# Impact of uncertainty visualizations on multi-criteria decision-making

by

**Amanda K. Newendorp**

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Majors: Industrial Engineering & Human-Computer Interaction

Program of Study Committee:
Stephen B. Gilbert, Major Professor
Michael C. Dorneich
Diane T. Rover

Iowa State University
Ames, Iowa
2024

# DEDICATION

This thesis is dedicated to my spouse, Tyler Owen, and my daughters, Audrey, Lauren, and Olivia, who have unconditionally supported my academic career over the last two years.

## TABLE OF CONTENTS

## LIST OF FIGURES

Page

# LIST OF TABLES

Page

# NOMENCLATURE

| | |
|---|---|
| AI | Artificial Intelligence |
| COA | Course of Action |
| HAT | Human-Agent Teaming |
| HCI | Human-Computer Interaction |
| ML | Machine Learning |
| MCDA | Multi-Criteria Decision Analysis |
| XAI | Explainable AI |

## ACKNOWLEDGMENTS

This thesis is the result of many hours spent thinking about different research proposals, but more importantly, of many hours spent discussing different research proposals. Thank you to my advisor, Stephen Gilbert, for sticking with me through months of indecision and for frequently reminding me that not everything needs to be perfect. I am deeply appreciative of all the insights, suggestions, and conversations you have shared with me. I would also like to express my appreciation to my committee members, Michael Dorneich and Diane Rover, for providing input and making time to support my research.

I want to thank my family, without whom I could not have done completed this thesis or degree. My spouse, Tyler Owen, has been nothing but encouraging from the moment I shared my idea of going back to school for another degree. Through many late nights and moments of self-doubt, he has been unwavering in his belief that I am doing the right thing and making a difference in the world. Thank you for everything.

To my daughters, Audrey, Lauren, and Olivia: I hope that one day you too will follow your dreams and choose the path that feels right. Thank you for motivating me to be my best self and welcoming me home from every trip to Ames with smiles, hugs, and giggles.

I want to express my deep appreciation for my parents, Bruce and Jan Newendorp, and brother, Brandon Newendorp, for helping me become the person I am today. They instilled a deep curiosity and love of learning that has been one of my guiding stars, and they have stood by my side through everything.

Finally, I want to thank my friends who encouraged me to take the leap of faith to return to school. This thesis might not exist without your boosts of encouragement and nudges in this direction.

## ABSTRACT

This study expands on previous research on visualizing uncertainty in decision support tools by exploring the effects of two factors. First, it explores how different methods of displaying uncertainty and automated recommendations affect the decision. Second, it investigates how these factors are influenced by individual differences.

In this within-subjects study, 346 participants were assigned to one of six conditions, each with a different version of a multicriteria information dashboard. Participants selected the optimal product over 12 timed trials and answered questions about their decision-making process.

Results showed that dashboards with detailed decision recommendations were associated with the highest trust, reliance, and decision accuracy. Including uncertainty on charts reduced decision confidence, accuracy, and speed, but also interacted with gender and recommendation style to impact how frequently participants sought more information to make decisions. Women reported lower decision confidence and more frequently sought additional information to make decisions. These results inform designers of data dashboards about how to promote more transparent and inclusive decision support tools that convey uncertainty in the input data.

# CHAPTER 1. BACKGROUND

## Introduction

Decision support tools such as data dashboards help humans make decisions, sometimes using artificial intelligence (AI) to recommend a specific course of action. Examples can be seen in multiple domains, such as medical diagnoses (e.g., Cai et al., 2019; Yang et al., 2019), military mission planning (e.g., Loft et al., 2023; Mercado et al., 2016), and operations research (Gupta et al., 2022). However, these AI-generated models are often based on probabilistic data, which means there is some uncertainty in the recommendations. This leads to a usability challenge; information uncertainty is notoriously difficult to depict in visualizations and for humans to interpret (Brodlie et al., 2012).

The challenge of clearly conveying uncertainty in data and recommendations presents an opportunity for the human-computer interaction (HCI), machine learning (ML), and AI communities to collaborate (Kaur et al., 2020; Tomsett et al., 2020) and develop explainable AI (XAI) techniques. XAI research seeks to create interfaces for AI and ML systems that humans can easily trust and interpret, giving them confidence to accept or reject the model output (Barredo Arrieta et al., 2020). One approach is to provide information to users about both the model output confidence and the sources of uncertainty in the input data. For example, uncertainty source information may enable a user to determine whether low model output confidence is the result of epistemic uncertainty (unavailable data) (Barredo Arrieta et al., 2020).

This research focuses on using data dashboards and recommendations to help humans make decisions with uncertain information, providing users with both a visualization of the input data and a decision recommendation. The visualizations in this study draw from research about uncertainty visualizations, which explores different methods of displaying uncertainty (Eberhard, 2021; Wilke, 2019) and encouraging deeper consideration of the data (Buçinca et al., 2021; Pfaff et al., 2013). The recommendation styles in this study draw from research about AI decision support, which has compared interfaces for displaying recommendations but has often simplified (Mercado et al., 2016; Stowers et al., 2020; Wohleber et al., 2023) or ignored (Behymer et al., 2015; Loft et al., 2023) uncertainty in the data. Few studies have explored decision support dashboard interfaces that include both a granular view of uncertainty in the input data and decision recommendations. This gap leads to the first research question. RQ1: When there is uncertainty in the input data, how does the method of displaying decision comparisons and recommendations on a data dashboard affect decision-making?

Individual differences in decision-making are also underexplored, both in research about uncertainty visualizations (Eberhard, 2021) and human-autonomy interaction (O'Neill et al., 2022; Wynne & Lyons, 2018). While individual differences can include a wide range of knowledge, skills, and abilities, this research focuses on gender-trending characteristics. Gender-trending characteristics are attributes that are disproportionately represented in women or men, but not unique to women or men. Gender-trending characteristics are central to GenderMag, a process that has been proven to identify gender-inclusivity software issues (Burnett et al., 2016). The five gender-trending characteristics in GenderMag, referred to as cognitive facets, are relevant to interactions with software and technology. Some of these facets (e.g., attitudes toward risk) are also associated with decision-making approaches, which can vary with gender (Apesteguia et al., 2012; Atkinson et al., 2003; Fehr-Duda et al., 2006; Jianakoplos & Bernasek, 1998). This topic leads to the second research question. RQ2: How does decision-making vary with gender-trending characteristics when using a data dashboard that includes decision comparisons and recommendations with uncertainty in the input information? In this research, participants were asked to make business decisions using versions of a decision support tool that differed in their depictions of uncertainty. Given these two research questions, it is important to understand previous research about decision support tools and dashboards, methods of visualizing uncertainty, and gender-trending individual differences.

## Decision Support Tools and Uncertainty

The concept of decision support tools was introduced in the early 1970s, and these tools have become increasingly capable over time (Shim et al., 2002). While earlier systems were based on structured models, such as decision support trees, many newer systems are based on ML models (Barredo Arrieta et al., 2020). Some decision support tools use multi-criteria decision analysis (MCDA), a process of making decisions by comparing differently weighted criteria among multiple alternatives to select the optimal option (Pelissari et al., 2018; Tsakalerou et al., 2022). While multiple approaches to performing MCDA exist, such as Pugh Matrix Analysis (PMM) (Pugh, 1981), Data Envelopment Analysis (DEA) (Cook & Seiford, 2009),  and Analytic Hierarchy Process (AHP) (Vaidya & Kumar, 2006), they share the underlying goal of making an optimized decision when there are multiple decision criteria to consider. However, although input data often includes uncertainty, traditional MCDA methods typically require exact input values (Pelissari et al., 2018). This uncertainty in MCDA input data can be a result of ambiguity (e.g., subjective scoring provided by humans), randomness (also referred to as stochasticity or aleatoric uncertainty), or partial information (also referred to as epistemic uncertainty) (Figure 1-1). Multiple methods exist to manage this uncertainty (H. Chen et al., 2011; Hyde et al., 2003; Pelissari et al., 2018; Tsakalerou et al., 2022), but they focus on ways to account for uncertainty in the MCDA algorithms and less on creating a usable interface with results, which is the focus of this research.

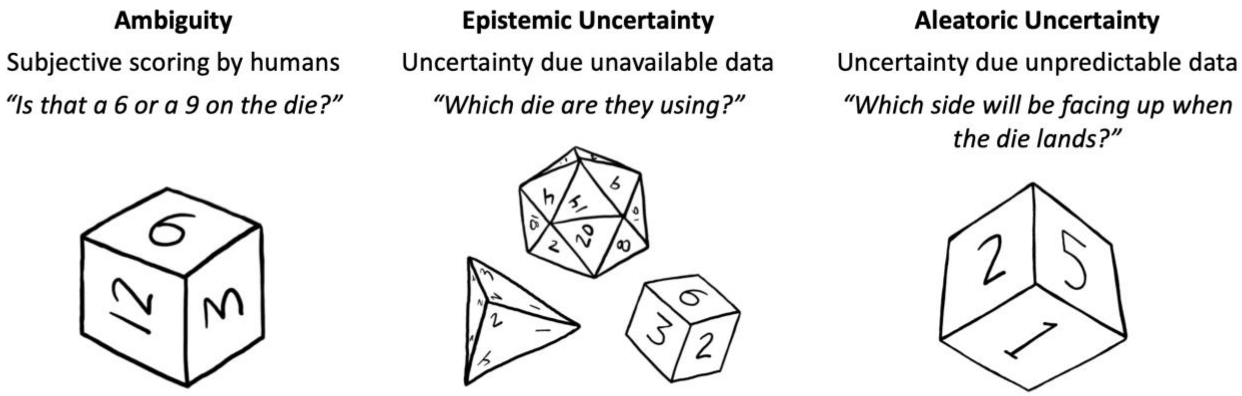

Ambiguity
Subjective scoring by humans
"Is that a 6 or a 9 on the die?"
Epistemic Uncertainty
Uncertainty due unavailable data
"Which die are they using?"
Aleatoric Uncertainty
Uncertainty due unpredictable data
"Which side will be facing up when the die lands?"

Figure 1-1: Uncertainty in input data can be due to ambiguity (human subjectiveness), epistemic uncertainty (unavailable data), and aleatoric uncertainty (unpredictable data).

Other decision support tools use AI and ML to make recommendations. In AI model outputs, two types of uncertainty are considered: aleatoric and epistemic uncertainty (Tomsett et al., 2020). Helping users understand what the system does and does not know, including the sources of uncertainty, can result in more interpretable, transparent decision support systems, and help users develop trust in the decision support system (Barredo Arrieta et al., 2020; Tomsett et al., 2020). This guidance to be more transparent about sources of uncertainty in AI model outputs aligns with research about human-autonomy teaming (HAT), in which appropriately calibrated trust is a result of automation transparency and reliability (J. Y. C. Chen et al., 2014). Transparency refers to the extent to which automation describes its own behavior, including its perception and comprehension of the current situation, as well as its prediction of the future situation (J. Y. C. Chen et al., 2014, 2018).

Recommendations generated by AI are also subject to both automation bias and algorithm aversion (Tomsett et al., 2020). Automation bias occurs when humans trust and rely upon automation instead of using their own judgment or critical thinking skills, which could be a result of choosing the least cognitively demanding action, believing automation to have superior recommendations, or neglecting the automated task in favor of other tasks (Parasuraman &

Manzey, 2010). Algorithm aversion occurs when humans trust and rely upon human recommendations over algorithm recommendations, even when they see the algorithm outperform the human (Dietvorst et al., 2015). These phenomena are important to consider, since they may interact with desirable constructs like trust and influence decision-making.

Human biases for or against AI-generated recommendations are also coupled with human biases about uncertainty. Among these are the representativeness bias (basing decisions on how well the data represent an existing mental model), the availability bias (making judgments about frequency or probability based on how readily examples come to mind), and the anchoring bias (making judgments based on the initial value or starting point) (Tversky & Kahneman, 1974). Decision support algorithms are not subject to these biases and, if trusted, can help users avoid making errors based on these biases.

Previous work in military HAT research has brought MCDA, information uncertainty, and AI decision recommendations together to help humans compare potential courses of action (COAs) in a military mission. Presenting these COAs to the user has included information visualizations, but uncertainty has typically been excluded (Bartik et al., 2019; Behymer et al., 2015) or simplified (Mercado et al., 2016; Stowers et al., 2020; Wohleber et al., 2023).

