utilization deficiencies—i.e., failures to benefit from typically effective strategies. For example, a production deficiency could be failing to strategically evaluate and consider the units of the axes when reading a graph (e.g., not encoding that an axis in is depicted in a logarithmic scale). In contrast, utilization deficiencies would result when participants did strategically consider the axes yet still did not use the information therein for some reason (e.g., didn't understand how to interpret the logarithmic scale). In these

ways and others, the cognitive approach to investigation of graph comprehension can provide insights into the sources, causes, and typical processes that give rise to individual differences in graph interpretation and comprehension.

Measuring Graph Literacy.

Within the cognitive and behavioral sciences, the most common way to measure graph literacy is with a psychometric performance instrument, namely the graph literacy scale (Galesic & Garcia-Retamero, 2011). The graph literacy scale is commonly used to measure graph skills assessed across three sub-skills, including reading points on a graph, comparing between points, and predicting trends with data presented in graphs (Galesic & Garcia-Retamero, 2011). The development of the graph literacy scale took place via an iterated test development and validation process. In a first study, researchers began with 60 German students and 60 German older adults who completed 42 items requiring the use of data presented in a graph, with additional individual difference measures used to assess convergent and discriminant validity (i.e., the degree to which the scale correlated with other similar scores and the degree to which the score dissociated from theoretically unrelated skills). The 42 items were all presented in the medical domain, and covered the four most widely used graph types, i.e. line, bar, pie, and icon arrays, as well as a range of complexity.

The 42-item measure required an average of 21 minutes to complete with a mean score of 34 out of 42 correct, and good internal consistency and reliability metrics (i.e., a Cronbach's α of .85). Psychometric optimization procedures based on these results identified 13 items that were selected for the final version of the measure based on five

criteria, namely: 1) the item had a percent correct less than 90; 2) the item-total correlation was at least 0.3 discriminability; 3) correlation with the existing graph literacy measure was at least 0.3; 4) the item covered the three graph abilities (read, compare, and predict trends) and covered the four commonly used graph types (line, bar, pie, and icon arrays); and 5) the measure had to take less than 10 minutes to complete.

The next phase of test development involved assessments on probabilistic national samples of the United States (492 participants) and Germany (495 participants). The 13-item measure was completed along with additional individual difference measures, including numeracy, and two graph performance tasks with two conditions (i.e., no graphs and graphs by the sample). The 13-item measure took between 9 and 10 minutes to complete with an average of 10.1 minutes for the United States sample and 9.2 minutes for the German sample. A reliability analysis provided additional evidence of internal consistency (i.e., α =.79 for the United States sample and α =.74 for the German sample).

Participant performance on the graph tasks were analyzed by creating four ability groups; high numeracy – high graph literacy, high numeracy – low graph literacy, low numeracy – high graph literacy, and low numeracy – low graph literacy. Graphs were found to be most helpful for participants in the low numeracy – high graph literacy group as compared to the low numeracy – low graph literacy group. This trend was also present in the high numeracy groups and has since been observed dozens of times across a wide variety of samples (e.g., surgeons, immigrants, at-risk young adults, etc.).

Benefits of Understanding Graphs.

A recent review of the use of visual aids to communicate health risks (Garcia-Retamero & Cokely, 2013) documents profound social, health, and economic benefits of visual aids that are mediated by improved comprehension of risks, in diverse samples including patients and doctors. While most people benefit in one form or another (e.g., increased confidence, improved user experience), the greatest benefits in terms of risk comprehension are often seen among people with low levels of practical mathematical skills—i.e., numeracy—particularly when those people also have moderate-to-high levels of graph literacy (Ellis, Cokely, Ghazal, & Garcia-Retamero, 2014; Garcia-Retamero et al., 2013; Garcia-Retamero & Galesic, 2010; Garcia-Retamero, Wicki, Cokely, & Hanson, 2014; Okan, Garcia-Retamero, Cokely, & Maldonado, 2012; Petrova, Garcia-Retamero, & Cokely, 2015). Unfortunately, even the most graph literate people can still be tricked by poorly designed graphs and graphs that are intentionally designed to distort and manipulate understanding and decision making (Okan et al., 2013; Woller-Carter et al., 2012). Note, however, that individuals who lack a basic understanding of graph literacy are not aided as much by graphs as those with a least a modest understanding of graphs (Garcia-Retamero & Galesic, 2010; Okan et al., 2012). And although guidelines exist to design graphs (Gillan, Wickens, Hollands, & Carswell, 1998; Jarvenpaa & Dickson, 1988; Kosslyn, 2006; Toth, 2006) currently there is no means to enforce the use of graph design guidelines and in fact guidelines are often violated by graph designers (Cooper, Schriger, Wallace, Mikulich, & Wilkes, 2003).

Improving Graph Comprehension

If the burden of designing easy to understand graphs cannot be practically met by graph designers, then one potential means of inoculating users against biased graphs is to increase the graph users' ability to understand graphs—i.e., improve their graph literacy. Currently, there are only a few easily accessible options for adults looking to improve their graph comprehension skills, as most online materials focus on graph skills in children. Short of reading Kosslyn's Graph Design for the Eye and Mind (2006) or Huff's How to Lie with Statistics (1952), my extensive review of the literature and a deliberative multi-year search for materials (including a visiting fellowship with one of the leading cognitive tutoring groups) indicates that the best currently available tutors are: 1) Carnegie Learning's MATHia (2011) , 2) MindTools.com's Charts and Graphs (2007) reading, and 3) SmartGraphs' Graph Literacy course (2011). While ambitious and generally well executed, these training systems vary in difficulty, content, mode of training, and scope of skills covered, and there is good reason to think they are not ideal for most adult learners who have limited time and resources.

Consider the options offered by the Carnegie Learning group that includes graph skills in a few specific modules of their program MATHia (2011). MATHia is an intelligent tutor designed to be used in concert with class lectures for middle and high school students. The MATHia modules dealing with graphs cover creating bar graphs and histograms, as well as, reading and comparing points in graphs. The training is text and task based. The text gives students a basic understanding of the material, which is then applied in the tasks. The tasks are adaptive in nature, such that objective and subject

difficulty levels of one's training are tracked along with skill mastery. Specifically, once a skill is mastered, the student no longer completes the task focusing on the skill. If a student struggles with the task, easier or even sub-tasks processes are then presented, along with appropriate feedback about success and failure. Some versions also include metacognitive scaffolding (e.g., helping students think about thinking during learning).

While the graph skills that are covered in MATHia are efficiently trained, there are a few downsides to MATHia that likely make it less ideal for adult learns and those interested in risk literacy applications. For example, MATHia was designed for middle and high school students making most of the task content irrelevant or uninteresting for many adults. Additionally, MATHia only covers a few graph skills, and only a handful of graph types. MATHia is also a "for purchase" training. For full use of MATHia by a school, access to MATHia, textbooks, and course syllabi must be purchased through the Carnegie Learning group. Researchers may access the content for program for research purposes by requesting a free trial and login; however, once access to MATHia is granted, finding the modules with the graph content requires a surprising amount of time, as graph skills are taught as a part of other courses and are rarely the focus of a course. Further complicating one's search for appropriate modules, the graph modules tend to be filed under statistics and data analysis rather than data visualization. Generally, this system may give the impression of graph literacy related content that is more academic than everyday users require or desire.

MindTools.com also offer a training resource, namely one reading titled Charts and Graphs (Hallett & MindTools.com, 2007). The course covers choosing the correct

graph type based on the data and goals of the user, along with the basics of x- and y-axes. The graphs covered in the training include line graphs, bar graphs, pie charts, and Venn diagrams. The content is purely text based, however, and does not include tasks to test skills or to provide interactive feedback. Nevertheless, the graph selection recommendations do follow most of the HFES guidelines (Gillan et al., 1998) and cover all but two of the common graph types.

The third option, Smart Graphs, has five courses available online. One course specifically focuses on Graph Literacy (Staudt et al., 2011). The Graph Literacy course is comprised of six modules: 1) Equivalent Graphs, 2) Interpolation, 3) Independent and Dependent Variables, 4) Graphs Tell a Story, 5) Hurricane Katrina, and 6) Growing Up. The first three models cover basic graph skills, while the last three give applied examples for using graphs. The content is not adaptive, but it is text and task based, allowing knowledge to be tested while direct connections to everyday applications are made clear. However, the content was designed for children making the tasks and applications content less relevant and likely less motivating for adults who are interested in training decision-making skills and risk literacy.

Taken together the currently available online tools mainly focus on children's graph skills, with the exception of MindTools.com reading. While the MindTools.com reading accords with the HFES graph guidelines (i.e., how to select the correct graph for your data), there are no tasks that allow self-testing for the content has to been learned. Moreover, it is unclear whether the graph literacy knowledge and learning will transfer to content not included in the text. I suspect that after completing any of the currently

available tools, most users would still not have a broad and representative coverage of the skills needed to improve graph literacy in such a way that it would also support risk literacy (e.g., unlikely to transfer to other tasks in support of superior decision making and risk evaluation). Completing the content of all three tools also requires four to six hours, depending on the users' previous knowledge and graph skills. In summary, there is currently no online tool available that covers all the graph skills needed to improve graph comprehension quickly, adaptively, and in a way that would likely be satisfying to diverse adults.

Intelligent Tutors.

Theoretically, intelligent tutors are more efficient learning support systems (e.g., instructional systems) because they embody key lessons of cognitive and learning sciences in interactive computer programs that adapt to the needs and capabilities of users in real time. VanLehn (2006), one of the leading authorities in the field describes a general framework and common language of intelligent tutoring systems (e.g., terms that are used widely in the intelligent tutor community and act as a common language between developers). The ***task domain*** refers to the knowledge and skills being taught. The ***tasks*** in a tutor are multi-step activities that can be rearranged depending on the tutor and student needs. A ***step*** is a user interface interaction executed to complete the task. The facts, rules, and principles of the domain presented by the tutor are ***knowledge components***. When the student applies a knowledge component to a task a ***learning event*** has occurred. Any action by the student that is inconstant with the instruction is labeled as incorrect. The two key components of an intelligent tutor are an ***outer loop***, which

controls task selection and presentation, and an ***inner loop***, which gives feedback and hints, while tracking skill acquisition. Systems without an inner loop are not considered intelligent tutors.

There are four methods for selecting tasks for the students of varying complexities. The least complex outer loop structure presents a list of tasks to the student, and the student then selects the tasks to complete. This is a common structure for online homework tasked assigned from instructors in different sections of the same class. Increasing the complexity of the outer loop can lead to training with a set order for all students. For example, the training might have some text to read followed by a video and comprehension quiz, before the student completes five guided tasks. The guided tasks are then followed by five unguided tasks before the completion of a unit test. The order of the material and the tasks are the same for all students and require similar amounts of time for all students. The tutor can also use a mastery learning outer loop structure, which requires that the student master the knowledge components of the unit prior to moving on to a new unit. Thus, the knowledge components must be labeled and traced by the tutor and a mastery level needs to be set by the developer and achieved by the student to move forward. The knowledge components are then monitored in the inner loop and feed to the outer loop. The most complex outer loop structure is referred to as macro-adaptation. To function, the tutor must know the knowledge components for each task and keep a running estimate of the current state of the knowledge components for all the tasks. Tasks are selected for presentation based on the amount of overlap between the mastered knowledge components and the knowledge components of the uncompleted tasks.

In order for marcoadaptation to work correctly between tasks and sessions, information about the student must be stored on a server. This information is commonly referred to as a student model, and often contains information in attribute-value pairs. The information in the student model can be as simple as a set task list or include information like GPA, standardized test scores, major, learning style in addition to task completion performance, number of hints requested, time to complete tasks, and number of failed attempts. This information can then be used to suggest targeted tasks for the student to improve specific skills or direct the learning modality of lessons.

The outer loop can also control the mode for the tasks. Some systems have a guided mode so that each step is demonstrated and explained to the student along with hints of next step. Essentially, hints are given to guide the student to the next step that is needed in order to complete the task. In contrast, the student led mode gives hints only when the student requests them. In most cases, the mode is selected based on the knowledge component mastery and number of tasks completed.

The inner loop tracks knowledge components at the level of the steps completed by the student and the knowledge components associated with each step. The inner loop also controls the timing, type, and amount of feedback the student is given, while completing the steps of the task. Feedback can vary from immediate feedback for each step to feedback only after submitting a task. The timing, amount, and type of hints given to the student are also controlled by the inner loop. A common practice in tutoring systems is to give hints to lead the student to the correct next step through multiple levels of hints ending in a bottom-out hint (i.e., telling the student what to enter and where). The

tracking of knowledge components, feedback, and hints are the essential elements that set intelligent tutoring systems apart from other tutoring systems (VanLehn, 2006). The design of user experience elements, content (or learning goals), and style features are primarily what differentiates intelligent tutoring systems from one another.

Current Research

This project is part of the NSF funded RiskLiteracy.org Decision Making Skills Training Program (SES-1253263). The current project was completed in two phases. The goal of Phase 1 was to develop and validate new psychometrically optimized individual difference assessment technologies (i.e., simple tests) for three key categories of component graph literacy skills. Assessment of these skills is required for higher fidelity modeling and measurement of skill levels, which in turn allows the tutor to determine task difficulty and student needs. Phase 2 was an experiment designed to test the effectiveness of the new tutor compared to an existing tutor, and to examine and map theoretically interesting questions about the benefits of different learning systems (e.g., does graph literacy training help people make better risky decisions more generally) and the benefits of more friendly user experiences (e.g., to what extent does efficiency promote learning success across different levels of skill).

Chapter 2: Methods and Results

Phase 1 includes 3 independent studies. The goal of the first study was to develop a short individual difference measure of one's ability to select or identify the most appropriate type of graph for presenting various data and information—i.e., graph type selection skill (SelectionGL). The study included existing individual difference for convergent and discriminate validity (see Study 1 Materials section for more details). The second study developed a short individual difference measure tentatively called "lying with graphs literacy" (LyingGL for short). The study included the same existing measures as Study 1 and the SelectionGL scale developed in Study 1. The final study used the existing measures from the previous studies, as well as the SelectionGL and LyingGL respectively in order to develop psychometric profiles of various graph and data tasks (i.e., building quantitative models of how hard and how unique different specific problems or test items are). In addition, a short measure of graph design skills (DesignGL) was developed.

The aim of phase one was to develop three additional individual differences measures specific to graph design and model task difficulty via psychometric indices of difficulty by discriminability. The new individual difference measures and the task difficulty ratings informed the design of the iGERD Tutor, which was tested in phase two of the research.