## Information Visualizations with Uncertainty

Depicting uncertainty in information visualizations is a significant challenge (Eberhard, 2021; Padilla et al., 2015; Wilke, 2019), adding another dimension and more visual noise to charts that often are already information-dense (Brodlie et al.,). In their review of research about uncertainty visualizations, Brodlie et al. (2012) elaborated on these challenges: uncertainty adds complexity to visualizations, and uncertainty can be depicted in different ways (e.g., probability density functions vs. bounded data) with input from multiple disciplines (e.g., domain researchers, statisticians, graphic designers, etc.).

Data uncertainty has been likened to data quality, in that both tell a user how much should use the data in their decision-making (MacEachren, 1992).  An effective and understandable depiction of uncertainty helps people make better decisions; it can help calibrate people on when to seek additional information (Dong & Hayes, 2012), encourage people to slow down and engage in the deeper thinking (Kaur et al., 2020), and make the information more clear (Aerts et al., 2003). However, people commonly struggle with correctly interpreting uncertainty in information visualizations.

When uncertainty is included in a visualization, people often misinterpret data points as representing precise values (Wilke, 2019). They employ strategies such as ignoring uncertainty (Padilla et al., 2015), favoring the option with the least uncertainty (Padilla et al., 2015), or transforming the data to fit their domain knowledge and mental model of the decision (Dilla & Steinbart, 2005). These strategies are not always consistent with making the optimal decision. Uncertainty can also be interpreted as being disproportionately important, since visualizations often place the highest emphasis on the greatest sources of uncertainty (e.g., a large error bar’s width may be perceived as the most important feature in a visualization) (Brodlie et al., 2012).

Although including uncertainty in information visualizations is associated with better decision-making, it is sometimes associated with reduced decision confidence (Eberhard, 2021). However, excluding information about uncertainty can create an illusion of confidence that is detrimental to decision-making (Brodlie et al., 2012). The prevalence of probabilistic data in decision recommendations, along with the challenges in displaying uncertainty emphasize the importance of continuing to explore methods of displaying uncertainty in HCI research (Aerts et al., 2003; MacEachren et al., 2012; Tomsett et al., 2020).

## Individual Differences and GenderMag

The GenderMag process has typically been used to identify gender-inclusivity bugs in software, but it has also been proposed as a way to systematically address biases against gender-trending approaches to using an AI decision support tool (Fern et al., 2024). Each of the five cognitive facets in GenderMag could play a role in the usage of AI decision support tools. The “attitudes toward risk” facet describes the trend of women tending to be less tolerant of risk than men, which is reflected in research about gender and risk tolerance in decision-making (Apesteguia et al., 2012; Atkinson et al., 2003; Fehr-Duda et al., 2006; Jianakoplos & Bernasek, 1998). In decision making, lower risk tolerance could also be associate with a higher likelihood of seeking more information before making a decision. Self-efficacy, or one’s confidence in their own ability to achieve a goal, with technology tends to be lower in women than in men. With self-efficacy’s close ties to confidence (Cramer et al., 2009), and previous work showing gender differences in confidence judgments (Lundeberg et al., 1994), the self-efficacy facet is likely related to decision confidence.

A third GenderMag facet, information processing style, describes two approaches to interpreting information and making judgments. The selective information processing style (more common in men) is characterized by making efficient judgments with key pieces of information, while the comprehensive information processing style (more common in women) is characterized by seeking and considering all available information before making a judgment (Burnett et al., 2016; Meyers-Levy, 1988). In decision-making, the comprehensive information processing style may be associated with higher rates of seeking more information to make a decision and slower decision response times, while the selective information processing style maybe associated with higher reliance on recommendations and faster decision response times.

The fourth and fifth GenderMag facets, motivations toward technology and learning by tinkering or by processing, are expected to be less relevant to decision-making. However, since these facets describe how readily a user will explore new technology (Burnett et al., 2016), they could influence decision-making when the data are presented in an interactive software interface, such as those proposed to make ML results (Kaur et al., 2020) and uncertainty visualizations (Aerts et al., 2003) more interpretable.

## Structure of Thesis

Chapter 1 describes the background of decision support tools, uncertainty in information visualizations, and gender-trending characteristics. Chapter 2 elaborates on related research about the topics discussed in Chapter 1. Chapter 3 explains the methods used in this experiment, and Chapter 4 presents the results of this experiment. Chapter 5 contains a discussion of the results, future work, and limitations.

# CHAPTER 2. RELATED WORK

To explore the two research questions, it is worthwhile to first understand relevant previous research. The overall goal of this literature review is to better understand existing research gaps and key findings about decision recommendations, uncertainty visualizations, and gender-trending characteristics. The results of this review informed decisions about the design of interface elements experiment methods (e.g., dependent variables, additional factors, like time pressure), and the formation of hypotheses.

This section first describes related research about decision support tools and recommendations, including both desirable user interface elements and experiment design methods. Next, it reviews research about uncertainty visualizations, such as which design elements most effectively convey the presence and amount of uncertainty to a layperson. Finally, this section describes related work about gender-trending characteristics, including their relevance to HCI and decision-making.

## Decision Support Tools and Recommendations

In some HAT research, decision support tools have combined MCDA with AI-generated recommendations, (e.g., for the formation and comparison of courses of action (COAs) in a military mission-planning context). Researchers in this domain are particularly interested in trust and reliance on recommendations, and the amount of information provided with a recommendation (transparency) is often included as an independent variable (Mercado et al., 2016; Stowers et al., 2020; Wohleber et al., 2023; Yuan et al., 2022). However, while uncertainty is sometimes depicted in high transparency conditions, it is not consistently included with recommendations. This section reviews four aspects of previous HAT studies that are relevant to this research: 1) The relationship between reliability, transparency, and trust, 2) Visualizations of recommendations and uncertainty, 3) Quantity of decision options and visualizations, and 4) Time pressure.

**Recommendation Reliability, Transparency, and Trust**

A typical study about AI-supported MCDA interfaces includes decision recommendations, often with a visualization to compare options across multiple decision factors. Research questions often include recommendation compliance (e.g., the extent to which participants' decisions align with the recommendations, and the extent to which participants correctly reject bad recommendations). Thus, the experiments are often designed to include some sub-optimal recommendations.

Compliance with these recommendations is closely related to users' trust in the AI system, and high trust is associated with both high transparency and high reliability (J. Y. C. Chen et al., 2014; O'Neill et al., 2022). Thus, it is useful to understand related HAT research about varying levels transparency (i.e., the extent to which the system explained its recommendations) before designing a decision recommendation interface. While it has been suggested that higher transparency can be associated with increased user workload and longer response times (J. Y. C. Chen et al., 2014), a systematic literature review of 17 empirical HAT studies about transparency found mixed results. Of the six studies in this review that involved participants responding to agent recommendations, the higher transparency condition was associated with longer response times in one study, shorter response times in two studies, and no significant changes in one study (van de Merwe et al., 2024). Response times were not analyzed in the other two studies about agent recommendations, and workload was not significantly different between the transparency levels in any of these six studies (van de Merwe et al., 2024). Overall, this review found that higher agent transparency is desirable and has a positive effect on users' performance and situation awareness. This aligns with earlier research suggesting that higher transparency is preferred by users and associated with higher trust in the automation (J. Y. C. Chen et al., 2014, 2017; Mercado et al., 2016; Tatasciore et al., 2023). However, in one study a high transparency condition that included simple depictions of uncertainty was associated with reduced trust (Stowers et al., 2020). Given that transparency is an important aspect of recommendation interfaces, it is important to next understand what a good recommendation interface looks like.

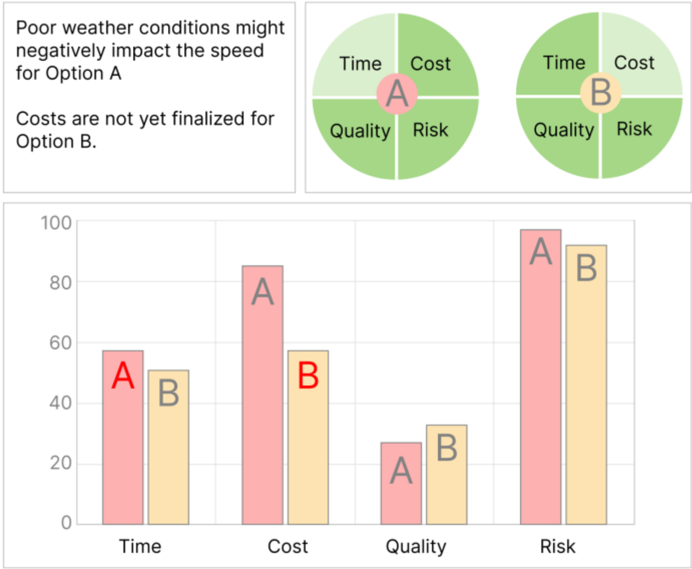
## Visualizations of Recommendations and Uncertainty

Transparency levels are often defined by the Situation Awareness-Based Agent Transparency (SAT) model, which defines three levels of agent transparency:

1. The agent's basic description of the current situation and its recommendation,
2. The agent describes its rationale,
3. The agent describes potential outcomes or future states. (J. Y. C. Chen et al., 2014, 2018).

However, other studies used the levels of transparency more generically to describe the amount of information provided by an agent, such as verbal and graphical explanations of the recommendation or some depiction of uncertainty. However, when uncertainty was depicted, it often took a binary form: either with or without a sentence to explain a source of uncertainty (e.g., "Poor weather conditions might negatively impact the speed for Option A") or opaque vs. translucent features in a graphic (Wohleber et al., 2023).

When an interface included textual, graphical, and iconographic information about recommendations, participants reported that they were most reliant on textual information in the low transparency condition, and most reliant on both the verbal and graphical information in the high transparency condition (Wohleber et al., 2023) (Figure 2-1). However, the graphical view only included information about the relative importance of each decision criteria in the low transparency condition, whereas it included detailed scoring information in the high transparency condition. Reliance on the iconographical information was low in both conditions. Overall, these results suggest that users appreciate both a graphical view of the options and a textual recommendation. While this study was not explicitly focused on depictions of uncertainty, it does provide insights about different methods of incorporating uncertainty in an information dashboard. Other studies have taken a similar approach to depicting uncertainty (Mercado et al., 2016; Stowers et al., 2020) or excluded uncertainty (Bartik et al., 2019; Behymer et al., 2015).



Poor weather conditions might negatively impact the speed for Option A

Costs are not yet finalized for Option B.

Figure 2-1: Previous studies about decision recommendation dashboards included simplified depictions of uncertainty, such as a textual description of uncertainty (top left), differently shaded wedges in an iconographic display (top right), and highlighting the labels of plans with uncertainty to be a different color in a graphical display (bottom). Adapted from Wohleber et al. (2023).

### Quantity of Decision Options and Visualizations

Next, it is useful to understand other aspects of recommendation dashboard design, such as the number of decision options, charts, and decision criteria to show at a time. The number of decision options provided to users has varied among previous studies, ranging from two (Loft et al., 2023; Mercado et al., 2016; Stowers et al., 2020; Wohleber et al., 2023) to three (Behymer et al., 2015) or more (Bartik et al., 2019; Buçinca et al., 2021). Showing one decision option at a time was not preferred because of the extra time it took to find more information, but preferences for showing four vs. eight decision options varied (Bartik et al., 2019). Some participants preferred the reduced clutter of four decision options, while others prefer the broader view offered with eight decision options (Bartik et al., 2019).

When asked to compare three decision options, participants preferred to see the decision options on single scatter plot, rather than seeing individual bar graphs or matrix charts for each decision option (Figure 2-2). However, the individual graphs and charts were preferred when participants were asked to assess a particular decision option (Behymer et al., 2015). Thus, the ideal form of presentation may depend on the user's task. While some MCDA problems include as many as seven decision criteria (Tsakalerou et al., 2022), many experiments feature three to five decision criteria (Bartik et al., 2019; Behymer et al., 2015; Loft et al., 2023; Mercado et al., 2016; Stowers et al., 2020; Wohleber et al., 2023). Among these papers, the most common approach was to include two decision options and four decision criteria on a single chart.