Phase 1: Individual Differences and Task Difficulty

The purpose of the Phase 1 studies was to create individual difference measures of key skills needed to design and comprehend graphs, SelectionGL and LyingGL. Additionally, these measures, and measures of graph literacy and numeracy, were then used to determine task difficulty. All of the studies in Phase 1 were conducted using Unipark surveys completed by diverse paid web panel participants recruited via Amazon's Mechanical Turk service. All participants in Phase 1 were paid for their participation based on a flat rate fee yoked to the average required to complete the surveys (e.g., about $1.00).

Study 1: SelectionGL.

SelectionGL is required for the iGERD tutor. The assessment of SelectionGL also informed quantitative structural models of more general construct of graph literacy (e.g., data reduction model like factor analysis or multivariable modeling in a general linear regression framework).

Participants. Data were collected from 257 participants. A final sample of 217 participants was used for analysis after 40 participants (15.6%) with incomplete data were removed. The mean age was 38 with a range of 18 to 85. The sample was comprised of 60.8% females, 38.2% males, and 0.9% of participants who preferred not to indicate their sex. Most (88%) participants had some college education or better, with only 12% of participants reporting less education; four participants with some high school and 22 had a high school diploma or equivalent. Most (65%) of the participants

were currently employed with 54% being in less than a management position. The mean annual household income range was between $35,000 and $49,999.

***Materials*.** Four existing individual difference measures were included for convergent and discriminate validity. (1) The subjective numeracy scale, developed by Fagerlin and colleges (2007), is an 8-item scale that asks participants to rate their skills working with numeric information and their preferences for risk information in words or a numeric format. (2) The subjective graph literacy scale, developed by Garcia-Retamero, Cokely, Ghazal, and Hanson (2014), is a 5-item scale that asks participants to rate their skills working with different graph types, as well as, reading and comparing points in a graph. (3) The Berlin Numeracy Test is an adaptive, objective measure of numeracy, one's ability to understand and use statistical information (Cokely et al., 2012; Cokely, Ghazal, Galesic, Garcia-Retamero, & Schulz, 2013; Cokely, Ghazal, & Garcia-Retamero, 2014), requiring participants to complete either 2 or 3 items to assess their numeracy. (4) The Graph Literacy Scale, developed by Galesic and Garcia-Retamero (2011), contains 13 items that require participants to read points in a graph, compare points in a graph, and project data in a graph in to the future.

The two subjective measures do not require participants to use their skills to complete tasks. In contrast, the objective measures of numeracy and graph literacy are measures of performance on tasks. The subjective measures are moderately correlated with their objective counterpart. Participants completed the subjective numeracy scale (Fagerlin et al., 2007), subjective graph literacy scale (Garcia-Retamero, Cokely, et al., 2014), Berlin Numeracy Test (Cokely et al., 2012), Graph Literacy scale (Galesic &

Garcia-Retamero, 2011) with 4 additional difficult questions, and 20 SelectionGL tasks. The SelectionGL tasks were presented in a random order and the answer options were randomized between questions. Demographic information was also collected. All new items are presented in Supplemental File A.

Procedure. Participants accessed the survey that was programmed in Unipark. The online program instructed them to read the informed consent form and to agree to participation prior to completing the survey (note all studies were approved by MTUs IRB—M0650). The participants completed the measure in the order specified above. After completing all measures, the participants read a debriefing statement and then were given the code required to receive payment for their participation.

Data analysis. All items were scored according to the procedures of the articles in which the scales were developed. The SelectionGL items were scored for correctness and a total score was calculated as the total number of items correct. Bivariate correlations were conducted to investigate the relations between the existing cognitive ability measures and the SelectionGL scale. To investigate the relations found with the bivariate correlations in more detail, linear hierarchical multiple regressions were conducted based on a priori theoretical assumptions, with SelectionGL score as the dependent variable. Step 1 of the regression entered numeracy and graph literacy as predictors, and Step 2 added subjective numeracy and subjective graph literacy, yielding a significant predictive model of SelectionGL in terms of numeracy, general graph literacy, and subjective graph literacy.

In order to represent the difficulty of the SelectionGL, the frequency correct was calculated for each item. After determining the difficulty of the items, bivariate correlations between each item and the total SelectionGL score were used to determine the discrimination of each item (i.e., modified classical test theory psychometric item analysis). Items with correlation coefficients greater than or equal to .300 were assumed to be above the cut score for minimum discriminability between skill levels. Cronbach's α was also calculated for the full scale. Analyses resulted in a four unique short form solutions (i.e., potential tests) for the brief SelectionGL. All short forms offered roughly interval differences in difficulty level across the full range of difficulty (e.g., the difference in difficulty from item 1 to 2 was about the same change in overall difficulty from item 3 to 4, and so on). Correlations to the existing cognitive ability scales and the four forms were compared to the full scale. Regressions were also used to determine which of the short forms best recovered or predicted performance across all items. Cronbach's α was also calculated for each of the short forms.

Results. The descriptive statistics and maximum score for each measure are displayed in Table 2.1. Higher values mean higher ability/preference on all scales. The scores on the graph literacy scale are higher than in past research due in part to the addition of four extra questions. Bivariate correlations were used to examine relations that might exist between scores on various cognitive ability measures. All measures showed the expected positive manifold associated with domain general cognitive abilities, with the exception of SelectionGL and subjective graph literacy (see Table 2.2 for correlation coefficients). Additionally, the correlation patterns of the existing

measures were consistent with those observed in past research (Garcia-Retamero, Cokely, et al., 2014; Woller-Carter et al., 2012).

Table 2.1

Descriptive Statistics for Cognitive Ability Measures' Scores in Study 1

Measure	Mean (SD)	Median	Maximum
Numeracy	2.21 (1.05)	2.00	4
Graph Literacy	13.24 (2.72)	14.00	17
Subjective Numeracy	4.36 (0.90)	4.5	6
Cognitive Abilities	4.12 (1.21)	4.25	6
Preference	4.61 (0.97)	4.75	6
Subjective Graph Literacy	4.23 (1.08)	4.2	6
SelectionGL	9.29 (3.21)	9.00	20

Note. SD = Standard Deviation

Table 2.2

Bivariate Correlation Coefficients between Measures in Study 1

	Numeracy	Graph Literacy	Subjective Numeracy	Subjective Graph Literacy
SelectionGL	.29**	.44**	.15*	.10^{+}
Numeracy		.38**	.16*	.22**
Graph Literacy			.25**	.20**
Subjective Numeracy				.62**

* $p < .05$. ** $p < .01$. $^{+}$ $p = .15$

A Cronbach's α of .59 was derived for the SelectionGL scale, a remarkably high coefficient given the number of graph types included and short length of the test. Multiple regression modeling revealed that Numeracy (β = .15, p = .03) and graph literacy (β = .39, p < .001) were each unique and robust predictors of SelectionGL performance (R^2 = .22, p < .001). Subjective numeracy (β = .06, p = .45) and subjective graph literacy (β = -.05, p = .53) were not unique predictors of SelectionGL in the full model, after controlling for numeracy and general graph literacy. Results suggest that SelectionGL is reasonably well represented as primarily reflecting skills that are linked to essential processes in general graph literacy, along with some unique contributions from numeracy that theoretically may reflect differences in underlying metacognitive and general problem solving skills (e.g., Cokely et al., 2012; Ghazal et al., 2014).

Item analysis of the SelectionGL items focused on the difficulty and discriminability of each item. The item difficulty was determined by the percent of participants correctly answering the item; while discriminability was determined using bivariate correlations between the score on each item with the total score (item-total correlations; see Table 2.3 for the item analysis statistics). Item 8 was the only item on which participants performed below chance (20%). The majority, about 37%, thought a bar graph should be used instead of an icon array. However, to make a bar graph, additional operations would be required to convert the numbers given into percentages, which in turn has implications for the reference class/denominator (i.e., obscuring it) that are required to make informed decisions from the data. Four items were unrelated to the

total score (4, 7, 10, & 17). Three of the items, which failed to correlate were bar graph items (4, 7, & 10), and the forth item was a line graph item (17).

The performance of four different short forms of the SelectionGL scale were compared to the full SelectionGL scale (see Table 2.4). Short form A had the second highest Cronbach's α (.43), but did not have a bar graph item. Short form B included a bar graph item but removed an item with greater discrimination resulting in the lowest R^2 of the short forms. Short form C included six items (3, 6, 8, 12, 16, and 19) making it the longest of the short forms. Short form D excluded the hardest item (8) resulting in a limited range of difficulty. Short form C was selected based on the overall superiority of its psychometric for the final short form (e.g., it had the best Cronbach's α and R^2 of the short forms covering the full range of difficulty, without any sacrifice to overall predictive or convergent validity).

Table 2.3

Percent of Correct Responses and the Item-Total Bivariate Correlation Coefficients for the SelectionGL Scale

Item	Graph	Correct	Total	Item	Graph	Correct	Total
1	IA	24.9	**.33****	11	P	50.7	**.37****
2	IA	35	**.44****	*12[a,b,c,d]*	*P*	*60.8*	***.48****
3[a,c,d]	*IA*	*30.9*	***.42****	13	P	66.8	**.49****
4	B	29	-0.07	14	P	48.8	.27**
5	L	61.8	**.39****	15	L	71	**.45****
6[b,c,d]	*B*	*54.4*	***.31****	*16[a,b,c,d]*	*L*	*72.4*	***.42****
7	B	49.3	0.06	17	L	31.3	0.08
8[a,b,c]	*IA*	*15.7*	*.20***	18	DT	45.6	**.44****
9	DT	52.1	**.48****	*19[a,b,c,d]*	*DT*	*43.8*	***.49****
10	B	36.9	-0.02	20	DT	48.4	**.57****

Note. IA = Icon Array. B = Bar Graph. L = Line Graph. DT = Decision Tree. P = Pie chart/graph. Item-total correlation coefficients indicated in bold are discriminating items. Superscript letters indicate which of the short forms the item was included. Italic items are included on the final short form.
** $p < .01$

Table 2.4

Comparison of SelectionGL Short Forms verses the Full Scale

Form	α	R^2	Numeracy	Graph Literacy	Subjective Numeracy	Subjective Graph Literacy
Full	.59	.22**	.15	.39**	.06	-.05
A	.43	.12**	.03	.33**	.07	-.04
B	.38	.11**	.07	.30**	.10	.02
C	**.42**	**.12****	**.05**	**.32****	**.09**	**.002**
D	.44	.13**	.08	.32**	.10	.01

Note. α = Cronbach's α. Standardized β's are presented for each of the cognitive abilities scales. Short form C, indicated in bold, was selected as the final short form.
** $p < .01$

Study 2: LyingGL.

A measure of LyingGL is required for the IGERD tutor. The measure will be used to determine the users' entry into the LyingGL module and task difficulty within the module.

Participants. Data were collected from 376 participants. A final sample of 299 participants was used for analysis after 20.5% of participants with incomplete data were removed. The mean age was 37 with a range of 19 to 75. The sample was comprised of 58.5% females, 39.8% males, and 0.7% of participants preferred not to indicate their sex. Most (89.3%) participants had some college education or better, with only 10.7% of participants reporting less education; one participant with no schooling completed, two participants with some high school, and 30 with a high school diploma or equivalent. Most (79.3%) of the participants were currently employed with 53.2% being in less than a management position. The mean annual household income range was between $35,000 and $49,999.

Materials. The materials from study 1 were again used with the following modification. SelectionGL short form C was used. Following the SelectionGL tasks, the LyingGL tasks were added. The LyingGL tasks are presented in Supplemental File B.

Procedure. The same procedure was used as in Study 1.

Data analysis. The same data analyses were conducted on the data from Study 2 as the data from Study 1 with the following exceptions. The LyingGL items were scored for correctness and a total score was calculated. LyingGL was the dependent and

SelectionGL was entered in Step 1 of the regressions. Only two short forms of the LyingGL scale were compared. Cronbach's α was also calculated for the SelectionGL scale to test the scale's reliability.

Results. The descriptive statistics and maximum score for each measure are displayed in Table 2.5. Scores on the pre-existing cognitive ability measures were similar to Study 1. The SelectionGL short form measure had a mean of 2.77, a standard deviation of 1.41, and a median of 3.00 in Study 1, which is also similar to the current scores.

Bivariate correlations were used to determine if relationships exist between scores on the cognitive ability measure. All measures are correlated with each other (see Table 2.6 for correlation coefficients). Additionally, the correlation patterns of the existing measures are consistent with Study 1; however, the correlations are stronger than Study 1.

A Cronbach's α of .52 was calculated for the LyingGL. Numeracy (β = .15, p = .008), graph literacy (β = .23, $p < .001$), and SelectionGL (β = .20, $p < .001$) predicted LyingGL performance (R^2 = .19, $p < .001$). Subjective numeracy (β = .11, p = .11) and subjective graph literacy (β = .00, p = .995) were not statistically significant predictors of LyingGL performance.

Table 2.5

Descriptive Statistics for Cognitive Ability Measures' Scores in Study 2

Measure	Mean (SD)	Median	Maximum
Numeracy	2.16 (1.10)	2.00	4
Graph Literacy	13.30 (2.68)	14.00	17
Subjective Numeracy	4.41 (0.91)	4.5	6
Cognitive Abilities	4.25 (1.15)	4.25	6
Preference	4.59 (1.00)	4.75	6
Subjective Graph Literacy	4.25 (1.09)	4.4	6
SelectionGL	2.60 (1.37)	3.00	6
LyingGL	5.30 (2.70)	5.00	19

Note. SD = Standard Deviation

Item analysis of the LyingGL items used the same process as Study 1 (see Table 2.7 for the item analysis statistics). Many of the items showed near chance performance and were removed from further consideration (see Table 2.7). Items with performance close to chance also failed to provide adequate psychometric sensitivity and were excluded from the short form models.

Table 2.6

Bivariate Correlation Coefficients between Measures in Study 2

	SelectionGL	Numeracy	Graph Literacy	Subjective Numeracy	Subjective Graph Literacy
LyingGL	.37**	.30**	.41**	.28**	.20**
SelectionGL		.27**	.39**	.19**	.24**
Numeracy			.43**	.30**	.26**
Graph Literacy				.37**	.28**
Subjective Numeracy					.61**

** $p < .01$.

The performance of two short forms of the LyingGL scale was compared to the LyingGL scale (see Table 2.8). Short form A fell short of the full-scale performance, while short form B surpassed the performance of the full scale due to the elimination of poor items (i.e., better interval scaling of underlying psychometric skill models). Due to its superior performance short form B was selected as the final short form, although it is worth noting that there was some evidence of a limited range of difficulty for the items. Therefore, to increase the range of difficulty one new item was added to Study 3 and can be found in Supplemental File C.

Table 2.7

Percent of Correct Responses and the Item-Total Bivariate Correlation Coefficients for the LyingGL Scale

Item	Lie	Correct	Chance	Total	Item	Lie	Correct	Chance	Total
1[b]	***II***	***39.1***	***25***	***.40****	10	X	15.7	14	.20**
2	**II**	**10.7**	**14**	**.32****	***11[b]***	***X***	***27.1***	***14***	***.43****
3	X	14.0	14	.21**	12	R	21.7	25	.26**
4[a,b]	***Y***	***34.1***	***25***	***.43****	**13**	**TA**	**6.4**	**14**	**.34****
5	TA	23.1	14	.23**	14	GT	29.8	14	.26**
6[b]	***Y***	***29.1***	***14***	***.47****	***15[a,b]***	***II***	***40.5***	***25***	***.43****
7[a,b]	***Y***	***54.2***	***25***	***.42****	**16[a]**	**GT**	**53.8**	**14**	**.39****
8[a,b]	***X***	***26.8***	***14***	***.45****	17	II	37.5	25	.27**
9	X	11.7	14	.01					

Note. II = Irrelevant Information. X = X-Axis Scale/Labels. Y = Y-Axis Scale/Labels. TA = Truncated Axis. R = Reverse Ordered Axis. GT = Graph Type. Item-total correlation coefficients indicated in bold are discriminating items. Superscript letters indicate which of the short forms the item was included. Italic items are included on the final short form.
** $p < .01$

Table 2.8

Comparison of LyingGL Short Forms verses the Full Scale

Form	α	R^2	Numeracy	Graph Literacy	SelectionGL	Subjective Numeracy	Subjective Graph Literacy
Full	.52	.19**	.15**	.23**	.20**	.11	.00
A	.36	.17**	.19*	.21**	.15^	.05	.02
B	**.56**	**.22****	**.17****	**.30****	**.19****	**.08**	**.03**

Note. α = Cronbach's α. Standardized β's are presented for each of the cognitive abilities scales. Short form B, indicated in bold, was selected as the final short form.
* $p \leq .05$, ** $p \leq .01$, ^ $p = .058$

The Cronbach's α for the SelectionGL measure was .44 in Study 2, compared to .42 in Study 1. The similar Cronbach's α's gives some indication of the reliability of this measure across samples.

Study 3: DesignGL and Task Difficulty.

The development of the iGERD tutor required a large pool of possible tasks. In order to determine where in the training a task should be presented, the task difficulty had to be determined.

Participants. Data were collected from 1045 participants. A final sample of 862 participants was used for analysis after 17.5% of participants with incomplete data were removed. The mean age was 37 with a range of 18 to 75. The sample was comprised of 59.7% females, 39.2% males, and 1% of participants preferred not to indicate their sex. Most (89.1%) participants had some college education or better, with only 10.9% of participants reporting less education; one participant with no schooling completed, one participant with nursery school to eighth grade completed, 11 participants with some high school, and 81 with a high school diploma or equivalent. Most (63.8%) of the participants were currently employed with 36.1% being in less than a management position. The mean annual household income range was between $35,000 and $49,999.

Materials. The materials from Study 2 were used with the following modification. LyingGL short form B was used. One additional easy item was added to the LyingGL scale (see Supplemental File C). A number sense scale (Siegler & Opfer, 2003) was also added between the Berlin Numeracy Test and the Graph Literacy Scale. The number

sense scale requires participants to estimate where a number falls on a number line between 0 and 1000 (see Supplemental File A for all items and scoring instructions). Following the LyingGL scale, the graph design tasks were added. The graph design tasks are presented in Supplemental File C.

Procedure. The same procedure as Study 1 and 2 was used for Study 3 with the following exceptions. Due to the large number of items needed for the iGERD, and the time constraints of Amazon Mechanical Turks, each participant was randomly assigned to complete 10 or 11 graph design items. Each item was completed by approximately 100 participants.

Data analysis. All items were scored according to the procedures outlined in Study 1 and 2. Graph design items were scored out of a possible two points for each item, one point for selecting the correct graph type and one point for selecting the correctly designed graph. A total graph design score was computed by summing the scores on the individual items, then dividing the sum by the total possible for the block and multiplying by 100. The DesignGL scores have a possible range from 0 to 100 as a result of the scoring procedure. In order to determine if there were any differences between the blocks, ANOVAs were conducted.

Bivariate correlations were used to investigate relationships between graph design scores and the existing cognitive ability measures. To investigate the relations found with the bivariate correlations in more detail, linear regressions were conducted, with graph design score as the dependent. Step 1 of the regression entered numeracy, number sense,

graph literacy, SelectionGL, and LyingGL scores as predictors and Step 2 added subjective numeracy and subjective graph literacy. Cronbach's α's was also calculated for the SelectionGL and LyingGL measures.

In order to determine the difficulty of the graph design items, the percent correct was calculated for each item. After determining the difficulty of the items, bivariate correlations between each item and the total graph design score were used to determine the discrimination of each item. Items with correlation coefficients greater than or equal to .300 were said to discriminate between skill levels. Cronbach's α was also calculated for the full scale.

After the overall item analysis on the graph design items was completed using the same method as in Study 1 and 2. In addition, data were analyzed using the R (R Core Team, 2014) ltm package (Rizopoulos, 2006) to complete a comprehensive item analysis.

Results. The descriptive statistics and maximum score for each measure are displayed in Table 2.9. Scores on the pre-existing cognitive ability measures were similar to Study 1 and 2.

A one-way ANOVA was used to determine if scores were different between blocks. Only one measure, DesignGL, for Block 7 was different from the other scores, F (7, 850) = 3.31, p = .002. A Bonferoni post-hoc test revealed Block 7 was statistically different from Blocks 1, 2, 4, 5, and 8.

Bivariate correlations were used to determine if relationships exist between scores on the cognitive ability measure. All measures are correlated with each other (see Table

2.10 for correlation coefficients). Additionally, the correlation patterns of the existing measures are consistent with Study 1 and 2.

Numeracy (β = .12, p < .001), graph literacy (β = .22, p < .001), SelectionGL (β = .29, p < .001), LyingGL (β = .13, p < .001), and number sense (β = -.10, p = .001) predicted DesignGL performance (R^2 = .36, p < .001). Subjective numeracy (β = .001, p = .98) and subjective graph literacy (β = .04, p = .31) were not statistically significant predictors of DesignGL performance.

Numeracy (β = .12, p < .001), graph literacy (β = .22, p < .001), SelectionGL (β = .29, p < .001), LyingGL (β = .13, p < .001), and number sense (β = -.10, p = .001) predicted DesignGL performance (R^2 = .36, p < .001). Subjective numeracy (β = .001, p = .98) and subjective graph literacy (β = .04, p = .31) were not statistically significant predictors of DesignGL performance.

Numeracy (β = .12, p < .001), graph literacy (β = .22, p < .001), SelectionGL (β = .29, p < .001), LyingGL (β = .13, p < .001), and number sense (β = -.10, p = .001) predicted DesignGL performance (R^2 = .36, p < .001). Subjective numeracy (β = .001, p = .98) and subjective graph literacy (β = .04, p = .31) were not statistically significant predictors of DesignGL performance.

Table 2.9

Descriptive Statistics for Cognitive Ability Measures' Scores in Study 3

Measure	Mean (SD)	Median	Maximum
Numeracy	2.18 (1.09)	2.00	4
Graph Literacy	13.06 (2.74)	14.00	17
Subjective Numeracy	4.34 (0.91)	4.5	6
Cognitive Abilities	4.13 (1.14)	4.25	6
Preference	4.55 (0.98)	4.75	6
Subjective Graph Literacy	4.14 (1.09)	4.2	6
SelectionGL	2.69 (1.36)	3.00	6
LyingGL	3.12 (1.87)	3.00	9
DesignGL	36.01 (15.42)	35.00	100
Number Sense	535.76 (600.08)	390.50	0

Note. SD = Standard Deviation

Table 2.10

Bivariate Correlation Coefficients between Measures in Study 3

	LyingGL	SelectionGL	Num.	Graph Lit.	Sub. Num.	Sub. Graph Lit.	Number Sense
DesignGL	.34**	.47**	.34**	.34**	.23**	.23**	-.30**
LyingGL		.30**	.29**	.35**	.23**	.27**	-.14**
SelectionGL			.25**	.39**	.19**	.20**	-.24**
Numeracy				.40**	.28**	.22**	-.20**
Graph Literacy					.34**	.29**	-.40**
Subjective Numeracy						.59**	-.16**
Subjective Graph Lit.							-.11**

** $p < .01$.

Numeracy ($\beta = .12$, $p < .001$), graph literacy ($\beta = .22$, $p < .001$), SelectionGL ($\beta = .29$, $p < .001$), LyingGL ($\beta = .13$, $p < .001$), and number sense ($\beta = -.10$, $p = .001$) predicted DesignGL performance ($R^2 = .36$, $p < .001$). Subjective numeracy ($\beta = .001$, $p = .98$) and subjective graph literacy ($\beta = .04$, $p = .31$) were not statistically significant predictors of DesignGL performance.

Item analysis was conducted using classical testing theory and item response theory with a two parameter model, difficulty and discriminability to determine a task order for the SelectionGL and DesignGL tasks as well as select items for the DesignGL measure. See Appendix A for item analysis tables. Twenty tasks were selected for both

the SelectionGL and DesignGL trainings. Tasks were selected and arranged in increasing difficulty based on the item analysis results. Tables 2.11 and 2.12 display the task orders and difficulty measures for the SelectionGL and DesignGL tasks respectively.

The DesignGL measure included 9 items selected based on their difficulty and discriminability. Items over 0.30 were considered to have acceptable discriminability to be included in the measure.

Table 2.11

SelectionGL Task Order and Difficulty

Order Number	Task Number	Graph Type	Difficulty
1	63	Decision Tree	0.99
2	36	Pie	0.89
3	34	Line	0.85
4	32	Line	0.78
5	29	Line	0.75
6	60	Decision Tree	0.70
7	58	Line	0.64
8	85	Bar	0.60
9	10	Bar	0.54
10	13	Bar	0.50
11	8	Pie	0.49
12	83	Bar	0.46
13	59	Icon Array	0.42
14	4	Icon Array	0.40
15	73	Icon Array	0.39
16	75	Icon Array	0.38
17	53	Pie	0.32
18	47	Pie	0.29
19	80	Decision Tree	0.15
20	20	Decision Tree	0.14

Note. Difficulty is equivalent to the percent of the sample responding correctly to the item.

Table 2.12

DesignGL Task Order and Difficulty

Order Number	Task Number	Graph Type	Difficulty
1	48	Pie	0.50
2	45	Pie	0.44
3	15	Line	0.39
4	66	Line	0.37
5	46	Pie	0.30
6	12	Bar	0.27
7	3	Icon Array	0.26
8	14	Bar	0.20
9	11	Bar	0.23
10	74	Icon Array	0.17
11	69	Pie	0.18
12	54	Decision Tree	0.14
13	33	Line	0.14
14	25	Line	0.12
15	72	Icon Array	0.08
16	77	Decision Tree	0.06
17	81	Bar	0.06
18	18	Bar	0.09
19	22	Icon Array	0.04
20	55	Decision Tree	0.02

Note. Difficulty is equivalent to the percent of the sample responding correctly to the item.

Table 2.13

DesignGL Items' Difficulty and Discriminability

Task Number	Graph Type	CTT Discriminability	CTT Difficulty	IRT Difficulty
7	Line	0.42	0.00	0.13
19	Line	0.33	0.31	0.31
28	Line	0.35	0.12	0.04
32	Line	0.33	0.07	0.04
42	Pie	0.49	0.61	0.73
43	Pie	0.46	0.73	1.00
50	Pie	0.48	0.50	0.49
59	Icon Array	0.49	0.20	0.19
76	Decision Tree	0.40	0.02	0.01

Note. CTT = classical testing theory; IRT = item response theory. CTT difficulty and IRT difficulty are equivalent to the percent of the sample responding correctly to the item.

Phase 2: iGERD Effectiveness and User Experience Testing

In Phase 2, I integrated the newly developed test items along with the test models to the effectiveness of the Intelligent Graphs for Everyday Risky Decisions (iGERD) Tutor verses an existing tutor. The longer term goal for the iGERD tutor will be to offer it as a freely available outreach program at RiskLiteracy.org, allowing users to learn how to use graphs for the risky decisions they face daily (e.g., as part of college courses in cognitive science, e-health or e-finance, risk communication, statistics, decision science, etc.). The tutor will be hosted and managed via a Learning Management System (LMS) on a dedicated server that will be established at a later date. The tutor will be structured to include one student model module and (at least) two general training modules.

Implementation Structure

LMS. Moodle (Dougiamas, 1999) will be used as the LMS for iGERD and act as the implementation platform. Moodle was selected because it is a free open-source LMS and is compatible with the intelligent tutor components. Moodle will present the modules to the user and track their progress through iGERD.

Intelligent Tutor. The intelligent tutor tasks were developed using the program Cognitive Tutor Authoring Tools (CTAT; Aleven, McLaren, Sewall, & Koedinger, 2009; Koedinger, Aleven, & McLaren, 2009). CTAT requires the developer to create a graphic user interface (GUI) in either Java or Flash before creating the individual tasks. The iGERD GUI's were developed in Flash using the CTAT specific components included in the CTAT package. Each training module has a GUI designed to support the tasks within

the module. The CTAT components of the GUI's are controlled by a CTAT behavior graph, which can be either programmed using production rules or example tracing. In the interest of efficiency, the iGERD tutor uses example tracing for both modules.

Creating a behavior graph using example tracing requires the developer to work through the task until completion. After completing the most direct path to the solution, the developer can add alternative paths by either editing the behavior graph directly in CTAT or by choosing the point in the behavior graph where an alternative action is possible and then demonstrating the new action sequence. Once all the possible paths are demonstrated in CTAT, the developer can label action correctness and add skill labels and hints to the actions. The developer can also specify if an action should be graded or if a group of steps can be done in any order. When a user is completing a task using a CTAT tutor, CTAT compares the users' input against what has been labeled as the correct action. The developer can create variables, formulas, and short computer code to make a demonstrated action adaptable to multiple tasks (e.g., a single task might present information about risks that need to be graphed in the context of baseball or health or shopping or savings statistics).

Modules. There are two types of module in the iGERD, student model creation and training. This distinction is important as the student model creation module was not designed using CTAT. This decision is a reflection of the fact that the content structure for the student model creation module is survey-based rather than task-based.

Student model creation. The student model creation module is comprised of existing and newly created, individual difference measures, and psychometrically

optimized based on the results of Study 3 (see the Materials sections for more details). The student model can be used to determine areas in which the student needs to improve in and direct where in the training the student should start.