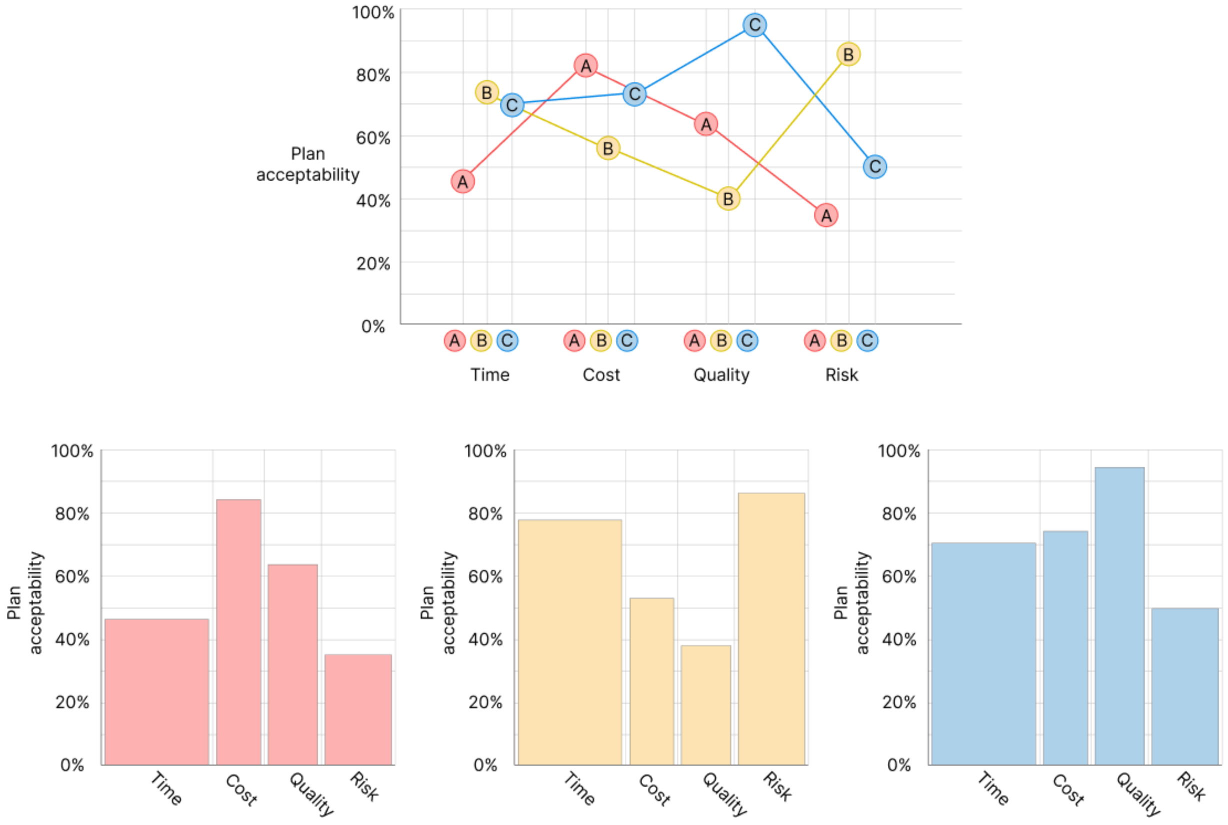


Figure 2-2: When comparing options, participants preferred to see multiple options on a single scatter plot (top). When answering questions about a specific option, participants preferred the individual bar graphs (bottom, bar width indicates criteria weights) or matrix charts (not shown). Adapted from Behymer et al., (2015).

## Time Pressure

Because lab experiments typically lack real-world time pressure and consequences, some experiments intentionally added additional demands on their participants in the form of secondary tasks (Allen et al., 2014; Tatasciore et al., 2023) or time pressure (Bartik et al., 2019; Loft et al., 2023; Mercado et al., 2016; Stowers et al., 2020). The approach to applying time pressure varied, ranging from scoring penalties for exceeding 30 seconds per decision (Bartik et al., 2019), to combining 45 second limits for each decision with a financial incentive for high performance (Loft et al., 2023), to setting a two-minute time limit for each mission (Mercado et al., 2016). While no studies in this review varied the amount of time pressure as an independent variable, there were recommendations to study time pressure in future work as a way to represent real-world pressure and consequences (Bartik et al., 2019; Wohleber et al., 2023).

## Information Visualizations with Uncertainty

Uncertainty is challenging to convey, but it is an important factor to consider when making decisions (Brodlie et al., 2012; Padilla et al., 2015; Pfaff et al., 2013). Much of the empirical work about uncertainty in information visualizations focuses on methods of displaying the uncertainty. Other research focuses on strategies to encourage deeper consideration of uncertainty. While uncertainty visualizations can take many forms, such as geospatial information (Aerts et al., 2003; MacEachren et al., 2005) and three-dimensional representations of scanned data (Zuk & Carpendale, 2006) this review focuses on uncertainty in graphs that can be used to compare options with an MCDA approach.

### Methods of Displaying Uncertainty

Broadly, there are two approaches to displaying uncertainty in MCDA scoring visualizations: with a probability distribution (e.g., depicting a bell curve or quartiles) or without a probability distribution (e.g., showing the minimum, maximum, and midpoint, also referred to as “strict uncertainty”) (Figure 2-3). Including a probability distribution can help people choose an optimal outcome in ambiguous comparisons (Correll & Gleicher, 2014; Dilla & Steinbart, 2005). However, people commonly struggle with interpreting probability distributions. There can be confusion about whether the distribution represents standard deviation, standard error, or a confidence interval (Dilla & Steinbart, 2005). Additionally, people may experience an anchoring bias by fixing on an arbitrary point in a probability distribution (Tversky & Kahneman, 1974). Visualizations depicting probability distributions are associated with less optimal decision-making than methods that do not depict probability distributions (Edwards et al., 2012) and with lower decision accuracy when combined with a concurrent cognitive loading task (Allen et al., 2014).

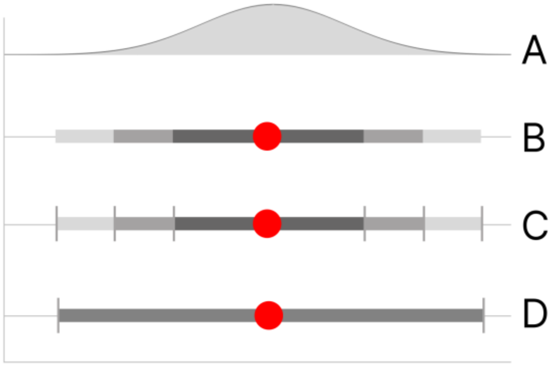
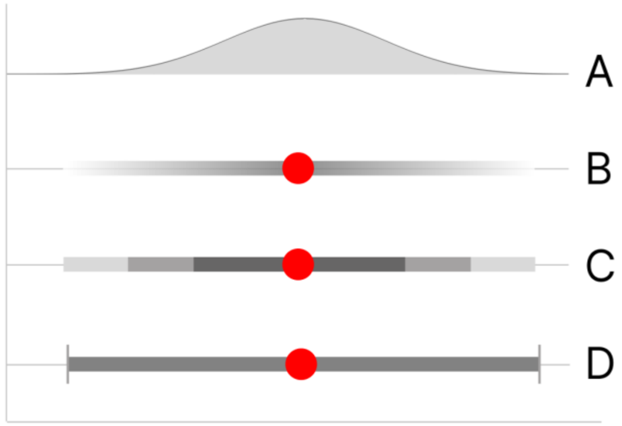


Figure 2-3: Uncertainty can be depicted with a probability distribution, such as by showing a bell curve (A) or quartiles (B, C). It can also be depicted without a probability distribution, showing the minimum, maximum, and midpoint (D). Adapted from Wilke (2019).

Trade-offs dominate the design choices in visualizations of uncertainty (Figure 2-4). Claude Wilke (2019) describes some of these trade-offs: Including confidence strips (color gradients) or bell curves conveys probability ranges, but discriminating between colors and interpreting the area under the curve is difficult. Graded error bars (showing multiple confidence intervals in different colors) help users understand that data points may fall outside the confidence interval, but they add visual clutter. End caps on error bars clearly communicate its end points, but they imply that all points fall within the range and add visual noise. Overall, Wilke recommends making trade-offs in information density and accuracy if it will result in a layperson being able to more intuitively understand the data.



| | Advantages | Disadvantages |
|---|---|---|
| A | Conveys the range of possible values, implies that data points can exist outside the confidence interval | Difficult to discriminate area under the curve or differences in color, adds visual noise<br>Adds visual noise |
| B | | |
| C | | |
| D | Clearly conveys end points of confidence interval | Adds visual noise, implies a fixed minimum and maximum value |

Figure 2-4: There are trade-offs in different chart styles, including how clearly they convey the distribution of data points, how accurately users can interpret the data, and how much visual noise they add (Wilke, 2019).

Selecting a graph type also comes with trade-offs; box plots or histograms clearly convey a range of uncertainty, but people are generally better at interpreting bar graphs (Eberhard, 2021). Graphical information is preferred over verbal or text depictions of uncertainty, and fuzziness or transparency are better at depicting uncertainty than changing colors (Eberhard, 2021).When using icons to depict uncertainty, fuzziness, location, value, arrangement, size, and transparency are more effective than color saturation at conveying uncertainty (MacEachren et al., 2012).

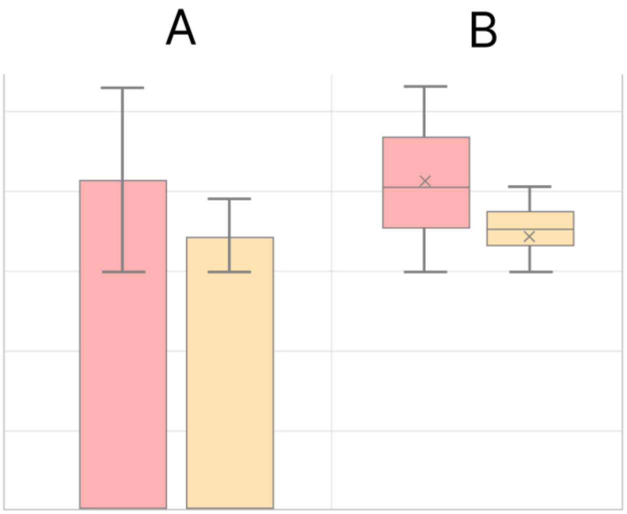


Figure 2-5: Bar charts (A) are more easily understood by a layperson, but boxplots (B) provide more information about probability distributions to convey the range of uncertainty (Eberhard, 2021).

This research suggests that for most users, a visualization that conveys the amount of strict uncertainty in a bar chart is the most interpretable solution. However, this comes with the trade-off of losing some of the data granularity if probability distributions are available.

### Strategies to Encourage Deeper Consideration of Uncertainty

Including uncertainty in information visualizations is sometimes associated with reduced decision confidence (Dong & Hayes, 2012), an effect that can be influenced by individual differences, such as experience with ML (Arshad et al., 2015). However, option awareness, which is defined by users having a strong awareness of decision options, factors, and potential outcomes, helps users make robust decisions (Pfaff et al., 2013). Researchers have also considered perceptions of uncertainty in the context of System 1 and System 2 thinking. System 1 thinking is fast and automatic, while System 2 thinking is slow and deliberative (Kahneman, 2013). Although rapid System 1 thinking can be desirable in some situations (e.g., under intense time pressure), in other situations it is worthwhile to encourage users to think more deeply before making a decision (e.g., for high-stakes decisions).

Although System 2 thinking may occur naturally when decision comparisons are highly ambiguous (Arshad et al., 2015), past research has also explored strategies to encourage the use of System 2 thinking in decision-making. Participants using an interface designed to encourage the use of interactive elements to seek additional information were more likely to correctly reject sub-optimal decision recommendations than participants in a static control condition (Buçinca et al., 2021). Including uncertainty in an MCDA-based decision support tool helped users determine whether they had enough information to make a decision (Dong & Hayes, 2012).

With ML data, including interpretable depictions of not just the amount of uncertainty, but the source of uncertainty (epistemic vs. aleatoric), could encourage users to employ System 2 thinking (Kaur et al., 2020) and build trust in decision recommendations (Tomsett et al., 2020). However, although some decision support tools indicate the system's confidence in its recommendation (Buçinca et al., 2021), little HCI research to date has focused on depicting these two sources of information uncertainty (Kaur et al., 2020; Tomsett et al., 2020).