Training. The training modules, focused on two broad categories of skills that echo the skills assessed and modeled in Phase 1, including: 1) selecting the correct graph type for the data given specific user's goals, and 2) creating easy to understand graphs based on data. Each training module is comprised of two sections. The first section gives users the content knowledge in a text based form and requires them to complete a comprehension quiz to move onto the next section. The second section will be the intelligent tutor tasks.

For each training module users also completed a user experience (UX) evaluation of the training content, tasks, and GUI (see Phase 2 Materials for more details). This data was used to assess possible areas of improvement for future versions of the iGERD and other tutors developed in our laboratory. Data was also used to build and tests structural cognitive process models that explain the various relations between emotion, usability, performance, workload, and learning (e.g., structural equation modeling or multifactorial modeling via the SPSS process macro).

SelectionGL. The first training module is designed to train users on how to select the correct graph for specific types of data and goals. The users are given the basic rule of when to use each graph type and complete a comprehension quiz on the content. The users must complete the comprehension quiz with at least an 80 percent to move onto the

tasks. Users then complete a series of questions to help provide hints to the user determine about the correct graph type. The questions use a scaffolding structure in order for the users to practice the rules needed to determine what graph type to use for the data and goals. The early questions are based on the type of data, while later questions focus on the users' goals. The final question of each task is to select the correct graph type. The idea behind using the scaffolding approach is to assist users in creating a cognitive model for the SelectionGL that is rule based (see *Figure 1.2* for screenshot of a SelectionGL task).

Graph design. The second training module focuses on designing easy to comprehend graphs. The text based design knowledge is taken from the graph design guidelines. Users need to complete the comprehension quiz with at least 80 percent to move on to the graph design tasks. The tasks require the user to select the correct graph type for the data and goals before moving onto designing the graph. Users answer a series of questions about the limits and ordering of the data. Then given these data facts, users select the best graph from four options (see *Figure 1.3* for a screenshot of a graph design task). All five types of graph problems were sampled in direct proportion to the results of psychometric modeling in experiment 3 of phase 1 (i.e., understanding some kinds of bar charts may entail the same skills involved in building some icon arrays—and so master of one would serve as an indicator of master of another, per psychometric model specifications).

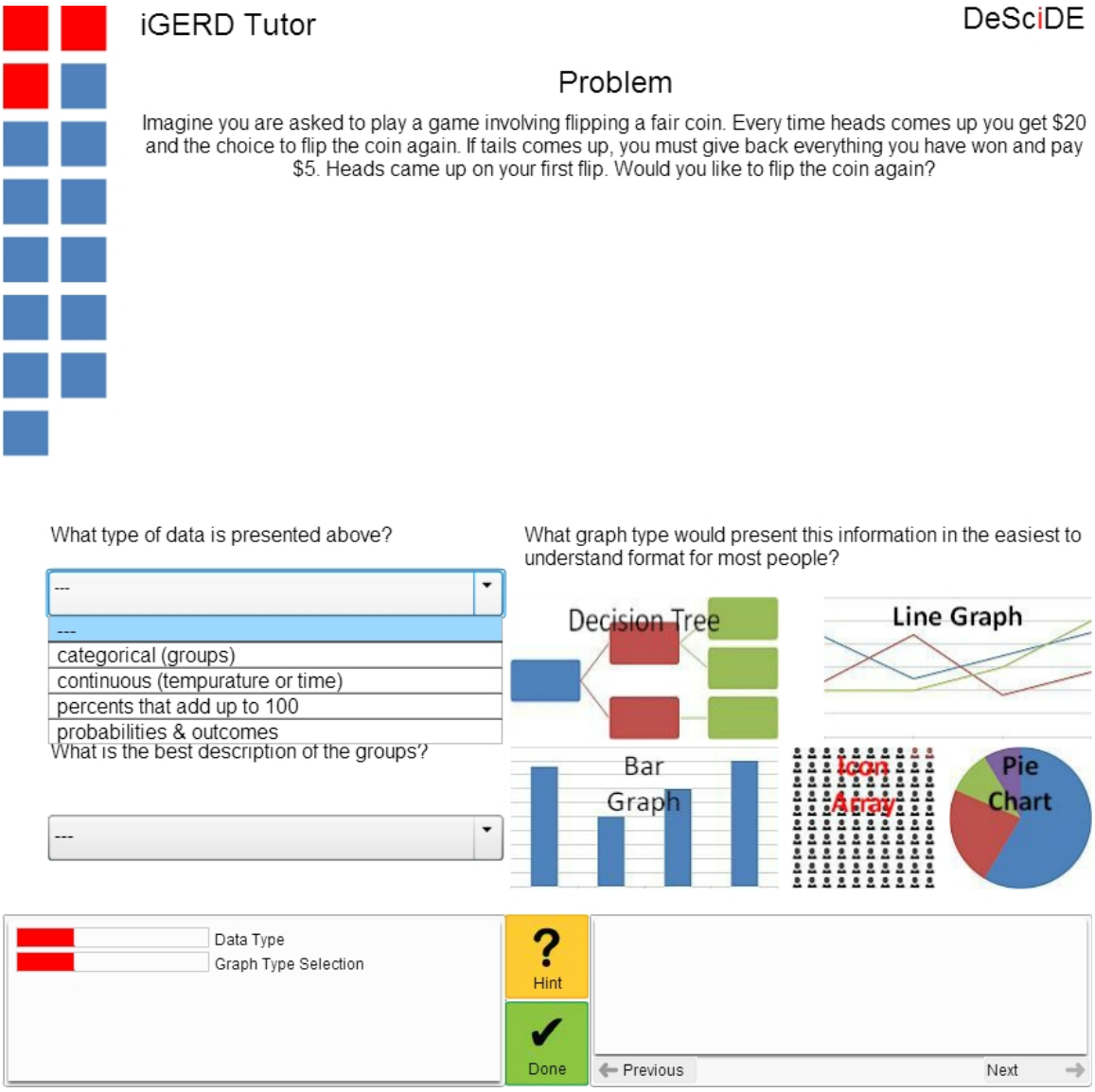


Figure 2.1. Screenshot of a SelectionGL task in iGERD.

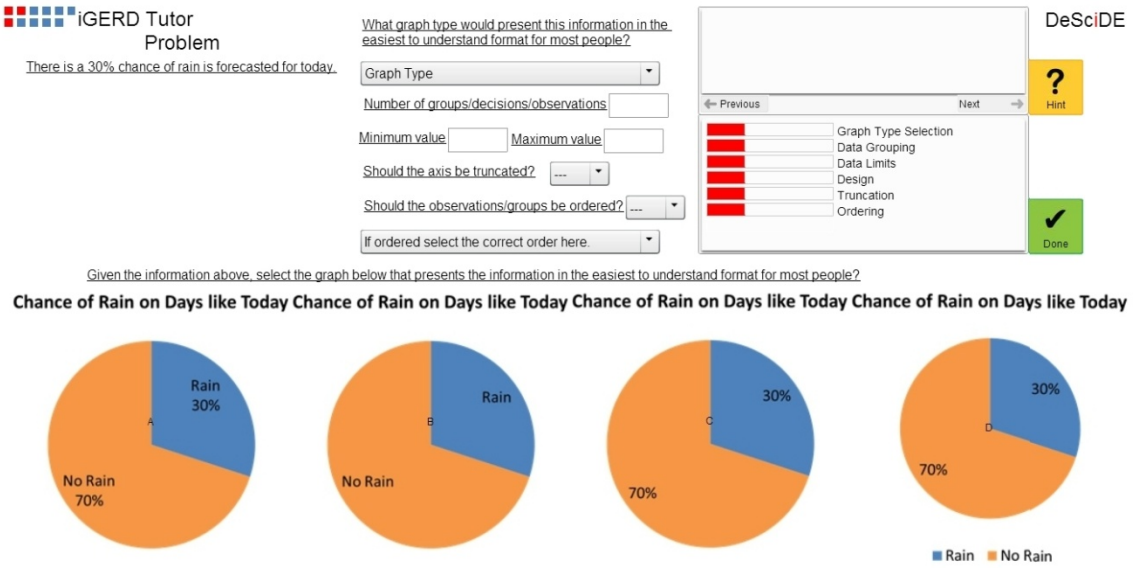


Figure 2.2. Screenshot of a graph design task in iGERD.

Hypotheses.

I conducted extensive and sophisticated statistical analyses to test a host of theoretical models and assumptions. Accordingly, below I present three hypotheses that represent the broad categories of analyses and assumptions I intend to tests and model.

H1: More Graph Literacy Learning. H1a The iGERD tutor is better for developing graph comprehension than a control tutor. H1b The iGERD tutor is better for learning to avoid tricky/biased graphs than a control tutor. Theoretically these results follow from classical learning theory, encoding specificity, and identical elements theory (e.g., the trained skill should develop better across the entire skill range, when training is focused on essential skills).

H2: Improved Risk Literacy and Decision Making. The iGERD tutor will be more likely to directly transfer into risk literacy related skills (e.g., improving numeracy and decision-making skill) than a control tutor. This result again follows from identical elements theory such that users may become better decision makers because they practice metacognitive skills and learn data representation, conceptualization and evaluation skills during graph literacy training. That is, those individuals who get good at thinking in terms of nested and other graphical representations, and those who also practice thinking about thinking during data graphing tasks are practicing the EXACT skills that are often needed to make good decisions (e.g., thinking about thinking and representing various complex data representations). I predict that although users will never be told they are practicing good habits that can empower decision-making, they will be gaining problem solving and metacognitive skills. Therefore, I expect at least some transfer that goes beyond self-efficacy type boosts in motivation (e.g., motivation that comes from succeeding in learning, which should be equally represented in both control and iGERD training). To my knowledge, this would be the first reported evidence of such an effect from training, although it would be consistent with hypotheses and theories suggested by other data (e.g., Cokely et al., 2012; Garcia-Retamero et al., 2013).

H3: Better User Experience. The iGERD tutor will be rated at least as useful, likable, and interesting as the control tutor, although the iGERD tutor will be

rated as more relevant to everyday decisions such as treatment options, politics, and finances.

Design and participants.

A modified mixed-factorial between and within participant design was used. Participants completed the experiment online. The participants were randomly assigned to either the control condition, i.e., training with STEM Foundations (a study skills training), or to the experimental, i.e., Graph Skills training.

Data were collected from 108 participants using Michigan Technological University Introductory Psychology subject pool for participant recruitment and reimbursement. The participants received partial credit toward the completion of their research participation requirement as compensation for their participation in the study. Experimentation began in early April and ended in mid-August. Data from 17 participants were excluded from analysis due to incomplete data, resulting in a control group of 39 participants and an experimental group of 52 participants.

Materials.

Pre-test. The pre-test was comprised of six measures presented in Unipark. Participants completed the subjective numeracy scale (Fagerlin et al., 2007), the subjective graph literacy scale (Garcia-Retamero, Cokely, et al., 2014), The Berlin Numeracy Test-Schwartz, a combination of numeracy tests validated for use with the general population of industrialized countries (i.e., BNT (Cokely et al., 2012) and

Schwartz et al., (1997) in the original forms, a number sense measure (Siegler & Opfer, 2003), and the graph literacy scale (Galesic & Garcia-Retamero, 2011).

Post-test. The post-test was comprised of decision tasks and individual difference measures presented in Unipark. The decision tasks included graph decisions, for both well-designed and biased graphs, and risky decisions. Sixteen well-designed graph decision tasks were taken from Okan, Galesic, and Garcia-Retamero (2015). Only tasks for graphs without conflicts were included and 2 items were added for increased difficultly (see Supplemental File D for new items). All of the well-designed graphs were artificial materials. The biased graph decision tasks were taken from Okan et al. (2013). All the biased graphs were ecological coming from print and electronic media. Three different decision-making measures were included. (1) Berlin Numeracy Components Test (BNT-C), developed by Ghazal (2014) and colleagues, which is an optimized, brief, and comprehensive numeracy test that provides a rapid and robust assessment of one's overall numeracy level as well as differences in one's component numeracy sub-skills (i.e. operations, probability, geometry, and algebra). (2) Numeracy Understanding for Medical Information (NUMi), developed by Schapira et al. (2012), is a 20 item measure of basic health numeracy including graph literacy and statistical numeracy. (3) The Decision Making Skill (DMS) measure (Ghazal, 2014) includes 73 items that measures strategic decision-making performance, risky choice, confidence, and consistency bias.

In addition to the decision tasks, participants also completed twelve individual difference measures. Individual difference measure covered three areas. (1) Graph skills were measured using the subjective graph literacy scale (Garcia-Retamero, Cokely, et al.,

2014), SelectionGL, LyingGL, and DesignGL. (2) Numeracy skills were measured using the subjective numeracy measure (Fagerlin et al., 2007), BNT version 2 (see Supplemental File D for BNT-2 items and scoring procedure), Schwartz form A (see Supplemental File D for Schwartz-A items and scoring procedure), and a number sense measure (Siegler & Opfer, 2003. (3) Personality measures were also included for convergent and discriminant validity. The personality measures included the Ten-Item Personality Inventory (TIPI; Gosling, Rentfrow, & Swann, 2003), a measure of cognitive impulsivity (i.e., the Cognitive Reflection Test (CRT; Frederick, 2005), a measure of one's determined persistent or "grit" (Duckworth, Peterson, Matthews, & Kelly, 2007), and six-item decision making style scale assessing maximization tendencies (Nenkov, Morrin, Ward, Schwartz, & Hulland, 2008).

Trainings. The STEM Foundations tutor (Open Learning Initiative, 2013) trains communication and study skills and was used as the control tutor. It was selected as the control for many reasons. First, the data from the tutor is easy to access and store for analysis. Secondly, it is a free training and compatible with Moodle. Finally, the tutor does not train any decision-making skills, which could have confounded the results of the experiment.

The iGERD and STEM modules were presented using Moodle. Participants were added to the system and enrolled in the course by the researcher.

User experience. At the end of each training module, the participants completed the user experience measures on the module they just completed. Four measures were used to access user experience. (1) An ease of use and evaluation of information

presented were developed specifically for the trainings based on the IBM Usability Standards (Lewis, 1995). Ten non-leading items were created with half being reverse ordered (see Supplemental File E for the ease of use and evaluation of information items and scoring procedure). (2) Graph learning scale developed using the IBM Usability Standards (Lewis, 1995), and assessed learning for each graph type (see Supplemental File E for the graph learning scale and scoring procedure). (3) The System Usability Scale (SUS) was developed by Brooke (1996), and includes 10 items that participants rate the ease of use of the system. The questions are generic and can be used with any system. (4) The NASA- Task Load Index (TLX) is a 6 item measure of task workload on six different dimensions, developed by Hart and Staveland (1988).

Procedure.