## Individual Differences in Decision Making

Individual differences also play a role in decision-making with information visualizations, but they are not yet fully understood (Eberhard, 2021). Training is very important, and people with greater domain expertise, numeracy, and cognitive capacity (e.g., working memory capacity) make better decisions with information visualizations (Eberhard, 2021). However, some studies point out that less knowledgeable users are disproportionately helped by having access to a good visualization (Eberhard, 2021). When information visualizations included uncertainty, ML researchers were more confident in their decision-making than non-ML researchers (Arshad et al., 2015).

In previous research about confidence judgements, individual differences have been shown to play a significant role in decision confidence. In one study, over-confidence and under-confidence in individuals (i.e., having a mismatch between reported decision confidence and actual decision accuracy) was consistent across decision domains (Klayman et al., 1999). In another study, participants' decision confidence was also consistent across tasks (Blais et al., 2005) and was not related to any of three cognitive style measures: the Need for Cognition (Cacioppo & Petty, 1982), Personal Need for Structure (Thompson et al., 2013), and Personal Fear of Invalidity (Thompson et al., 2013) scales. This research suggests that it is worthwhile to explore the effects of not just dashboard interface style, but also individual differences, on decision confidence.

There may also be gender-trending differences in decision-making. In research about financial decision-making, women achieved similar results but used different strategies than men (Atkinson et al., 2003; Fehr-Duda et al., 2006), and women tended to be more risk-averse than men (Jianakoplos & Bernasek, 1998; Lauriola & Levin, 2001; Powell & Ansic, 1997). Men and women in an undergraduate psychology class both displayed overconfidence in their decisions, but men were particularly overly confident when they had the incorrect answer (Lundeberg et al., 1994). This research suggests that it is worthwhile to continue studying individual differences in decision-making so that training and interfaces, such as data dashboards and decision support tools, can be more effectively utilized by all users.

## GenderMag and HCI

The GenderMag process, which has proven to be a useful method of identifying gender-inclusivity software bugs (Burnett et al., 2016), offers a lens through which gender differences in decision making can be studied (Fern et al., 2024).

The GenderMag process helps software developers test their designs for gender-inclusivity by evaluating them through the lenses of three diverse personas, which represent users with characteristics more prevalent in women (Abi), more prevalent in men (Tim), and a mix of characteristics (Pat). These characteristics are based on five cognitive facets: attitude toward risk, information processing style, self-efficacy, motivations, and learning by process or by tinkering (Table 2-1). The facets were identified to meet three requirements: 1) they must be relevant to interacting with technology and software, 2) they must be backed by strong research, and 3) they must be easily understood by software and design professionals, without requiring a background in psychology or gender research (Burnett et al., 2016). The research basis for each of the five facets is described below.

Table 2-1: GenderMag features three personas: Abi, Pat, and Tim. Abi's characteristics are more common in women, Tim's characteristics are more common in men, and Pat represents a midpoint between the two (Burnett et al., 2016).

| **GenderMag Personas and Facets** (Burnett et al., 2016) | | | |
|---|---|---|---|
| **Facets** | **Abi** | **Pat** | **Tim** |
| **Attitude Toward Risk** | More risk-averse | Somewhat risk-averse | More risk-tolerant |
| **Self-efficacy** | Less confident in their ability to learn and use new technology, blames themself for problems | Somewhat confident in their ability to learn and use new technology, sometimes tries to troubleshoot problems | Confident in their ability to learn to use new technology and software, blames problems on the developer |
| **Information Processing Style** | Comprehensive - prefers to collect more information before making decisions | Comprehensive - prefers to collect more information before making decisions | Selective – prefers to use select, key information to make decisions |
| **Motivations** | Primarily uses technology to accomplish a goal | Learns new technology when needed, but doesn't usually spend their free time doing this | Likes to explore technology with no specific goal in mind |
| **Learning by Process or by Tinkering** | Prefers to learn by following a process or tutorial | Prefers to learn by tinkering and exploring but may revert to known methods | Prefers to learn by tinkering and exploring |

## Attitude Toward Risk

Attitude toward risk is often listed as a key theme in research about gender differences (Meyers-Levy & Loken, 2015; Reeves, 2022). In finance, men are more risk-tolerant (Charness & Gneezy, 2012), while women CEOs are associated with lower corporate risk-taking (Faccio et al., 2016). A domain-specific survey of risk perceptions found that college-aged women were more risk-averse than men in financial, health/safety, recreational, and ethics domains, but less so in a social domain (Weber et al., 2002). However, their results suggested that risk perceptions were more closely tied to the domain-specific benefits and risks than to a general personal attitude toward risk. A similar trend was seen in a study about financial risk-taking, in which risk-tolerance was similar between men and women, but only if the women also reported high confidence in their task-related abilities (He et al., 2008).

In HCI and decision making, risk tolerance is important because it is closely related to uncertainty. In one study, software was designed to be more inclusive of risk-averse users by including a new “maybe” label. As participants reviewed a list of potential bugs in a spreadsheet using this software, they could label cells as “maybe” containing an issue and revisit those cells later, as opposed to immediately committing to a binary yes/no label (Grigoreanu et al., 2008). In decision-making, information uncertainty is a source of risk. This previous work suggests that high uncertainty may lead to different decision-making strategies, such as more deliberation or a greater desire to first gather more information in risk-averse people.

## Self-efficacy

Self-efficacy is defined by a person’s confidence in their ability to achieve a desired outcome; people are less likely to persevere with a challenging task when they have low self-efficacy (Bandura, 1977). In the GenderMag process, the self-efficacy facet refers more specifically to self-efficacy with computers and technology (Burnett et al., 2016). Women tend to have lower self-efficacy in STEM fields than men (Durndell & Haag, 2002; Wang & Yu, 2023).(Hartzel, 2003). In a debugging environment, women tended to report lower self-efficacy, but women with high self-efficacy had similar performance to men (Beckwith et al., 2005). Because self-efficacy is closely related to confidence in oneself (Cramer et al., 2009) and one's ability to achieve a desired goal (Bandura, 1977), the previous work in this review suggests that low self-efficacy may be associated with lower decision confidence and higher reliance on recommendations.

## Information Processing Style

Information processing style refers to the way a person processes new information and makes judgments (Burnett et al., 2016). This facet is based on the selectivity hypothesis, a theory that explains observed gender differences in information processing (Meyers-Levy, 1988). This theory suggests that women are more likely to use a comprehensive information processing style in which they systematically consider and integrate all available cues, potentially at the cost of reduced speed. Men are more likely to use a selective information processing style, focusing on more salient cues and heuristics, potentially coming at the cost of reduced accuracy.

In studies to test the selectivity hypothesis, women were more likely to use all available information from an advertisement to make a decision, while men were more likely to use more salient themes and schema from the same advertisement to make a decision (Meyers-Levy & Maheswaran, 1991; Meyers-Levy & Zhu, 2010; Noseworthy et al., 2011). Women were more likely than men to identify incongruent products and had worse performance on a parallel verbal processing task, even when they were asked to prioritize the verbal processing task (Noseworthy et al., 2011). In a study with congruent and incongruent fictional television segments, results suggested that women made greater use of incongruent information in their judgments than men, although women and men had equivalent recall of the incongruent information (Meyers-Levy & Maheswaran, 1991; Meyers-Levy & Sternthal, 1991). Altogether, this research suggests that there are gender-trending differences in how people process information when making judgments and raises questions about whether this phenomenon also occurs in other decision-making contexts. This previous work suggests that when using a decision support dashboard, selective information processors may rely more on highly salient features (e.g., recommendations) and make more rapid decisions, while comprehensive information processors may methodically consider all available information, resulting in slower decisions.

### Motivations

The motivations facet refers to a person’s motivations for using technology (Burnett et al., 2016). The Abi persona, which represents characteristics that are more common in women, uses technology as a means to accomplish a goal. The Tim persona, representing characteristics prevalent in men, uses technology for its own sake. This facet is based on research showing that girls are more likely to use electronics (Cassell, 2002; Kelleher, 2009) and construction materials (Hallström et al., 2015) to achieve a separate goal, such as creating objects for a game or to tell a story, while boys are more likely to treat act of building and exploring as the goal. At Carnegie Mellon, changes to the computer science curriculum including a new emphasis on the social impact and interdisciplinary of computer science work (i.e., emphasizing different motivations for studying computer science), was associated with an increase from 7% to 42% female enrollment in the computer science department (Fisher & Margolis, 2002). Similarly, Harvey Mudd College implemented changes to recruit more women to their program, including a new emphasis on the breadth of computer science applications. After making these changes, the percentage of women in their computer science program increased from 12 percent to about 40 percent in five years (Corbett & Hill, 2015). In decision making, the motivations facet may only be relevant in certain decision domains (e.g., deciding whether or not to adopt a new type of technology).

**Learning by Process or by Tinkering**

The fifth facet describes preferences for learning about new technology. Learning by process is characterized by waiting for instructions, closely following directions, and hesitating to explore unfamiliar features. Learning by tinkering is characterized by an open-ended approach and readily exploring the available tools and technology. Among industry professionals and software developers, men showed more interest in tinkering with new, advanced features and tools, while women showed a stronger preference for using existing tools and enthusiasm for adding help wizards (Burnett et al., 2010). In an undergraduate computer science summer program, 15.4% of female participants described themselves as tinkerers, while 70% of male participants described themselves as tinkerers (Krieger et al., 2015). The researchers noted that the female participants' disinclination for tinkering may be related to risk aversion and a concern for breaking things. This raises questions about how HCI principles could be used to develop interfaces that lower the risk of making mistakes, and whether this might help more users feel safe to explore and tinker with new technology. In the context of decision making, preferences to learn by process or by tinkering may be relevant to interactive decision support systems, like an information dashboard.

## Summary

This review of decision support tools, methods of visualizing uncertainty, and gender-trending characteristics highlights three research gaps. First, while research about decision support tools provides guidance on how to display decisions recommendations and associated data to users, most of these studies do not focus on decision-making with information uncertainty. This highlights an opportunity to study both A) different methods of displaying uncertainty data alongside recommendations, and B) the effect that this display of uncertainty has on decision-making. This research could then be used to improve decision support tools that use probabilistic data, such as those utilizing AI-generated models.

Second, in research about uncertainty visualizations, much is known about the trade-offs in creating visualizations that are both data-rich and interpretable, less is known about their usage in decision recommendation tools. While previous research often measures how accurately participants interpret the data and draw conclusions from a visualization, these studies typically do not include decision recommendations or measure constructs such as trust and reliance.

Third, previous research about gender-trending characteristics, such as attitude toward risk and information processing style, indicate that there may be gender-related differences in decision-making, reactions to uncertainty (or risk associated with a decision), and usage of decision support tools. However, papers about decision support and uncertainty visualization domains often do not disaggregate results by gender, so there is little data to support or discredit an association between gender-trending characteristics and decision-making. This research aims to address these three gaps by exploring the interactions between uncertainty visualizations, decision recommendation styles, and gender-trending characteristics.

# CHAPTER 3. METHODS

## Overall Approach

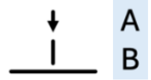
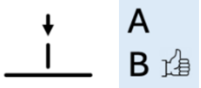
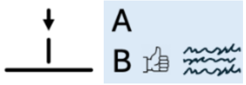
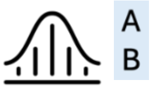
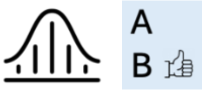
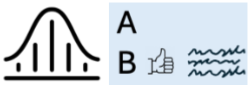
The purpose of this research is to explore the effects of two factors on decision-making when uncertainty is present in the input data. First, it explores how different methods of displaying decision options and recommendations affect the decision. Second, it investigates the impact of individual differences on the decision. Specifically, this research addressed two research questions.

- RQ1: When there is uncertainty in the input data, how does the method of displaying decision comparisons and recommendations on a data dashboard affect decision-making?
- RQ2: How does decision-making vary with gender-trending characteristics when using a data dashboard that includes decision comparisons and recommendations with uncertainty in the input data?

These research questions were addressed in the context of product evaluation using an MCDA method to compare two products. The goal of this research is to help designers create more interpretable, trustworthy, and inclusive interfaces for decision support tools that use probabilistic data.

To explore these questions, a 2x3 factorial between-subjects experiment was conducted to compare two types of MCDA charts and three types of decision recommendations (Table 3-1). The charts showed either a) a single, fixed score for each criterion (without uncertainty), or b) a range of possible scores for each criterion (with uncertainty). The recommendation style was either a) no recommendation, b) a basic recommendation, or c) a detailed recommendation that also describes the sources of uncertainty. In the detailed recommendation conditions, a third independent variable was considered: the predominant source of uncertainty (defined below).

Table 3-1: The experiment included six conditions, combining two styles of charts and three styles of recommendation for Product A vs. B.

| | No recommendation | Basic recommendation | Detailed recommendation |
|---|---|---|---|
| No uncertainty | 1  | 2  | 3  |
| Uncertainty | 4  | 5  | 6  |

The two chart styles (with and without uncertainty) were developed because including additional information, such as uncertainty, in a chart can help users make more thoughtful and sound decisions (Brodlie et al., 2012; Buçinca et al., 2021; Dong & Hayes, 2012; MacEachren, 1992; MacEachren et al., 2005). However, that enhanced decision making but may come at the cost of lower decision confidence (Arshad et al., 2015; Dong & Hayes, 2012), higher response times (Pfaff et al., 2013), and lower decision accuracy due to incorrectly interpreting data (Brodlie et al., 2012; Dilla & Steinbart, 2005; Padilla et al., 2015). This experiment uses charts with and without uncertainty data, so that these effects of uncertainty in visualizations could be compared in the context of MCDA-based data dashboards.

The two recommendation styles were developed to explore the effects of different levels of decision support and recommendation transparency, which is associated with higher performance (van de Merwe et al., 2024) and higher trust of recommendations (J. Y. C. Chen et al., 2014; O'Neill et al., 2022), performance (van de Merwe et al., 2024). The basic recommendation represented a low transparency design with only an overall score for the two options. The detailed recommendation represented a high transparency design, including the overall scores along with a description of the sources of uncertainty (the third independent variable). The sources of uncertainty were show as being predominantly due to either unavailable data (epistemic uncertainty) or unpredictable data (aleatoric uncertainty). The sources of uncertainty were included as a form of XAI to make the recommendations and probabilistic data more interpretable (Barredo Arrieta et al., 2020; Kaur et al., 2020; Tomsett et al., 2020).

Participants were assigned to one of the six conditions in this between-subjects study and presented with brief training and then a series of decisions to make based on the information displayed. To be consistent with MCDA-based recommendation studies reviewed in Chapter 2 and to manage the overall experiment complexity, each scenario included two decision options and four decision criteria. The entire experiment was conducted via a Qualtrics survey, and to facilitate the recruitment of many participants, the protocol was designed to take approximately fifteen minutes. Because of this time constraint, the six conditions were varied between subjects. This approach also enabled participants to spend their time using the visualizations to make decisions, rather than learning how to interpret multiple visualization styles.

## Participants

Participants were recruited through a mass email sent to all Iowa State University students and faculty. Anyone aged 18 or older was welcome to participate, and participants were given the opportunity to enter a drawing for one of three $25 eGift cards upon completion of the survey. IRB approval for this study was obtained with protocol 24-005 (APPENDIX A). The total participants responding to the survey was 773. After outlier removal (described below), 346 participants remained. Participant demographics are described in Table 3-2. As with all IRB-approved surveys, participants were not required to answer the demographics questions, leading to lower totals on some categories.

Table 3-2: Count of total participants by gender, age, college, and student classification demographics.

| Gender | | Age | | Academic College | | Student Type | |
|---|---|---|---|---|---|---|---|
| Female | 243 | 18-24 | 148 | Agriculture & Life Sciences | 72 | Undergraduate | 136 |
| Male | 84 | 25-34 | 72 | Business | 33 | Graduate | 59 |
| Other | 13 | 35-44 | 37 | Design | 19 | Non-Student | 144 |
| No response | 6 | 45-54 | 43 | Engineering | 53 | No response | 7 |
| | | 55-64 | 30 | Human Sciences | 53 | | |
| | | 65-74 | 10 | Liberal Arts & Sciences | 85 | | |
| | | No response | 3 | No response | 31 | | |

## Domain for Experiment: Business Decisions

The information visualizations in this experiment were based on a fictional business scenario, in which a company called MegaMart asked for help deciding which of two products (A or B) to start selling in their stores. In some conditions, participants were also told that MegaMart a new tool called ProdStar to make recommendations. See APPENDIX B for exact transcripts of the scenario training videos, which varied slightly by condition (see details below). MegaMart used four decision criteria to evaluate the two products: timing, cost, market demand, and product quality. As in a real business environment, the relative importance of these criteria could vary, so the four criteria were assigned an importance of low, medium, or high in each scenario. This decision domain (business and marketing) was selected for its neutrality (i.e., domains like AI-assisted medical diagnoses and military mission planning could require specialized knowledge). The MegaMart scenario and decision criteria were selected to be understandable to a layperson and not subject to strong gender stereotypes (e.g., a scenario that instead compared computer hardware specifications may be disproportionately challenging for some participants).

## Experiment Protocol

### Overview of Protocol

After reviewing the informed consent form, confirming they are 18 years of age or older, and agreeing to participate, participants were presented with a 2–3-minute training module. Then they were asked to answer questions based the information visualization in 12 trials. This number of trials was based on recording response times during pilot testing and then designing the overall protocol to take approximately 15 minutes. The upper time limit was intended to minimize attrition in participants since the survey was conducted online.

The order in which the 12 trials were presented was randomized. Next, participants were asked to answer questions about their experience with the data dashboard, such as their reliance on various elements of the information visualizations and an overall assessment of usability. Finally, the survey collected information about individual differences, including an assessment of the GenderMag facets. This protocol is shown in Figure 3-1.

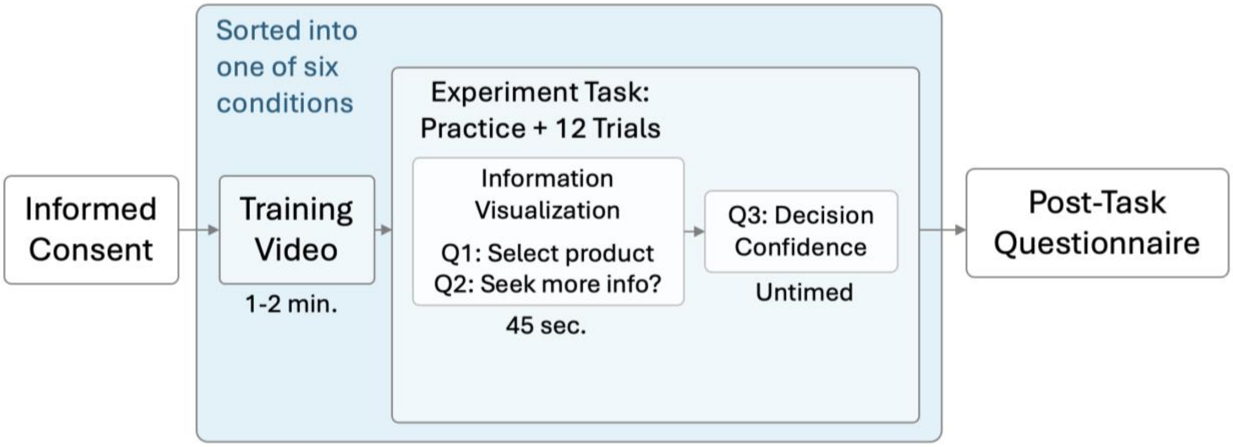


Figure 3-1: After signing the informed consent form and agreeing the participate, the experiment protocol includes a training, one practice product decision trial, 12 product decision trials, and a questionnaire.

## Training Module

The training module started with a video, which explained the MegaMart product decision scenario, described the charts, and introduced the ProdStar recommendation tool. The training video was unique for each of the six conditions, but a modular set of video segments was used to create each video (e.g., the training videos for Conditions 1-3 used the same video segment describing the chart style, and Conditions 4 and 6 used the same video segment describing the recommendation style). Transcripts and images of the training videos are included in APPENDIX B. After viewing the video, participants were presented with a practice version of the information visualization, with the same time constraint and the same three questions.

## Primary Experiment Task: Decision Scenarios

For each of the 12 trials, participants were asked three questions:

1. Which product should MegaMart choose to sell in their store? (A or B)
2. Should MegaMart seek additional information before making a final decision? (no or yes)
3. You selected Product [A,B]. How confident are you that you selected the optimal product? (1-5 Likert scale, ranging from “not confident at all” to “extremely confident”)

Questions 1 and 2 were presented on the same page as the bar chart and (if applicable) the recommendation. Question 1 was used to measure decision accuracy, and Question 2 was used to study participants’ usage of information about the sources of uncertainty in the detailed recommendation. In line with previous recommendations to simulate real-world task demands by applying time pressure in decision support research (Bartik et al., 2019; Loft et al., 2023; Mercado et al., 2016; Wohleber et al., 2023), and to manage the overall experiment timing, participants were allotted 45 seconds to make each decision. This timing was reviewed during the pilot testing, in which most participants typically answered the questions well within the time limit. After 45 seconds, the survey auto-advances to the next page, with or without both questions being answered. The question about decision confidence on the next page was not timed.

## Independent Variables

The six between-subjects conditions in this experiment comprised of combinations of two independent variables: chart style (two levels) and recommendation style (three levels). These variables were used to create six versions of a decision support dashboard (Figure 3-5). A third independent variable (predominant source of uncertainty) was considered in Conditions 3 and 6.

### Independent Variable #1: Chart Style

Based on the results of Eberhard's systematic review about data visualizations (Eberhard, 2021) and Wilke's guidance (Wilke, 2019), this experiment used bar charts to show the scores for each decision criterion to make it understandable to a layperson (Figure 3-2). The importance of each decision criterion was displayed on the right side of the chart. Participants in Conditions 1, 2, and 3 saw a version of the charts with no uncertainty, in which the bars indicated an absolute score for each criterion. Participants in Conditions 4, 5, and 6 saw a version of the charts that included uncertainty, in which the bars showed a range of possible scores. This range was described to participants as being a 95% confidence interval, so 95% of the time the actual value would be expected to fall in this range. The confidence interval was shown with an opaque bar, similar to a boxplot, and a partially transparent bar filled the space from zero to the lower end of the confidence interval. The partially transparent bar was included to provide a more consistent, bar chart-like experience to all participants. Because the scores were on a normalized scale from zero to one hundred percent acceptable, the partially transparent bar also gives some indication of the overall acceptability (e.g., if the lower bound of the confidence interval is at 75% acceptable, some participants may benefit from the partially transparent bar suggesting that there is some certainty that it is *at least* 75% acceptable). This approach could counter the phenomenon of error bar width (uncertainty) dominating over perceptions of certain data (Brodlie et al., 2012).