Participants accessed the link to the trainings and used the login information provided to them when they signed-up to participate in the study. Participants were randomly assigned to one of the trainings by the researcher. Participants first read and indicated they had read, understood their rights, and agreed to participate in the study. Participants completed the pre-test prior to completing the training at their own pace. At the end of each training module, the participants completed user experience measures on the module they just completed. Information about how the training was completed (e.g., massed v. spaced, plus total time and intervals) was also collected. After completing all of the assigned training modules and user experience surveys, participants completed the post-test.

Results and Discussion.

Did training improve graph literacy? Generalizability of training effectiveness was first modeled using independent-samples t-tests to compare overall change in Graph Literacy (i.e., TrainingGL = SelectionGL plus DesignGL) by condition (i.e., experimental v. control groups).[1] As predicted, the experimental group exhibited the large and significant TrainingGL improvement compared to the control group, t (90) = 5.74, $p \leq$ .001, d = 1.21. Next, I examined all pre-training variables for differences in skills that could complicate interpretation and parameterization of t test results. Pre-training graph literacy was found to be the only pre-training variable found to differ significantly between the groups, t (89) = 2.04, p = .044 (*Cohan's d* = 0.43). Because of the observed difference in pre-training graph literacy, I modeled the relationship between TrainingGL and condition (training v. control) in a multiple regression, statistically controlling for and estimating any influence of pretest graph literacy scores. As expected, the model indicated that the large differences TrainingGL associated with training remained relatively unchanged even when statistically controlling for differences in initial levels of graph literacy, t (89) = 5.23, $p \leq .001$, d = 1.10).

Did training improve general decision making skills? Given the modest sample size and thus modest statistical power, in the light of *a priori* assumptions based on pre

[1] Both SelectionGL, t (89) = 5.74, $p \leq .001$ and a *Cohan's d* = 1.22, and DesignGL, t (89) = 2.91, $p \leq .001$ and a *Cohan's d* = 0.62, were significantly higher for the experimental group compared to the control group.

and post analysis the decision-making tasks were split into two groups of skill tasks as follows: (1) "visualizable" decision tasks that could benefit from an understanding of graphs (e.g., spatially relevant and visualizable decision problems) and (2) "non-visualizable" decisions tasks that should be unrelated to understanding of graphs (i.e., limited to no spatial or visualizable decision context). Specifically, the visualizable tasks included decision skills related to avoiding ratio bias (i.e., icon array), resisting framing effects (i.e., bar graphs), avoiding sunk cost (i.e., decision trees), as well as making decisions based on data presented in biased graphs (e.g., general graph literacy skills), and making decisions about how best to lie with graphs (e.g., metacognitive understanding of graph literacy).[2] The non-visualizable tasks included intertemporal choice, confidence calibration, consistency of risk perception, applying (untrained) decision rules, exhibiting path independence, making ecological risky decisions based on data, and recognizing social norms. Variables were Z scored and integrated into an equally weighted composite overall score for (1) visualizable decisions and (2) non-visualizable decisions. Simple regression modeling revealed a clear and robust effect of training on visualizable decision tasks (see next section for multiple regressions) that was absent in non-visualizable tasks. To refine estimates of training intervention effect sizes and distributions, I used a package in the statistical software language R to bootstrap

[2] Reflecting the small sample size independent-samples t-tests of each decision task indicated only nominal and non-significant statistical trends ($p \geq 0.059$). Bootstrapping results are presented in Appendix B an estimate an overall moderate average effect size for each of the various visualizable decision tasks at around the $d = .4$ level.

estimates of results with simulation of 10,000 resamples (with replacement) from the original data set for each of the visualizable and non-visualizable tasks. Means were calculated for the experimental and the control group for each resample and a difference between the scores was calculated for each resample. Figure 2.1 displays the observed large differences in density functions that obtained for each of the measures under bootstrapped estimation. A 95% confidence interval for each distribution is indicated by dashed lines. The raw estimated effect based only on the true sample difference between the group means is indicated by the solid line representing a good approximation of the population central tendency.

a. Visualizable b. Non-Visualizable

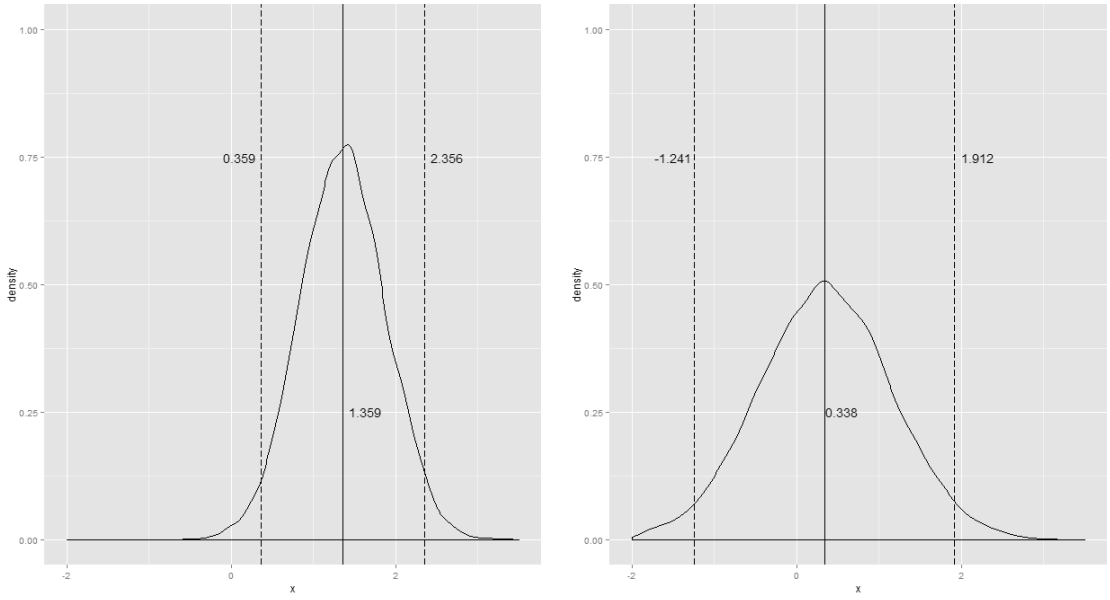


Figure 2.3. Bootstrapping Density Functions of Experimental Minus Control Means. Dashed line indicate the end points of the 95% confidence interval and the solid line

indicates the original mean difference. Robust effects are implied when confidence intervals do not contain zero.

How did user feel about their training experience? Differences in user experience between the groups were investigated using independent-samples t-tests. The iGERD training had significantly higher scores on graph learning (t (85) = 4.31, $p \leq$ .001), NASA TLX performance (t (85) = 2.48, p = .015), and NASA TLX frustration (t (85) = 2.10, p = .039) than the control training. However, the control training outscored the iGERD on ease of use (t (85) = -3.82, $p \leq .001$), SUS (t (85) = -3.95, $p \leq .001$), and NASA TLX physical (t (85) = -2.40, p = .018).

Table 2.14

Hieratical Regression Statistics for Model Comparison

	Non-Visualizable Score					Visualizable Score				
Model & variables	β	R	R^2	ΔR^2	F	β	R	R^2	ΔR^2	F
Model 1.										
Group	.054	.054	.003	.003	.265	.271**	.271	.074	.074	7.07**
Model 2.										
Group	.050					.187				
Graph Literacy	.494***	.486	.236	.233	13.58***	.398***	.475	.225	.152	12.79***
Model 3.										
Group	.201					.084				
Graph Literacy	.349***	.553	.306	.070	12.76***	.299**	.508	.258	.033	10.08***
TrainingGL	.350**					.240*				

Note. * $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$

In order to determine if the usability results affected performance on the visualizable tasks, a series of hierarchical linear regression models were conducted. Two sets of models were constructed. In both cases, the first model predicted decision task using only group. The next model added pre-training graph literacy as a control together with group. The full model added subjective measures of performance or ease of use measures to both pre-training graph literacy and group. Essentially, the full model allowed me to estimate the degree to which the usability measures mediated the relation between condition and decision-making, controlling for pre-training graph literacy scores. Table 2.15 contains the regression statistics for both subjective performance and ease of use scores. For the subjective performance measures, the full model failed show a reliable relationship between the subjective performance measures and decision task performance, F (2, 76) = 1.39, p = .255, R^2_{change} = .028. Also, for the ease of use measures, the full model failed show a reliable relationship between the subjective performance measures and decision task performance, F (2, 76) = .745, p = .478, R^2_{change} = .015. These models show the importance of performance over ease of use. However, once a system meets the performance requirements ease of use should be optimized to reduce frustration and increase user retention.

Table 2.15

Hieratical Regression Statistics for Model Comparison

	Subjective Performance					Ease of Use				
Model & variables	β	R	R^2	ΔR^2	F	β	R	R^2	ΔR^2	F
Model 1.										
Group	.278**	.278	.077	.077	6.59**	.278**	.278	.077	.077	6.59**
Model 2.										
Group	.202*	.467	.218	.141	10.89***	.202*	.467	.218	.141	10.89***
Graph Literacy	.383***					.383***				
Model 3.										
Group	.293**					.188				
Graph Literacy	.332**	.496	.246	.028	6.20***	.357***	.	.	.	.
Graph Learning	-.141					N/A	N/A	N/A	N/A	N/A
NASA TLX Performance	-.083					N/A				
Ease of Use	N/A	N/A	N/A	N/A	N/A	-.206	.483	.233	.015	5.78***
SUS	N/A					.145				

Note. ** $p \leq .01$, *** $p \leq .001$

Chapter 3: Discussion

The current research created, refined, and tested an online training system, the iGERD tutor, for improving graph skills that are essential parts of effective decision-making skill and risk literacy. Past research has shown risk literacy can be improved, sometimes dramatically, via presentation of graphical materials (Garcia-Retamero & Cokely, 2011). Improved graph comprehension skills are directly linked to overcoming and avoiding distorted and biased graphs (Woller-Carter et al., 2012). In turn, reducing the cost of unnecessary/ineffective treatments and screenings defrayed to the public.

This research shows that the iGERD tutor is currently the most efficient and effective training to improve decision making in adults. The training was effective for improving performance on visualizable tasks due to the identical elements in common between the training tasks and the decision making tasks (Holding, 1965; Thorndike & Woodworth, 1901; Yamnill & McLean, 2001). These elements may include new long-term working memory structures that allow for more information to be stored and manipulated while completing the decision task as compared to participants who did not complete the iGERD training (Anderson, Reder, & Simon, 1996; Barton, Cokely, Galesic, Koehler, & Haas, 2009; Cokely & Kelley, 2009; Cokely, Kelley, & Gilchrist, 2006; Cokely, Schooler, & Gigerenzer, 2010; Ericsson, 1985; Ericsson & Charness, 1994; Ericsson & Delaney, 1999; Ericsson & Kintsch, 1995; Ericsson, Prietula, & Cokely, 2007; Gigerenzer & Edwards, 2003; Keller, Cokely, Katsikopoulos, & Wegwarth, 2010).

The exact elements and long-term working memory effects will require further empirical investigation to identify and understand. However, the iGERD training has a moderate improvement over the control training. This improvement may be the result of many things including long-term working memory changes, but the truth is that the iGERD does improve decision making more than the control training.

The approach used to develop the iGERD system should be implemented in the development of new trainings in order to accurately assess training effects of the system. 1) Identify and validate the need for a new training. 2) Identify the skills to be trained. 3) Assess the skills prior to training, normally requiring the development of a new assessment tool. 4) Develop the training leveraging current technology and existing trainings. 5) Training the skills. 6) Compare training improvements to the current training standard. Currently businesses spend between $58.6 billion and $200 billion a year on employee training (Yamnill & McLean, 2001) however, few know the true effect of the training due to poor assessment. Most researchers miss the key step of skill assessment when developing new trainings, resulting in a system without a good indication of the training's effectiveness.

The training developed here has implications beyond the improvement of graph comprehension skills, as the training also lead to improvements in decision making performance. While these results are currently limited by a small sample size, future research should focus on testing the effectiveness of iGERD in the general population, as well as, other well educated samples. The results are expected to hold true for other educated samples as the effect sizes are moderate. The training effects are expected to be

even greater for the general population as there are generally lower levels of numeracy and graph literacy compared to educated samples, allowing more room for improvement.

In order to determine the longevity of the training effects, a longitudinal study will be required. The study would require that participants' decision making skills be tested at points after completing the post-training assessment. Current training studies have shown other training effects to last as long as 12 weeks (Morewedge et al., 2015).

The usability testing of the iGERD has pointed to a few problems with the system that require improvement to increase the ease of use of the iGERD system. In addition, I served as the iGERD Help Desk and assisted users with IT issues. One of the biggest issues was system crashes, as a result of unexpected user interactions with the system. The first simple and cost-effective solution is to create a short tutorial video to accompany the trainings. The video would walk the user through the completion of a task not used in the training and point out potential pitfalls such as not needing to enter data in every field for every task, and hitting the done button to move to the next task. Other options would require the system to be rebuilt and launched on a different platform that allows for dynamic student interfaces.

Future research should also implement an adaptive task structure to the iGERD system to improve user experience and reduce training time. In an adaptive system, users are not forced to complete tasks they already understand, making adaptive systems have higher user experience ratings than traditional systems. Higher user experience ratings also lead to higher completions rates and motivation levels compared to non-adaptive trainings. However, additional tasks will need to be tested and added to the iGERD

system to allow for users to progress at their own speed. This will require additional studies like Study 3 to be completed for additional tasks and training modules.

Conclusion. History suggests that many of the solutions to our most pressing social and economic challenges are technology and information driven. As science advances, will new technologies and bigger pools of risk data make decisions better or will they overwhelm decision makers? In the current research I documented one potentially powerful means of improving people's ability to navigate our complex and data drenched world. While I'm grateful and impressed by the current results, the theoretical and practical value of the current project shouldn't outshine another important lesson. Beyond the benefits of developing the specific iGERD system and contributing to the RiskLiteracy.org decision skills training program, the current project illustrates the timeliness, value and power of adaptive systems that bridge the psychological and technological. A user-friendly future requires user-friendly systems that can respond and provide appropriate feedback in real time. Opportunities abound for those who dare to innovate and create adaptive systems using scientific approaches to measurement, assessment, and design of interactive cognitive systems.

References

Aleven, V., McLaren, B. M., Sewall, J., & Koedinger, K. R. (2009). A new paradign for intelligent tutoring systems: Example-tracing tutors. *International Journal of Articial Intelligence in Education, 19*(2), 105-154.

Anderson, J. R., Reder, L. M., & Simon, H. A. (1996). Situated learning and education. *Educational Researcher, 25*(4), 5-11.

Barton, A., Cokely, E. T., Galesic, M., Koehler, A., & Haas, M. (2009). *Comparing Risk Reductions: On the Interplay of Cognitive Strategies, Numeracy, Complexity, and Format.* Paper presented at the 31st Annual Conference of the Cognitive Science Society.

Brisbane, A. (1911, March 28). Speackers give sound advice. *Syracuse Post Standard*.

Brooke, J. (1996). SUS - A quick and dirty usability scale. *Usability eveluation in industry, 189*, 194.

Carnegie Learning. (2011). MATHia. Pittsburgh, PA: Carnegie Learning, Inc.

Carpenter, P. A., & Shah, P. (1998). A Model of the Perceptual and Conceptual Processes in Graph Comprehension. *Journal of Experimental Psychology: Applied, 4*(2), 75-100.

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring Risk Literacy: The Berlin Numeracy Test. *Judgment and Decision Making, 7*(1), 25-47. Retrieved from <Go to ISI>://WOS:000300401900003

Cokely, E. T., Ghazal, S., Galesic, M., Garcia-Retamero, R., & Schulz, E. (2013). How to measure risk comprehension in educated samples. In R. Garcia-Retamero & M.

Galesic (Eds.), *Transparent Communication of Health Risks* (pp. 165-191). New York: Springer.

Cokely, E. T., Ghazal, S., & Garcia-Retamero, R. (2014). Measuring numeracy. In B. L. Anderson & J. Schulkin (Eds.), *Numerical reasoning in judgments and decision making about health* (pp. 11-38): Cambridge University Press.

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

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. doi:10.3758/BF03213914

Cokely, E. T., Schooler, L. J., & Gigerenzer, G. (2010). Information use for decision making *Encyclopedia of library and information sciences* (3 ed., pp. 2727-2734).

Cooper, R. J., Schriger, D. L., Wallace, R. C., Mikulich, V. J., & Wilkes, M. S. (2003). The Quantity and Quality of Scientific Graphs in Pharmaceutical Advertisements. *J Gen Intern Med, 18*(4), 294-297. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12709097

Dougiamas, M. (1999). Moodle (Version 2.7+) [Software Package]. Perth, Australia: Moodle Pty Ltd.

Ellis, K. M., Cokely, E. T., Ghazal, S., & Garcia-Retamero, R. (2014). Do People Understand their Home HIV Test Results? Risk Literacy and Information Search.

Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 58, 1323-1327.

Ericsson, K. A. (1985). Memory skill. *Canadian Journal of Psychology, 39*(2), 188-231. doi:10.1037/h0080059

Ericsson, K. A., & Charness, A. C. (1994). Expert performance: Its structure and acquisition. *American Psychologist, 49*(8), 725-747. doi:10.1037/0003-066X.49.8.725

Ericsson, K. A., & Delaney, P. F. (1999). Long-term working memory as an alternative to capacity models of working memory in everyday skilled performance. In A. Miyake, Shah, P. (Ed.), *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control* (pp. 257-297): Cambridge University Press.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review, 102*(2), 211-245. doi:10.1037/0033-295X.102.2.211

Ericsson, K. A., Prietula, M. J., & Cokely, E. T. (2007). The making of an expert. *Harvard Business Review, 85*(7-8), 114-121.

Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jancovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring Numeracy without a Math Test: Development of the Subjective Numeracy Scale. *Med Decis Making, 27*(5), 672-680. doi:10.1177/0272989X07304449

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*(4), 25-42. doi:Doi 10.1257/089533005775196732

Friendly, M. (2008). A brief history of data visualization. In C. Chen, W. Hardle, & A. Unwin (Eds.), *Handbook of data visualization* (pp. 15-56). Berlin Heidelberg: Springer.

Galesic, M., & Garcia-Retamero, R. (2011). Graph literacy: a cross-cultural comparison. *Med Decis Making, 31*(3), 444-457. doi:10.1177/0272989X10373805

Galesic, M., Garcia-Retamero, R., & Gigerenzer, G. (2009). Using icon arrays to communicate medical risks: overcoming low numeracy. *Health Psychology, 28*(2), 210-216. doi:10.1037/a0014474

Garcia-Retamero, R., & Cokely, E. T. (2011). Effective communication of risks to young adults: using message framing and visual aids to increase condom use and STD screening. *J Exp Psychol Appl, 17*(3), 270-287. doi:10.1037/a0023677

Garcia-Retamero, R., & Cokely, E. T. (2012). Advances in efficient health communication: Promoting prevention and detection of STDs. *Current HIV Research, 10*(3), 262-270. doi:10.2174/157016212800618084

Garcia-Retamero, R., & Cokely, E. T. (2013). Communicating Health Risks With Visual Aids. *Current Directions in Psychological Science, 22*(5), 392-399. doi:10.1177/0963721413491570

Garcia-Retamero, R., & Cokely, E. T. (2014a). The influence of skills, message frame, and visual aids on prevention of sexually transmitted diseases. *Journal of Behavioral Decision Making, 27*(2), 179-189. doi:10.1002/bdm.1797

Garcia-Retamero, R., & Cokely, E. T. (2014b). Using visual aids to help people with low numeracy make better decisions. In B. L. Anderson & J. Schulkin (Eds.),

Numerical Reasoning in Judgments and Decision Making about Health (pp. 153-174): Cambridge University Press.

Garcia-Retamero, R., & Cokely, E. T. (2015a). Brief Messages to Promote Prevention and Detection of Sexually Transmitted Infections. *Current HIV Research, 13*(5), 408-420.

Garcia-Retamero, R., & Cokely, E. T. (2015b). Simple but powerful health messages for increasing condom use in young adults. *The Journal of Sex Research, 52*(1), 30-42.

Garcia-Retamero, R., & Cokely, E. T. (Submitted). Designing visual aids that promote health risk literacy: A theoretical framework and simple how-to guides. *Front Psychol*.

Garcia-Retamero, R., Cokely, E. T., & Galesic, M. (2013). Reducing the Effect of Framed Messages About Health. In R. Garcia-Retamero & M. Galesic (Eds.), *Transparent Communication of Health Risks* (pp. 165-191). New York: Springer.

Garcia-Retamero, R., Cokely, E. T., Ghazal, S., & Hanson, B. (2014). Measuring subjective graph literacy in diverse samples and cultures. *Medical Decision Making, submitted*.

Garcia-Retamero, R., & Galesic, M. (2010). Who profits from visual aids: overcoming challenges in people's understanding of risks. *Soc Sci Med, 70*(7), 1019-1025. doi:10.1016/j.socscimed.2009.11.031

Garcia-Retamero, R., & Galesic, M. (Eds.). (2013). *Transparent Communication of Health Risks: Overcoming Cultural Differences* (First ed.). New York: Springer-Verlag.

Garcia-Retamero, R., Okan, Y., & Cokely, E. T. (2012). Using Visual Aids to Improve Communication of Risks about Health: A Review. *The Scientific World Journal, 2012*, 1-10. doi:10.1100/2012/562637

Garcia-Retamero, R., Wicki, B., Cokely, E. T., & Hanson, B. (2014). Factors predicting surgeons' preferred and actual roles in interactions with their patients. *Health Psychology, 33*(8), 920-928.

Ghazal, S. (2014). *Component Numeracy Skills and Decision Making*. Dissertation. Cognitive and Learning Sciences. Michigan Technological University.

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

Gigerenzer, G. (2012). Risk literacy. In J. Brockman (Ed.), *This will make you smarter: New scientific concepts to improve your thinking* (pp. 259-261). New York, NY: Harper Perennial.

Gigerenzer, G., & Edwards, A. (2003). Simple tools for understanding risks: from innumeracy to insight. *British Medical Journal, 327*(7417), 741-744. doi:10.1136/bmj.327.7417.741

Gillan, D. J., Wickens, C. D., Hollands, J. G., & Carswell, C. M. (1998). Guidelines for Presenting Quantitative Data in HFES Publications. *Human Factors: The Journal*

of the Human Factors and Ergonomics Society, 40(1), 28-41. doi:10.1518/001872098779480640

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality, 37*(6), 504-528. doi:Doi 10.1016/S0092-6566(03)00046-1

Hallett, T., & MindTools.com. (2007). Charts and Graphs: Choosing the right Format. Retrieved from MindTools website: https://www.mindtools.com/pages/article/Charts_and_Diagrams.htm Retrieved from https://www.mindtools.com/pages/article/Charts_and_Diagrams.htm

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology*, 139-183.

Holding, D. H. (1965). *Principles of Training* (First ed.). Oxford: Pergamon Press.

Huff, D. (1952). *How to Lie with Statistics*. New York, NY: W. W. Norton & Company, Inc.

Jarvenpaa, S. L., & Dickson, G. W. (1988). Graphics and Managerial Decision-Making - Research Based Guidelines. *Communications of the ACM, 31*(6), 764-774. doi:10.1145/62959.62971

Keller, N., Cokely, E. T., Katsikopoulos, K. V., & Wegwarth, O. (2010). Naturalistic heuristics for decision making. *Journal of Cognitive Engineering and Decision Making, 4*(3), 256-274. doi:10.1518/155534310X12844000801168

Koedinger, K. R., Aleven, V., & McLaren. (2009). Cognitive Tutor Authoring Tools (Version 3.3.0) [Computer Program]. Pittsburgh, PA: Carnegie Mellon University.

Kosslyn, S. M. (2006). *Graph design for the eye and mind*. New York, NY: Oxford University Press.

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science, 11*(1), 65-99. doi:10.1016/S0364-0213(87)80026-5

Lewis, J. R. (1995). IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use. *International Journal of Human-Computer Interaction, 7*(1), 57-78. doi:10.1080/10447319509526110

Lipkus, I. M., & Hollands, J. G. (1999). The visual communication of risk. *Journal of the National Cancer Institute Monographs, 1999*(25), 149-163. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10854471

McCarley, J. S., Steelman, K. S., Grable, J. E., Palmer, L., Yeske, D., & Chaffin, C. R. (2015). The psychology of decisions: A short tutorial. In C. R. Chaffin (Ed.), *CFP Board Financial Planning Competency Handbook* (2 ed., pp. 843-858). New Jersey: Wiley.

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. doi:10.1177/2372732215600886

Mt-Isa, S., Hallgreem, C. E., Asiimwe, A., Downey, G., Genov, G., Hermann, R., . . . Tzoulaki, I. (2013). *Review of visualisation methods for the representation of benefit-risk assessment of medication: Stage 2 of 2*. Retrieved from http://www.imi-protect.eu/documents/ShahruletalReviewofvisualisationmethodsfortherepresentationofBRassessmentofmedicationStage2A.pdf

Nelson, D. E., Hesse, B. W., & Croyle, R. T. (2009). *Making data talk: Communicating public health data to the public, policy makers, and the Press*. New York, NY: Oxford University Press.

Nenkov, G. Y., Morrin, M., Ward, A., Schwartz, B., & Hulland, J. (2008). A short form of the Maximization Scale: Factor structure, reliability and validity studies. *Judgment and Decision Making, 3*(5), 371-388.

Okan, Y., Galesic, M., & Garcia-Retamero, R. (2015). How people with low and high graph literacy process graphs: Evidence from eye tracking. *Journal of Behavioral Decision Making*. doi:10.1002/bdm.1891

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. doi:10.1002/bdm.751

Okan, Y., Garcia-Retamero, R., Cokely, E. T., & Maldonado, A. (2015). Improving risk understanding across ability levels: Encouraging active processing with dynamic icon arrays. *American Psychological Association, 21*(2), 178-194.

Okan, Y., Woller-Carter, M. M., Garcia-Retamero, R., & Cokely, E. T. (2013). *Predicting errors in graphical interpretation: Evidence from medical, financial, and political decision making*. Paper presented at the Subjective Probability, Utility, and Decision Making Conferance, Barcelona, Spain.

Open Learning Initiative. (2013). STEM Foundations. Retrieved from https://oli.cmu.edu/jcourse/lms/students/syllabus.do?section=cccf4d9e80020ca6004f90b2bdd9abda

Peters, E. (2012). Beyond comprehension: The role of numeracy in judgements and decisions. *Current Directions in Psychological Science, 21*(1), 31-35. doi:10.1177/0963721411429960

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*(7), 847-858.

Pinker, S. (1990). A Theory of Graph Comprehension. In R. Freedie (Ed.), *Artificial Intelligence and The Future of Testing* (pp. 73-126). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

R Core Team. (2014). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Reyna, V. F., Nelson, W. L., Han, P. K., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychol Bull, 135*(6), 943-973. doi:10.1037/a0017327

Rizopoulos, D. (2006). ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses. *Journal of Statistical Software, 17*(5), 1-25. Retrieved from http://www.jstatsoft.org/v17/i05/

Schapira, M. M., Walker, C. M., Cappaert, K. J., Ganschow, P. S., Fletcher, K. E., McGinley, E. L., . . . Jacobs, E. A. (2012). The Numeracy Understanding in Medicine Instrument: A Measure of Health Numeracy Developed Using Item Response Theory. *Medical Decision Making, 32*, 851-865.

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

Siegler, R. S., & Opfer, J. E. (2003). The development of numerical estimation: Evidence for multiple representations of numerical quantity. *Psychlological Science, 14*(3), 237-243. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12741747

Spiegelhalter, D., Pearson, M., & Short, I. (2011). Visualizing uncertainty about the future. *Science, 333*(6048), 1393-1400. doi:10.1126/science.1191181

Staudt, C., Win, D., & Dorsey, C. (2011). Graph Literacy. *SmartGraphs.* Retrieved from http://concord.org/projects/graph-literacy#curriculum

Thorndike, E. L., & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions. *Psychological Review, 8*(3), 247-261.

Toth, T. (2006). *Graphing Data for Decision-Making*. Retrieved from http://udspace.udel.edu/handle/19716/2666

Tufte, E. (2001). *The visual display of quantitative information* (Second ed.). Cheshire, CT: Graphics Press.

VanLehn, K. (2006). The Behavior of Tutoring Systems. *International Journal of Artifical Intelligence in Education, 16*(3), 227-265.

Woller-Carter, M. M., Okan, Y., Cokely, E. T., & Garcia-Retamero, R. (2012). Communicating and Distorting Risks with Graphs: An Eye-Tracking Study. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 56*(1), 1723-1727. doi:10.1177/1071181312561345

Yamnill, S., & McLean, G. N. (2001). Theories Supporting Transfer of Training. *Human Resource Development Quarterly, 12*(2), 195-208.