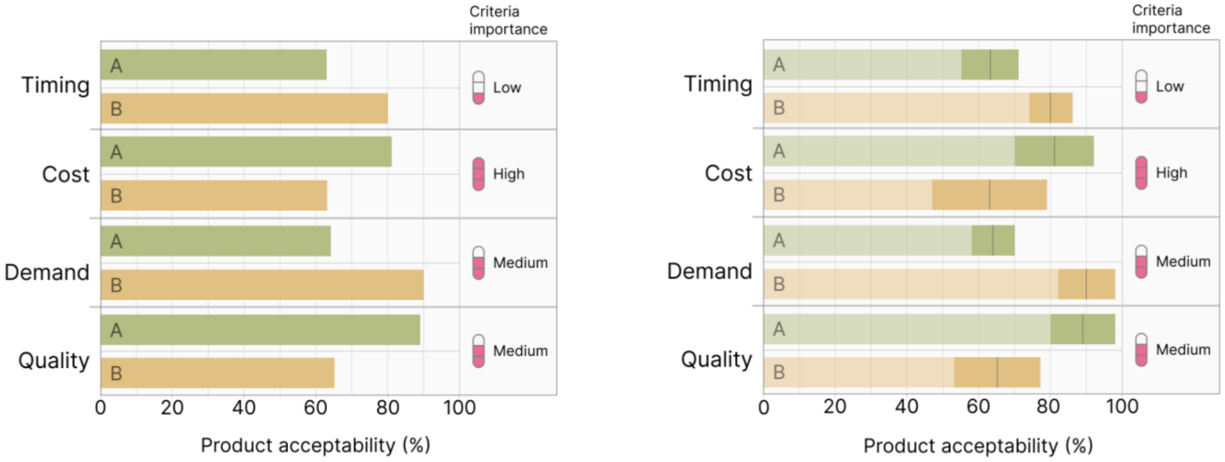
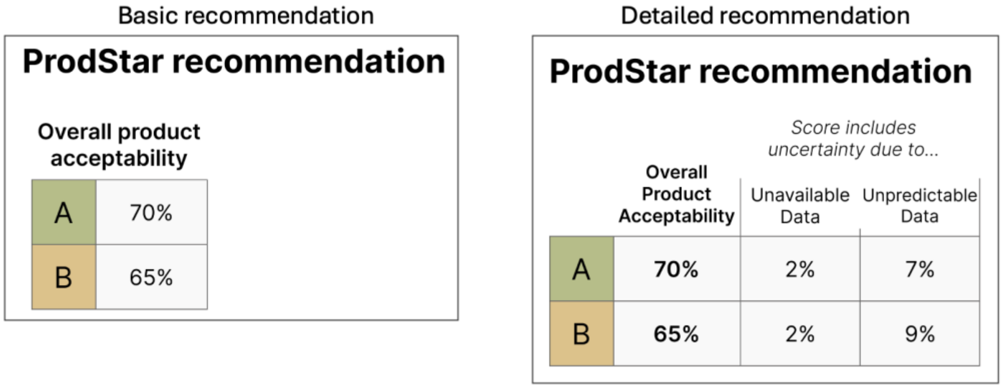


Figure 3-2: Bar charts were used to display scores for the four decision criteria and the importance of each criterion. Participants in Conditions 1, 2, and 3 saw a version with absolute scores (left), and participants in Conditions 4, 5, and 6 saw a version that included a range of uncertainty (right).

The colors on the charts were selected to be accessible to users with color blindness, and pattern fills were avoided to minimize visual noise on the information-dense charts. The colors were selected to be similarly salient to avoid the possibility of a vividly colored bar being perceived as more important than a pale colored bar.  The scores, amount of uncertainty, and relative criteria importance varied among the twelve scenarios. There were four decision criteria on the charts: timing, cost, demand, and quality.

### Independent Variable #2: Decision Recommendation Style

This experiment included three recommendation styles (Figure 3-3). In Conditions 1 and 4, the participants did not have access to a recommendation. Because the ProdStar recommendations always suggested the optimal product, Conditions 1 and 4 were used to establish baseline data about how often participants selected the optimal product given only the scoring data on the charts. In Conditions 2 and 5, participants were shown a basic recommendation, which compared the overall acceptability of the two products. In Conditions 3 and 6, participants were shown a detailed recommendation, which included the sources of uncertainty. The sources of uncertainty were described as being due to either unavailable data or unpredictable data. The sources of uncertainty were included in response to recommendations to explore interface designs that describe uncertainty (unavailable data) and aleatoric uncertainty (unpredictable data) (Kaur et al., 2020; Tomsett et al., 2020). The two styles of recommendation could also be described in HAT terms as low vs. high transparency. For both recommendation styles, the training video described ProdStar recommendations as being imperfect ("Like a weather prediction, ProdStar isn't always right") to give participants a sense of uncertainty in the data and to encourage further reflection on all elements of the ProdStar data dashboard.



Basic recommendation

**ProdStar recommendation**

| | Overall product acceptability |
|---|---|
| A | 70% |
| B | 65% |

Detailed recommendation

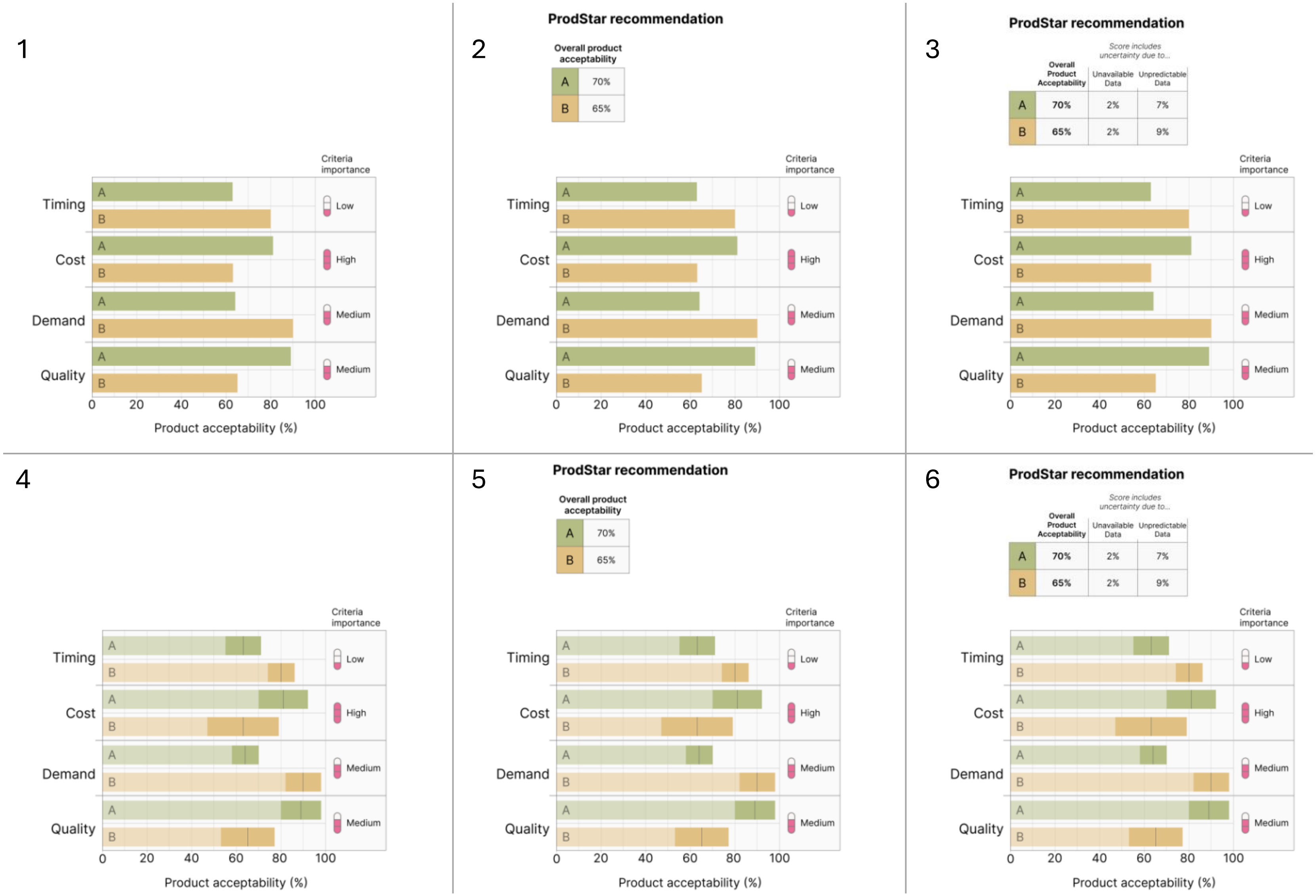
**ProdStar recommendation**

| | | *Score includes uncertainty due to...* | |
|---|---|---|---|
| | Overall Product Acceptability | Unavailable Data | Unpredictable Data |
| A | **70%** | 2% | 7% |
| B | **65%** | 2% | 9% |

Figure 3-3: Participants in Conditions 1 and 4 saw no decision recommendation, participants in Conditions 2 and 5 saw a basic recommendation (left), and participants in Conditions 3 and 6 saw a detailed recommendation with the sources of uncertainty (right).

**Independent Variable #3 (Conditions 3 and 6 only): Predominant Source of Uncertainty**

Conditions 3 and 6, which included a description of the source of uncertainty, also included a within-subjects independent variable. In 6 of the 12 trials, the uncertainty was shown to be predominantly due to unpredictable data. In the other six the trials, uncertainty was shown to be predominantly due to unavailable data. This approach enabled the researchers to explore whether the type of uncertainty influenced decision-making (e.g., whether participants determined that seeking additional information is only appropriate when uncertainty is due to unavailable data).



ProdStar recommendation

| Overall product acceptability | |
|---|---|
| A | 70% |
| B | 65% |

ProdStar recommendation

| | | Score includes uncertainty due to... | |
|---|---|---|---|
| | Overall Product Acceptability | Unavailable Data | Unpredictable Data |
| A | 70% | 2% | 7% |
| B | 65% | 2% | 9% |

ProdStar recommendation

| Overall product acceptability | |
|---|---|
| A | 70% |
| B | 65% |

ProdStar recommendation

| | | Score includes uncertainty due to... | |
|---|---|---|---|
| | Overall Product Acceptability | Unavailable Data | Unpredictable Data |
| A | 70% | 2% | 7% |
| B | 65% | 2% | 9% |

Figure 3-4: The six training conditions included all combinations of chart style (with or without uncertainty) and recommendation style (none, basic, or detailed)

## Scenario Creation and Recommendation Reliability

The scenarios in this experiment were designed such that one product was always 5% better than the other, and the recommendations were 100% reliable (i.e., ProdStar always recommended the optimal product). Products A and B were each the optimal choice for six of the twelve the scenarios. In initial pilot testing of earlier scenarios, product margins ranged from 10-20%, and the recommendations were reliable in 70% of the scenarios. However, this margin was large enough that pilot participants consistently detected the bad recommendations and easily made optimal decisions without the help of ProdStar. Pilot testing also indicated that response times were related to the size of the score margin. Therefore, new scenarios were created with the consistent, smaller margin of 5%. Because the optimal product was less obvious in the new scenarios, the Conditions 1 and 3 (the no recommendation conditions) were added to generate baseline data of decision-making without a recommendation.

To create the scenarios, a random number was generated to define the scores for each decision criterion (30-100), the percentage of uncertainty for each criterion (0-20), and the importance of each criterion (1-3, corresponding to low, medium, and high). The lower bound of 30 for decision score and upper bound of 20 for the percentage of uncertainty were chosen to represent a more believable scenario; products with very low acceptability scores or very high amounts of uncertainty would be unlikely contenders in a business decision. The 3-level scale of criteria importance was chosen to manage the overall data and visualization complexity, and because in a real-world scenario this would likely be an inexact, subjective ranking based on business priorities and managerial opinions. This approach typically resulted in similar scores for the two products, so the values were adjusted until the margin was exactly 5%.

**Overall Product Acceptability**

The overall product acceptability was calculated by summing the weighted scores and amount of uncertainty for each of the four criteria. This method is described by Equation 1.

Equation 1: Equation to calculate overall product acceptability.

$$Overall\ Product\ Acceptability = \sum_{i=1}^{4} S_i * \frac{W_i}{(W_1 + W_2 + W_3 + W_4)} * \left(\frac{1}{1 + U_i}\right)$$

Where:

$W$ = Criterion weight (1, 2, or 3)

$S$ = Criterion score (30-100%)

$U$ = Amount of uncertainty (0-20%)

$i$ = Decision criterion (1, 2, 3, or 4)

An example of product scores is shown in (table). For the timing criterion, Product A has a criterion weight of 1, a criterion score of 63%, and 8% uncertainty. This is used to calculate a weighted score for the timing criterion (Equation 2), as shown in Equation 3. The overall score is equal to the sum of the four weighted criterion scores (Equation 4), as shown in Equation 5. In these equations, percentages are represented as decimals (e.g., 63% is entered as 0.63).