Appendix A: Item Analysis of Study 3

Table A.1

SelectionGL Item Analysis Results

Task Number	Graph Type	CTT Discriminability	CTT Difficulty	IRT Difficulty
1	Icon Array	0.493	23.4	30.6
2	Icon Array	0.496	17.1	1.2
3	Icon Array	0.541	16	42.4
4	Icon Array	0.4	25.2	40.0
5	Icon Array	0.432	40.8	59.8
6	Icon Array	0.396	21.1	25.8
7	Line	0.398	43.2	23.9
8	Pie	0.325	35.1	49.2
9	Bar	0.21	33.6	39.3
10	Bar	0.176	32.2	54.0
11	Bar	0.342	30.5	53.9
12	Bar	0.242	21	48.1
13	Bar	0.286	42.1	50.1
14	Bar	0.451	43.7	76.9
15	Line	0.517	28.2	76.5
16	Icon Array	0.261	19.1	33.5
17	Decision Tree	0.601	38.3	39.2
18	Bar	0.292	36.2	45.3
19	Line	0.191	23.4	54.2

Task Number	Graph Type	CTT Discriminability	CTT Difficulty	IRT Difficulty
20	Decision Tree	0.298	12.8	13.8
21	Decision Tree	0.485	34.6	36.1
22	Icon Array	0.222	13.7	17.9
23	Icon Array	0.314	24	31.1
24	Line	0.539	49	72.1
25	Line	0.427	50	66.7
26	Line	0.431	56.2	57.3
27	Line	0.447	54.3	73.5
28	Line	0.568	58.9	83.8
29	Line	0.346	41.3	75.3
30	Line	0.382	38.3	52.3
31	Line	0.423	31.9	40.9
32	Line	0.43	61.5	77.6
33	Line	0.42	32.7	49.6
34	Line	0.533	55.8	85.4
35	Line	0.466	41.7	43.3
36	Pie	0.497	42	89.1
37	Pie	0.44	2.9	74.9
38	Pie	0.572	21.3	87.7
39	Pie	0.447	14	62.8
40	Pie	0.368	38.3	86.2
41	Pie	0.435	14.7	91.7
42	Pie	0.558	17	88.4
43	Pie	0.3	18.1	96.0
44	Pie	0.369	40.4	73.4

Task Number	Graph Type	CTT Discriminability	CTT Difficulty	IRT Difficulty
45	Pie	0.417	7.4	51.7
46	Pie	0.553	31.8	69.5
47	Pie	0.308	10.6	28.6
48	Pie	0.417	17.5	66.9
49	Pie	0.585	17.5	77.2
50	Pie	0.578	9.5	61.0
53	Pie	0.249	16.3	32.2
54	Decision Tree	0.405	17.5	31.9
55	Decision Tree	0.387	47.6	50.1
56	Line	0.562	36.2	40.2
57	Line	0.487	47.9	72.1
58	Line	0.417	42.1	63.5
59	Icon Array	0.264	21.9	42.1
60	Decision Tree	0.601	54.4	69.8
61	Decision Tree	0.526	36.2	52.7
62	Decision Tree	0.529	61.5	68.3
63	Decision Tree	0.728	36	99.4
64	Decision Tree	0.453	39.8	41.7
65	Line	0.396	21.5	50.5
66	Line	0.384	28.7	69.5
67	Bar	0.406	46.5	71.5
68	Bar	0.142	29	41.0
69	Pie	0.361	22.3	40.5
70	Pie	0.286	15.2	51.9

Task Number	Graph Type	CTT Discriminability	CTT Difficulty	IRT Difficulty
71	Icon Array	0.369	13.7	30.9
72	Icon Array	0.376	26	39.9
73	Icon Array	0.26	24.5	38.9
74	Icon Array	0.417	10.7	33.5
75	Icon Array	0.531	27.1	37.7
76	Decision Tree	0.328	14.7	16.3
77	Decision Tree	0.467	29.9	40.1
78	Decision Tree	0.077	17	19.1
79	Decision Tree	0.515	26.7	14.0
80	Decision Tree	0.238	14.7	15.2
81	Bar	0.273	42	48.1
82	Bar	0.313	22.3	48.9
83	Bar	0.289	40.4	45.6
84	Bar	0.307	41.1	62.9
85	Bar	0.446	51.5	60.4

Note. CTT = Classical Testing Theory; IRT = Item Response Theory. Difficulty is equivalent to the percent of the sample responding correctly to the item.

Table A.2

DesignGL Item Analysis Results

Task Number	Graph Type	CTT Discriminability	CTT Difficulty	IRT Difficulty
1	Icon Array	0.34	13.1	9.7
2	Icon Array	0.264	12.4	9.9
3	Icon Array	0.198	27.7	25.6
4	Icon Array	0.469	14	7.2
5	Icon Array	0.263	18.4	15.9
6	Icon Array	0.444	7	1.7
7	Line	0.419	0	13.3
8	Pie	0.277	14	12.2
9	Bar	0.273	5.6	5.6
10	Bar	0.114	21	21.9
11	Bar	0.091	23.2	23.0
12	Bar	0.203	27.6	27.3
13	Bar	0.071	7.9	7.9
14	Bar	0.379	29.1	19.7
15	Line	0.425	39.8	38.9
16	Icon Array	0.087	14.9	14.7
17	Decision Tree	0.06	5.6	3.9
18	Bar	0.118	9.5	9.3
19	Line	0.329	30.8	30.8
20	Decision Tree	0.033	2.1	0.3
21	Decision Tree	-0.028	3.8	0.0

Task Number	Graph Type	CTT Discriminability	CTT Difficulty	IRT Difficulty
22	Icon Array	0.055	4.2	4.1
23	Icon Array	-0.182	6.7	7.6
24	Line	0.057	17.3	15.1
25	Line	0.263	13.2	12.4
26	Line	0	0	0.0
27	Line	0.12	18.1	17.2
28	Line	0.354	11.6	3.6
29	Line	0.258	24.5	23.7
30	Line	0.217	13.8	13.7
31	Line	0.264	10.6	8.6
32	Line	0.328	6.7	3.6
33	Line	-0.077	16.3	13.5
34	Line	0.426	17.9	9.6
35	Line	0.033	2.9	2.6
36	Pie	0.342	37.1	34.0
37	Pie	0.259	70.9	72.6
38	Pie	0.354	57.4	62.2
39	Pie	0.212	46.7	46.5
40	Pie	0.23	45.7	44.9
41	Pie	0.224	72.6	78.3
42	Pie	0.487	60.6	72.8
43	Pie	0.457	73.3	100.0
44	Pie	0.334	29.8	29.3
45	Pie	0.367	44.2	43.7
46	Pie	0.17	31.8	30.3

Task Number	Graph Type	CTT Discriminability	CTT Difficulty	IRT Difficulty
47	Pie	0.4	18.3	16.7
48	Pie	0.29	50.3	49.6
49	Pie	0.377	55.3	57.4
50	Pie	0.483	49.5	49.5
53	Pie	0.443	15.4	13.8
54	Decision Tree	0.151	15.5	14.4
55	Decision Tree	0.104	2.1	2.0
56	Line	0.253	5.7	0.0
57	Line	0.246	18.1	18.0
58	Line	0.092	20.2	19.7
59	Icon Array	0.492	20.2	19.4
60	Decision Tree	-0.212	2.6	1.5
61	Decision Tree	0.15	15.2	14.2
62	Decision Tree	0.179	1.9	0.5
63	Decision Tree	0.31	26.3	24.6
64	Decision Tree	-0.065	3.9	1.2
65	Line	0.397	29	25.0
66	Line	0.149	36.4	36.6
67	Bar	0.33	22.8	14.5
68	Bar	0.192	12.1	12.1
69	Pie	0.162	18.4	18.4
70	Pie	0.299	36.2	36.6
71	Icon Array	0.387	18.9	17.3
72	Icon Array	0.273	14.4	8.2

Task Number	Graph Type	CTT Discriminability	CTT Difficulty	IRT Difficulty
73	Icon Array	0.381	14.7	6.4
74	Icon Array	0.382	23.3	17.0
75	Icon Array	0.383	13.1	6.3
76	Decision Tree	0.404	2.1	1.4
77	Decision Tree	0.326	12.1	5.8
78	Decision Tree	-0.126	2.1	1.3
79	Decision Tree	0	0	0.0
80	Decision Tree	-0.12	1.1	0.9
81	Bar	-0.064	6.3	6.2
82	Bar	0.001	26.6	25.0
83	Bar	-0.055	5.8	3.8
84	Bar	0.13	21.1	21.0
85	Bar	-0.231	6.8	6.5

Note. CTT = Classical Testing Theory; IRT = Item Response Theory. Difficulty is equivalent to the percent of the sample responding correctly to the item.

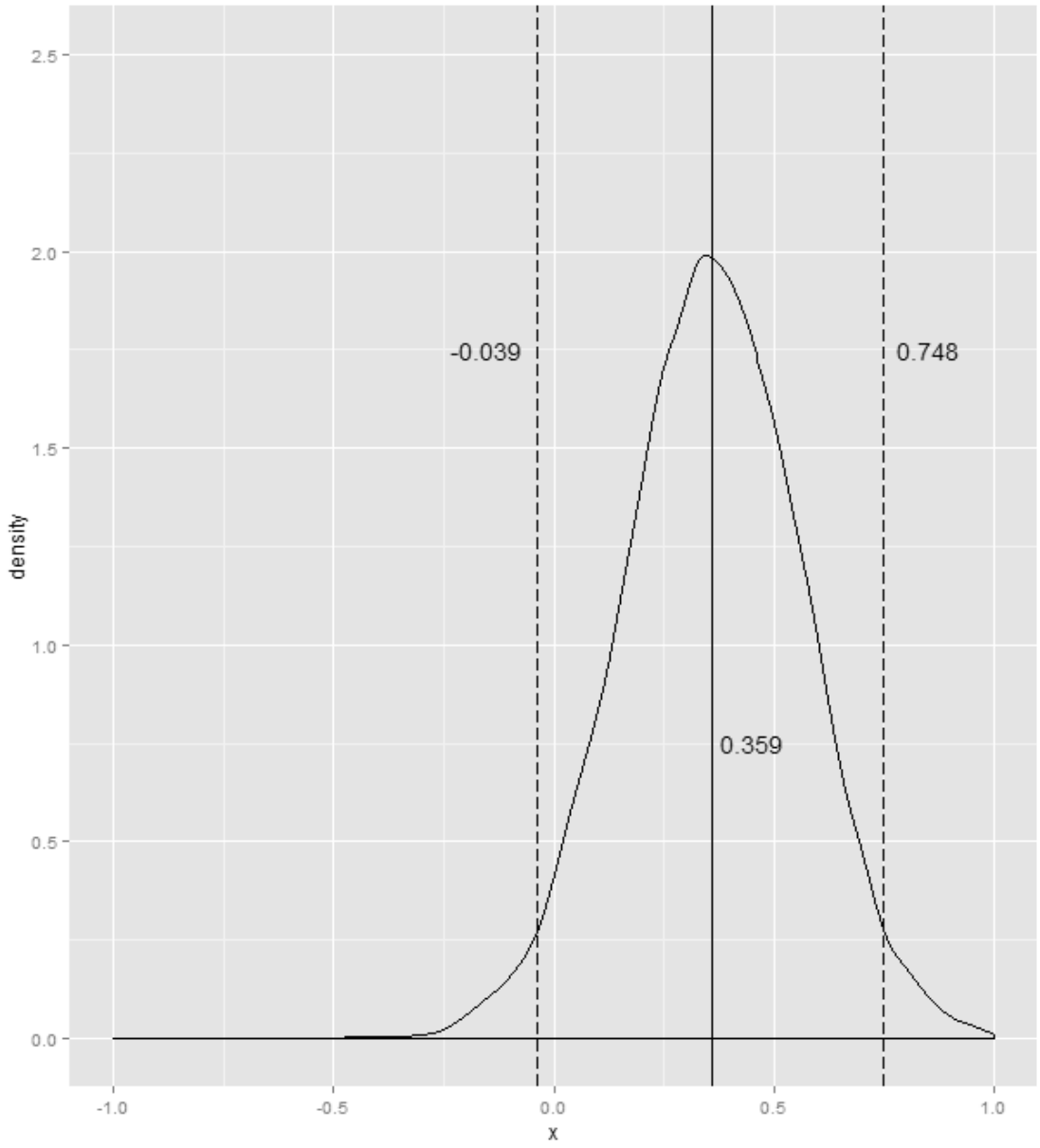
Appendix B: Bootstrap Analysis of Phase 2

A bootstrapping analysis was conducted to determine if the marginally significant trends in performance score were likely the result of insufficient statistical power. Following procedures by Larget, (2014), R was used simulate 10,000 resamples, sampling with replacement from the original data set for each of the visualizable measures. Means were calculated for the experimental and the control group for each resample and a difference between the scores was calculated for each resample. Figure G.1 displays density functions for each of the measures. A 95% confidence interval was calculated for each distribution and is indicated by dashed lines. The original difference between the group means is indicated by the solid line.

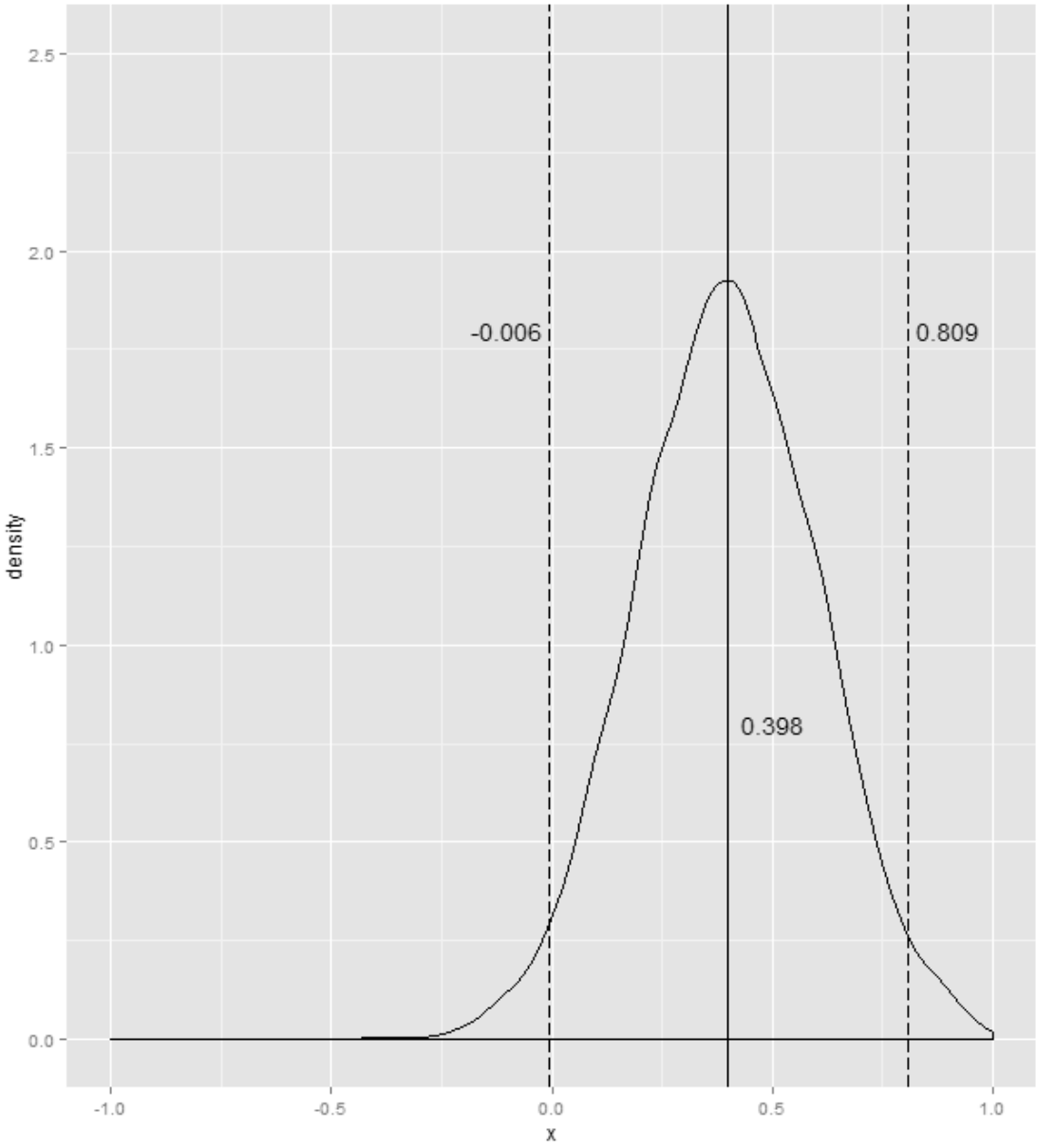
As expected, given the observed marginally significant trends in the raw data all estimated confidence intervals included zero. However, density functions indicated that the overlap for biased graphs (0.039), ratio bias (0.006), and sunk cost (0.026) were less than 0.05 strongly suggesting that the observed differences are likely to be robust and significant when replicated with larger samples. In contrast, LyingGL (0.389) and resistance to framing (0.209) fell below 0.4 suggesting these effects are not likely to be robust or reliable in other larger samples with similar sample characteristics. It is an empirical question the degree to which all these effects may be robust when sampling members of the general population instead of relatively mathematically skilled, well-educated Michigan Tech students. Theoretically, given normal statistical issues and reduction of power that emerges when there is restriction of range, I speculate that real effects are underestimated in the current sample. More research is needed to test this

assumption with more diverse and representative samples. Nevertheless, bootstrap simulations suggest that any observed improvement in one's ability to lie with graphs following similar training protocols is likely to be trivial, if significant. Theoretically, all other effects seem likely to approach one standard deviation (i.e., approach the aggregate overall effect size of Visualizable decision task).

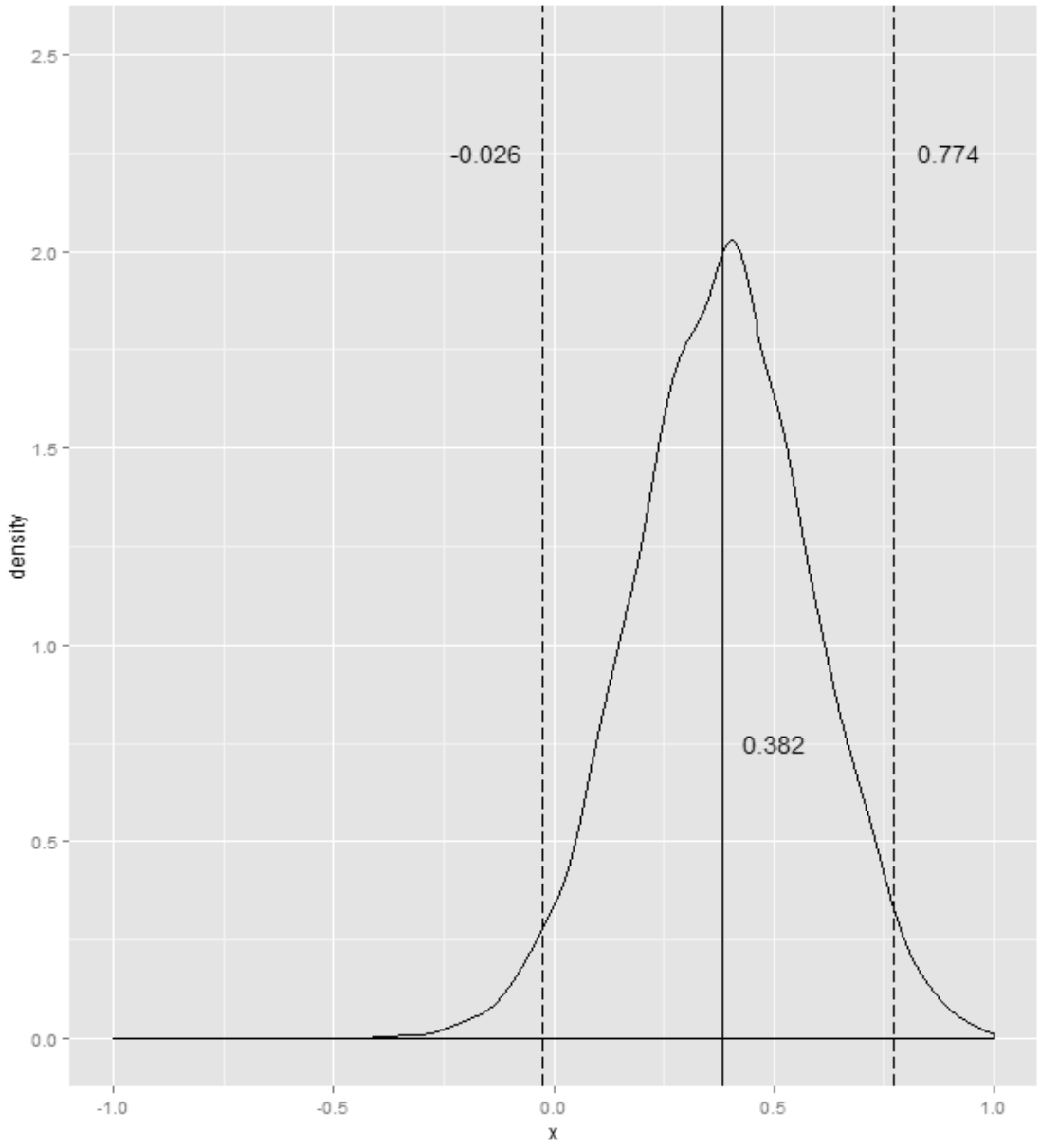
A. Biased Graphs



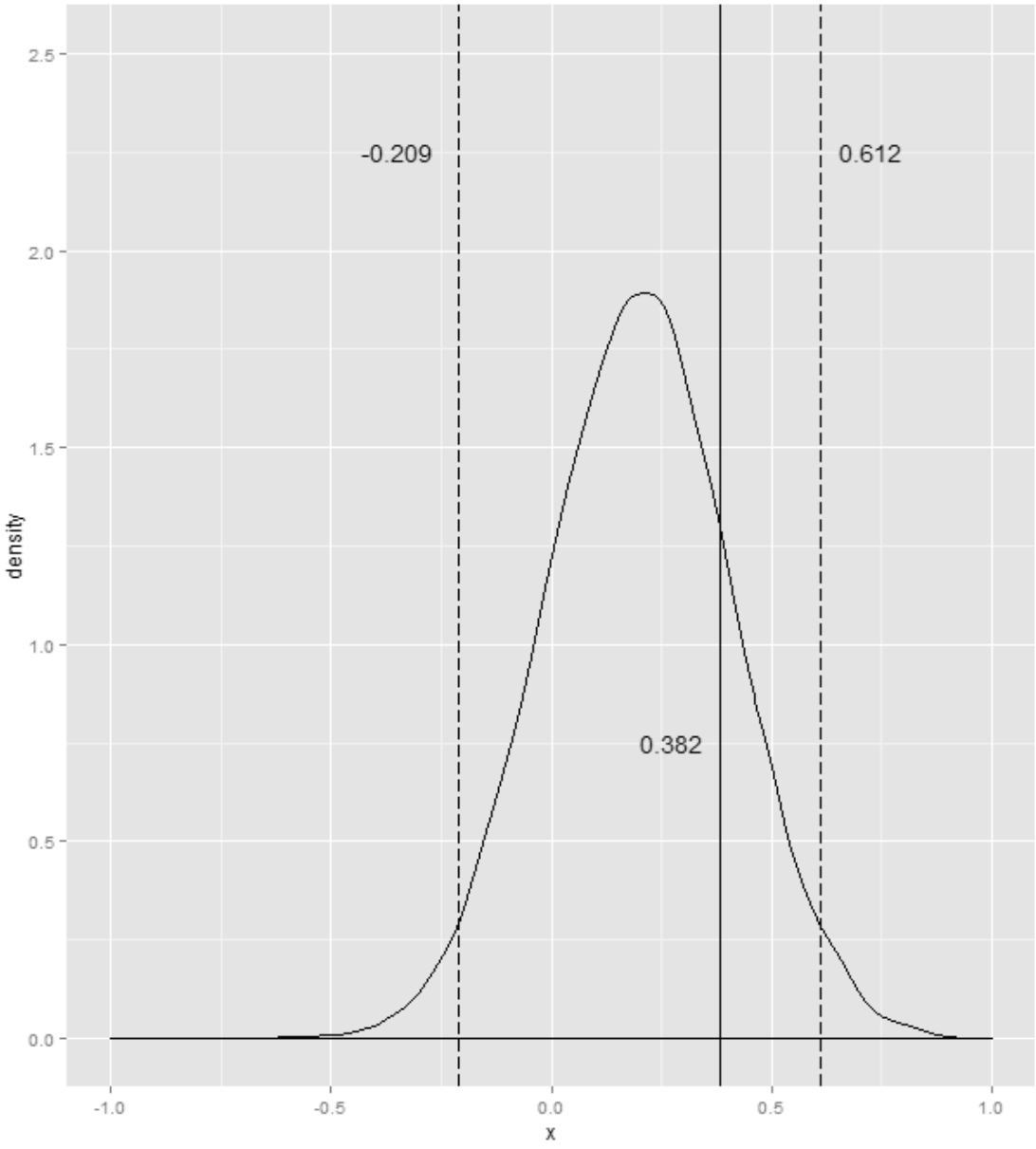
B. Ratio Bias



C. Sunk Cost



D. Resistance to Framing



E. LyingGL

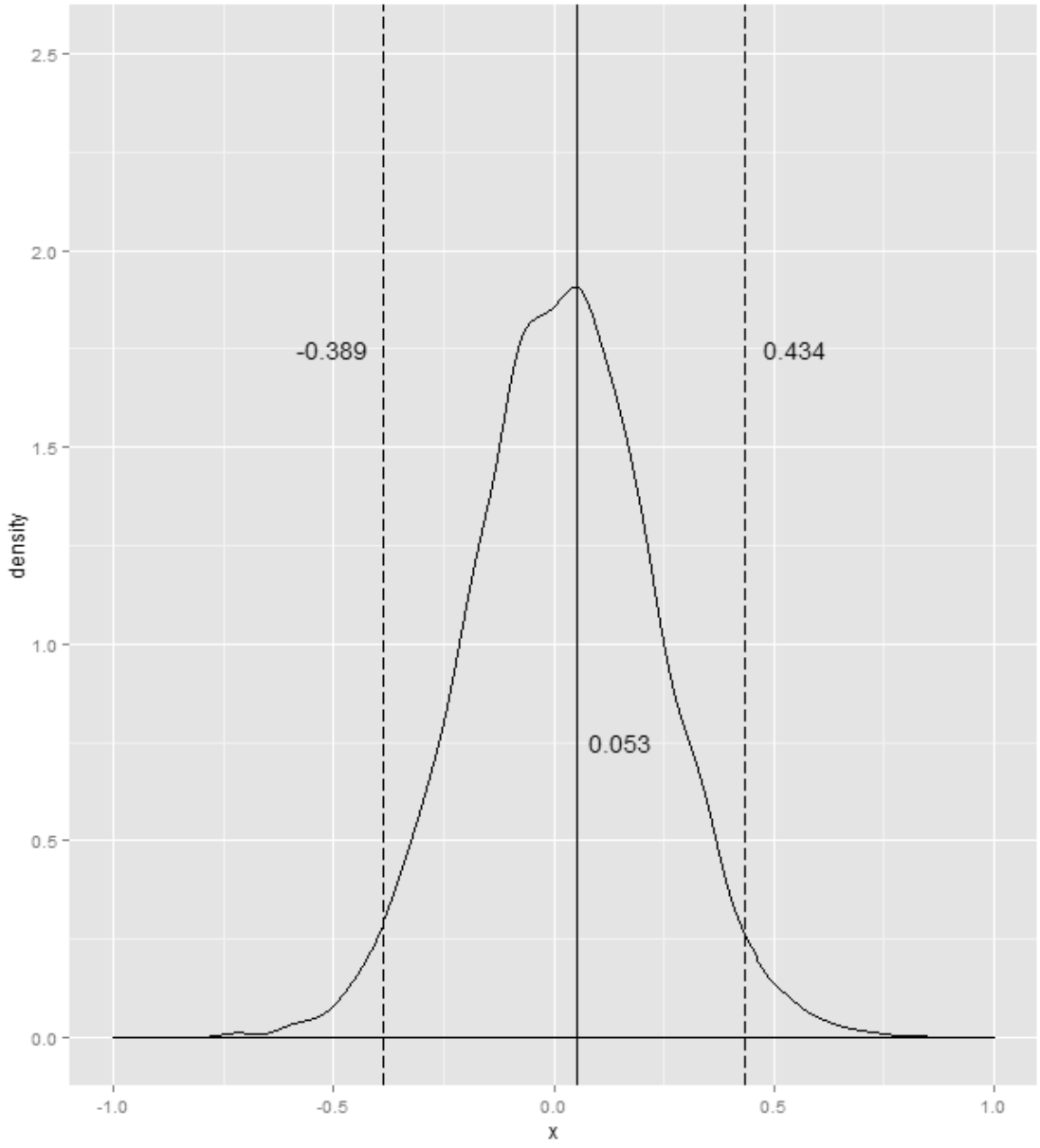


Figure B.1. Bootstrapping Density Functions of Experimental Minus Control Means. Dashed line indicate the end points of the 95% confidence interval and the solid line indicates the original mean difference.