Table 3-3: The input data for each product and criterion includes a weight, score, and amount of uncertainty. The weighted scores and weighted uncertainty were used to calculate the overall product score and the overall amount of uncertainty.

| | **Input Data** | | | | **Intermediate Calculations** | |
|---|---|---|---|---|---|---|
| | **Criterion** | **Weight (1=low, 3=high)** | **Score** | **Amount of Uncertainty** | **Weighted Score** | **Weighted Uncertainty** |
| **A** | Timing | 1 | 63% | 8% | 7.3% | 1.0% |
| | Cost | 3 | 81% | 11% | 27.4% | 4.1% |
| | Demand | 2 | 64% | 6% | 15.1% | 1.5% |
| | Quality | 2 | 89% | 9% | 20.4% | 2.3% |
| | | Sum = 8 | Overall Product Acceptability = 70.2% | | | 8.9% |
| | | | | | | |
| **B** | **Criterion** | **Weight (1=low, 3=high)** | **Score** | **Amount of Uncertainty** | **Weighted Score** | **Weighted Uncertainty** |
| | Timing | 1 | 80% | 6% | 9.4% | 0.80% |
| | Cost | 3 | 63% | 16% | 20.4% | 6.0% |
| | Demand | 2 | 90% | 8% | 20.8% | 2.0% |
| | Quality | 2 | 65% | 12% | 14.5% | 3.0% |
| | | Sum = 8 | Overall Product Acceptability = 65.1% | | | 11.8% |
| **Scoring Margin = 5.0%** | | | | | | |

Equation 2: Equation to calculate a weighted score for each criterion.

$$Weighted\ Score_i = S_i * \frac{W_i}{(W_1 + W_2 + W_3 + W_4)} * \left(\frac{1}{1 + U_i}\right)$$

Equation 3: Example of using Equation 2 to calculate the weight score for criterion 1 (timing) of Product A.

$$Weighted\ Score_{1,A} = \ .63 * \frac{1}{(1 + 3 + 2 + 2)} * \left(\frac{1}{1 + 0.08}\right) = .07$$

Equation 4: The overall product acceptability is equal to the sum of the four weighted criterion scores.

$$Overall\ Product\ Acceptability_A = \sum_{i=1}^{4} Weighted\ Score_{\,i,A}$$

Equation 5: In this example, Equation 4 can be used to calculate an overall acceptability of 70.2% for Product A.

$$Overall\ Product\ Acceptability_A = \ 0.073 + 0.274 + \ 0.151 + 0.204 = \ .702$$

## Total Amount of Uncertainty Per Product

The detailed ProdStar recommendation was based on a weighted sum of the total amount of uncertainty for each product. This method is described by Equation 6.

Equation 6: Equation to calculate the total amount of uncertainty for the detailed ProdStar recommendation

$$Total\ Amount\ of\ Uncertainty = \sum_{i=1}^{4} U_i * \frac{W_i}{(W_1 + W_2 + W_3 + W_4)}$$

The first decision criterion (timing) for Product A is again used as an example. Using Equation 7, a weighted criterion uncertainty of 1.0% is calculated for the timing of Product A (Equation 8).

Equation 7: Equation to calculate the weighted uncertainty for each criterion.

$$Weighted\ Uncertainty_i = U_i * \frac{W_i}{(W_1 + W_2 + W_3 + W_4)}$$

Equation 8: In this example, Equation 7 is used to calculate the weighted uncertainty for the first criterion (timing) of Product A.

$$Weighted\ Uncertainty_1 = \ .06 * \frac{1}{(1 + 3 + 2 + 2)} = 0.08$$

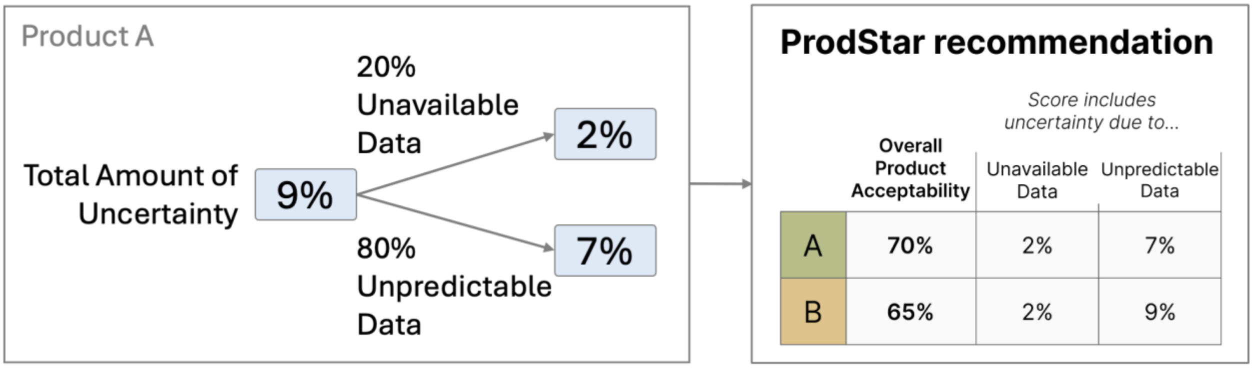
Equation 9: The total amount of uncertainty equals the sum of the four criterion’s weighted uncertainty.

$$Total\ Amount\ of\ Uncertainty_A = \sum_{i=1}^{4} Weighted\ Uncertainty_{i,A}$$

Equation 10: In this example, the total amount of uncertainty for Product A is 8.9%.

$$Total\ Amount\ of\ Uncertainty_A = 0.01 + 0.041 + 0.015 + 0.023 = 0.089$$

The total amount of uncertainty was split into epistemic uncertainty (unavailable data) and aleatoric uncertainty (unpredictable data). In half of the trials, the majority (a random number in the range of 65-90%) of the uncertainty was attributed to unpredictable data, and in the other half of the trials, the majority of the uncertainty was attributed to unavailable data. This information was displayed in the detailed ProdStar recommendation during Conditions 3 and 6. In this example, 80% of the uncertainty was due to unpredictable data, resulting in 7% uncertainty due to unpredictable data and 2% uncertainty due to unavailable data (Figure 3-5).



ProdStar recommendation

*Score includes uncertainty due to...*

| | Overall Product Acceptability | Unavailable Data | Unpredictable Data |
|---|---|---|---|
| A | 70% | 2% | 7% |
| B | 65% | 2% | 9% |

Figure 3-5: In this example, the total amount of uncertainty (8.9%) was rounded to one significant figure and distributed between uncertainty due to unavailable data (20%) and unpredictable data (80%). These values were displayed in the detailed ProdStar recommendations.

## Dependent Variables

The dependent variables included measures of task performance, such as decision accuracy, response time, and decisions to seek additional information. They also included subjective ratings, such as decision confidence, trust and perceived accuracy of recommendations, usability and net promoter score of visualization, and reliance on each element of the visualizations. The dependent variables are described and defined in

Table 3-4, and the full survey is included in APPENDIX C.

Table 3-4: Descriptions and definitions of the dependent variables. Details of the survey data are in APPENDIX C.

| **Construct** | **Dependent variable** | **Description** | **Definition** |
|---|---|---|---|
| Task Performance | Decision Accuracy | Percentage of optimal product selections | Count of trials with the optimal product selected divided by the number of completed trials |
| | Response Time | Mean decision response time | Mean decision response time (out of 45 seconds) across all 12 trials |
| | Decisions to Seek More Information | Percentage of trials to seek more information | Number of trials with recommendations to seek more information divided by the number of completed trials |
| Judgments of Information | Decision Confidence | Mean decision confidence | Mean decision confidence across all 12 trials |
| | Trust of Recommendations | Self-reported trust of recommendations (conditions 2, 3, 5, 6 only) | "How much did you trust the recommendations to be accurate?" (1-5 Likert scale) |
| | Perceived Accuracy of Recommendations | Perceived accuracy of recommendations (conditions 2, 3, 5, 6 only) | "How accurate do you think the recommendations were?" (0% accurate - 100% accurate) |
| Judgments of Visualization | SUS Score | System Usability Scale (SUS) score (Brooke, 1996) | Survey questions from validated SUS scale (score from 0-100) |
| | Net Promoter Score | Net Promoter Score (Reichheld, 2003) | "How likely is it that you would recommend this type of information display to a friend or colleague that is interested in decision-making support?" (1-10 Likert scale) |
| | Reliance on Visualization Elements | Reliance on:<br>-Bar chart<br>-Range of uncertainty on bar chart<br>-Criteria weights<br>-Recommendations<br>-ProdStar description of<br>-Sources of Uncertainty | Survey questions: How much did you rely on ___? (1-5 Likert scale) |

## Individual Differences

To explore RQ2, the experiment included an assessment of gender-trending individual differences. This included the GenderMag facet survey, which uses 20 questions to assess an individual's preferences for five cognitive facets (risk tolerance, information processing style, self-efficacy, motivations, and learning by process or by tinkering) (Vorvoreanu et al., 2019). The survey also asked participants to describe their gender, age, student classification, and with which college their major or home department is affiliated. Because machine learning researchers were found to interpret information visualizations with uncertainty differently than non-ML researchers (Arshad et al., 2015), the survey also asked participants to describe their knowledge about managing probabilistic data (1-5 Likert scale, ranging from "not knowledgeable at all" to "extremely knowledgeable."

## Hypotheses

The following hypotheses were developed based on the previous research described above as well as the methodology of this study. Previous research has shown that when information visualizations include uncertainty, users are more likely to decide they need to seek more information (Dong & Hayes, 2012) by encouraging them to think more deeply about the information (Kaur et al., 2020), but also to report lower decision confidence (Arshad et al., 2015; Dong & Hayes, 2012).

*H1: Participants will make recommendations to seek more information in a larger percentage of trials when provided with bar charts with uncertainty displayed than when provided bar charts without uncertainty displayed.*

*H2: Participants will report lower confidence in their decisions when provided with bar charts with uncertainty displayed than when provided bar charts without uncertainty displayed.*

Because of the small score margin between the two products, the ProdStar recommendation was expected to help participants make better decisions. Therefore, the decision accuracy was expected to be higher in the conditions with a recommendation than the conditions without a recommendation.

*H3: Participants will achieve higher decision accuracy when a ProdStar decision recommendation is provided than when not.*

The detailed recommendation that includes uncertainty has more information transparency, and high transparency is associated with higher trust in HAT research (J. Y. C. Chen et al., 2014; O'Neill et al., 2023). Therefore, the detailed recommendation was expected to be perceived as being more trustworthy than the basic recommendation. Trust was measured using the Likert survey question, "How much did you trust the recommendation to be accurate?"

*H4: Participants will report higher trust of the recommendations when a detailed recommendation is provided than when a basic recommendation is provided.*

The detailed recommendation describes the sources of uncertainty as being predominantly due to unavailable data (epistemic uncertainty) or predominantly due to unpredictable data (aleatoric uncertainty). Providing this additional information was expected to calibrate participants on whether they should seek additional information.

*H5: Participants will make recommendations to seek more information in a larger percentage of trials when the uncertainty is shown as being predominantly due to unavailable data than when the uncertainty is predominantly due to unpredictable data.*

Because the Abi GenderMag facets include lower risk tolerance, a comprehensive information processing style, and lower self-efficacy, participants with more Abi facets are expected to have longer response times and lower decision confidence than participants with more Tim facets.

*H6: Participants with the Abi GenderMag facets will have longer response times and report lower confidence in their decisions*.

Additionally, the comprehensive information style is characterized by systematically reviewing all available information, while the selective information processing style is characterized by making more efficient decisions using only the most important information. Therefore, participants with the comprehensive information processing style were predicted to more frequently make recommendations to seek additional information, and participants with the selective information processing were predicted to report higher reliance on the ProdStar recommendations.

*H7: Participants with the comprehensive information processing style will make recommendations to seek more information in a larger percentage of trials than participants with the selective information processing style*.

*H8: Participants with the selective information processing style will report higher reliance on decision recommendations than participants with the comprehensive information processing style.*

## Data Analysis

The analysis focused on the six dependent variables that were relevant to the hypotheses: decision accuracy, decision response time, decisions to seek more information, and decision confidence (continuous variables), as well trust of recommendations and reliance on recommendations (ordinal variables). The remaining dependent variables (SUS score, Net Promoter Score, and reliance on other dashboard elements) were collected for exploratory analysis and will be reserved for future work. There were three experimental independent variables: chart style, recommendation style, and predominant source of information uncertainty (only applicable in Conditions 3 and 6). Additionally, the analysis treated individual differences as independent variables: gender, the five GenderMag facets, and knowledge about probabilistic data. The effects of chart style, recommendation style, and gender were analyzed for each of the six dependent variables. The GenderMag facets, knowledge about probabilistic data, and predominant source of uncertainty were used for analyses only when they were relevant to a hypothesis.

Group differences for the continuous variables were analyzed with independent-samples t-tests (for independent variables with two levels) and one-way ANOVAs (for independent variables with three more levels). When the assumption of homogeneity of variances was violated, a one-way Welch ANOVA was run instead (Laerd Statistics, 2017). Ordinal variables were analyzed with Mann-Whitney U tests. Two-way and three-way ANOVA tests were conducted to evaluate interaction effects. Because ANOVA tests and independent-sample t-tests are both considered to be relatively robust with outliers and non-normal data, these tests were still used when the data included outliers or deviations from normality (Laerd Statistics, 2017). For ANOVA tests, effect sizes were reported as partial $\eta^2$ (.01 = small effect size, .06 = medium effect size, .14 = large effect size) (Cohen, 1988). Effect sizes for independent-samples t-tests were reported as Cohen’s d (.2 = small effect size, .5 = medium effect size, .8 = large effect size) (Cohen, 1988). Effect sizes for Mann-Whitney tests were reported with Pearson’s r correlation (.1 = small effect size, .3 = medium effect size, .5 = large effect size) (Cohen, 1988).

## Outlier Removal

Because this experiment was deployed through a campus-wide Qualtrics survey, some participant attrition was expected. Thus, the data was inspected to identify and remove outliers (Figure 3-6). First, participants who completed less than 80% of the survey were removed.

Participants who completed less than 100% but more than 80% of the survey were included because some participants completed the primary experiment task (the Product A vs. B decisions) but did not complete the entire post-task questionnaire. This criterion reduced the number of participants from 773 to 484.

Next, participants with any decision response times below seven seconds were removed. Seven seconds was selected as the filtering criterion because this was the shortest decision response time that could be achieved by the expert experimenters. This filter reduced the number of participants from 484 to 372.

Finally, participants were filtered by the time spent on the training page; participants who were not on that page of the survey long enough to receive the experiment task instructions were removed. Because the training video length varied between conditions, the data was sorted by condition and time spent on the training page. Any participants who spent less time on the training page than the video length were removed. This removed another 26 participants, leaving a total of 346 participants for the analysis. While data for participants who did not sufficiently complete the study were removed as outliers per the criteria described in this section, other data points will continue to be represented as outliers in the data analysis and results based on their relationship to the mean.

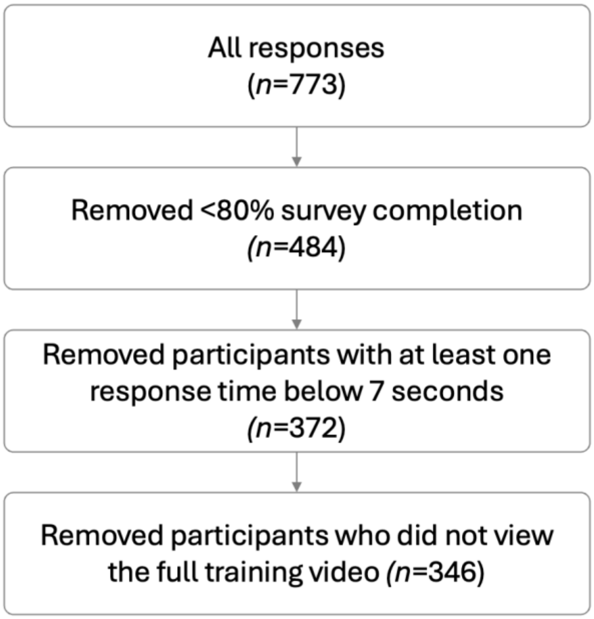

All responses
(n=773)
Removed <80% survey completion
(n=484)
Removed participants with at least one response time below 7 seconds
(n=372)
Removed participants who did not view the full training video (n=346)

Figure 3-6: Data was filtered to remove participants who did not complete more than 80% of the survey, spent less than 7 seconds on at least one of the 12 trials, or did not view the full training video.

### Individual Differences Independent Variables

Three types of individual difference variables were used as independent variables in the analysis: gender, GenderMag facets, and knowledge of probabilistic data. The total sample size was smaller in analyses of individual difference independent variables than experimental variables due to some participants not completing the full GenderMag and demographics surveys. Participants' gender was collected through a survey question, but due to the small number of non-binary or other gender responses (13 out of 346) participants, after removing outliers), gender was treated as dichotomous variable with only women and men.

To characterize participants' GenderMag facets, they were asked to complete the 20-question GenderMag facet survey, and facet scores were calculated according to the published protocol (Vorvoreanu et al., 2019). The GenderMag facets are typically defined within an experiment population, not against an absolute value. Therefore, the median score for each facet was calculated (after removing outliers); participants with a higher or lower score were assigned to the corresponding Abi or Tim facet. For participants with scores on the median, the recommended protocol of assigning them to the larger group was followed (e.g., if the population has more Tims, participants on the median are defined as Tims) (Vorvoreanu et al., 2019) (Figure 3-7).

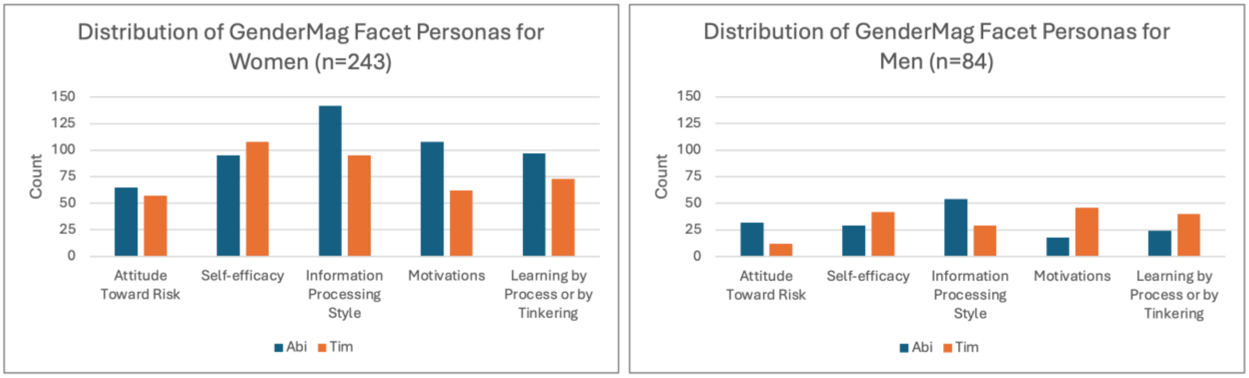


Figure 3-7: Distribution of GenderMag facets for women and men.

Knowledge of probabilistic data was measured through a single survey question: “How knowledgeable are you about managing probabilistic data (e.g., experience with statistics or machine learning)?” Responses were measured on a 1-5 Likert scale, ranging from “Not knowledgeable at all” to “Extremely knowledgeable.” In the analysis, this was treated as a dichotomous variable, with responses of 4 and 5 grouped as having high knowledge of probabilistic data (19%) and responses of 1, 2, or 3 grouped as having low knowledge of probabilistic data (81%). The threshold for high versus low was determined following the same process as the GenderMag facets, by grouping participants above and below the median, with participants right on the median going to the larger group (Figure 3-8).

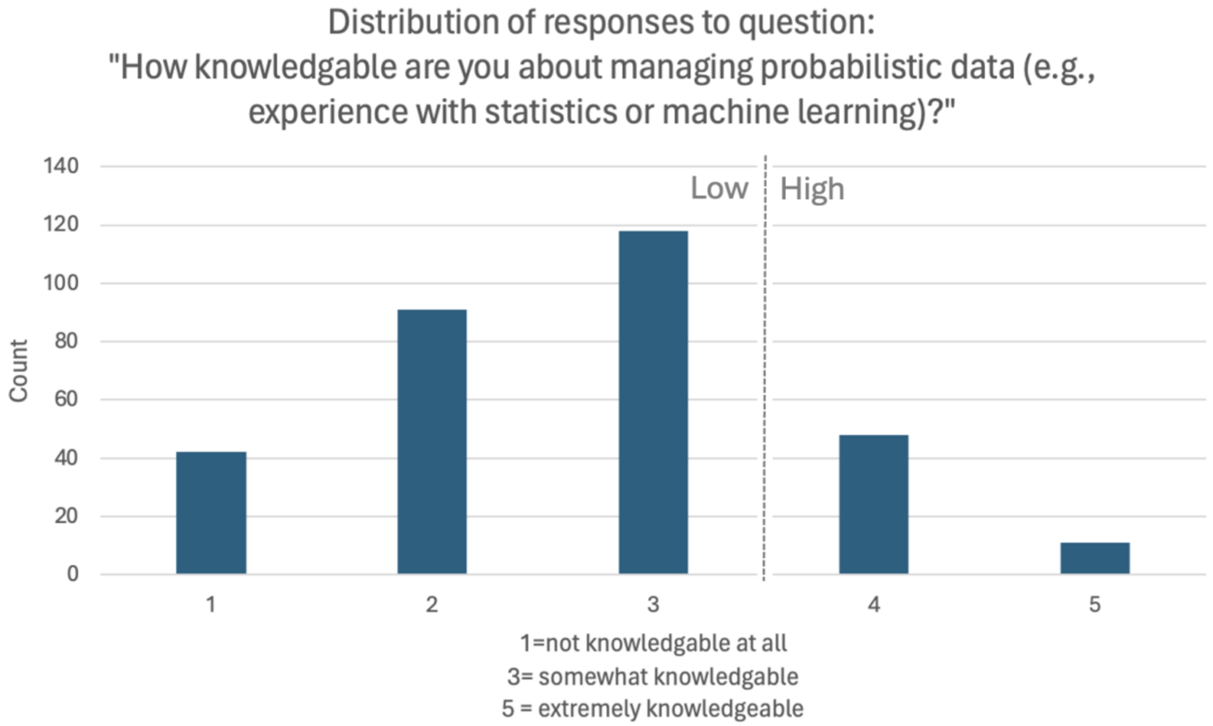


Figure 3-8: Knowledge of probabilistic data within the experiment population.

# CHAPTER 4. RESULTS

The analysis results are grouped by dependent variable, starting with continuous variables (decision accuracy, decision response time, decisions to seek more information, and decision confidence), then ordinal variables (trust of recommendations and reliance on recommendations). For each dependent variable, the effects of chart style, recommendation style, and gender were analyzed. Other variables (predominant source of information, GenderMag facets, and knowledge of probabilistic data) were included only when relevant to a hypothesis.

## Decision Accuracy

Decision accuracy was defined as the percentage of trials in which the participant selected the optimal product (A or B). For each participant, decision accuracy was calculated by dividing the count of their correct responses by the count of their completed responses. Because a response was not required on the A vs. B product selection questions and the page auto-advanced after 45 seconds, some participants did not complete all 12 trials. Therefore, decision accuracy for a participant who completed all 12 trials and selected the optimal product 9 times would be 75% (9/12 = 75%), and decision accuracy for a participant who completed 10 trials and selected the optimal product 9 times would be 90% (9/10 = 90%). This approach was used to avoid treating incomplete responses as being equivalent to incorrect responses.

There was one hypothesis related to decision accuracy: H1: Participants will achieve higher decision accuracy when a ProdStar recommendation is provided than when it is not. This hypothesis was tested with an ANOVA analysis. ANOVA analyses were also performed to explore the relationship between chart style, gender, GenderMag facets, and knowledge of probabilistic data. Results are reported below for recommendation style, chart style, gender, and knowledge of probabilistic data. No statistically significant results were identified for the effects of the GenderMag facets on decision accuracy.

### Decision Accuracy and Chart Style

Mean decision accuracy was statistically significantly higher for charts without uncertainty than charts with uncertainty (Figure 4-1). An independent-samples t-test was conducted to determine if the decision accuracy was different for the different chart styles. There were two chart styles: without uncertainty displayed (*n* = 176, Conditions 1, 2, and 3) and with uncertainty displayed (*n* = 170, Conditions 4, 5, and 6). There were five outliers and one extreme outlier, as assessed by boxplot; data was not normally distributed for each group, as assessed by Shapiro-Wilk test ($p < .001$); and there was homogeneity of variances, as assessed by Levene's test of homogeneity of variances ($p = .372$). Decision accuracy increased from the charts with uncertainty displayed (*M* = 71.4%, *SD* = 16.0%), to without uncertainty displayed (*M* = 82.3%, *SD* = 15.2%), and the differences between these chart styles was statistically significant, $t(1, 344) = 6.505$, $p < .001$, $d = 0.7$ (medium effect size).

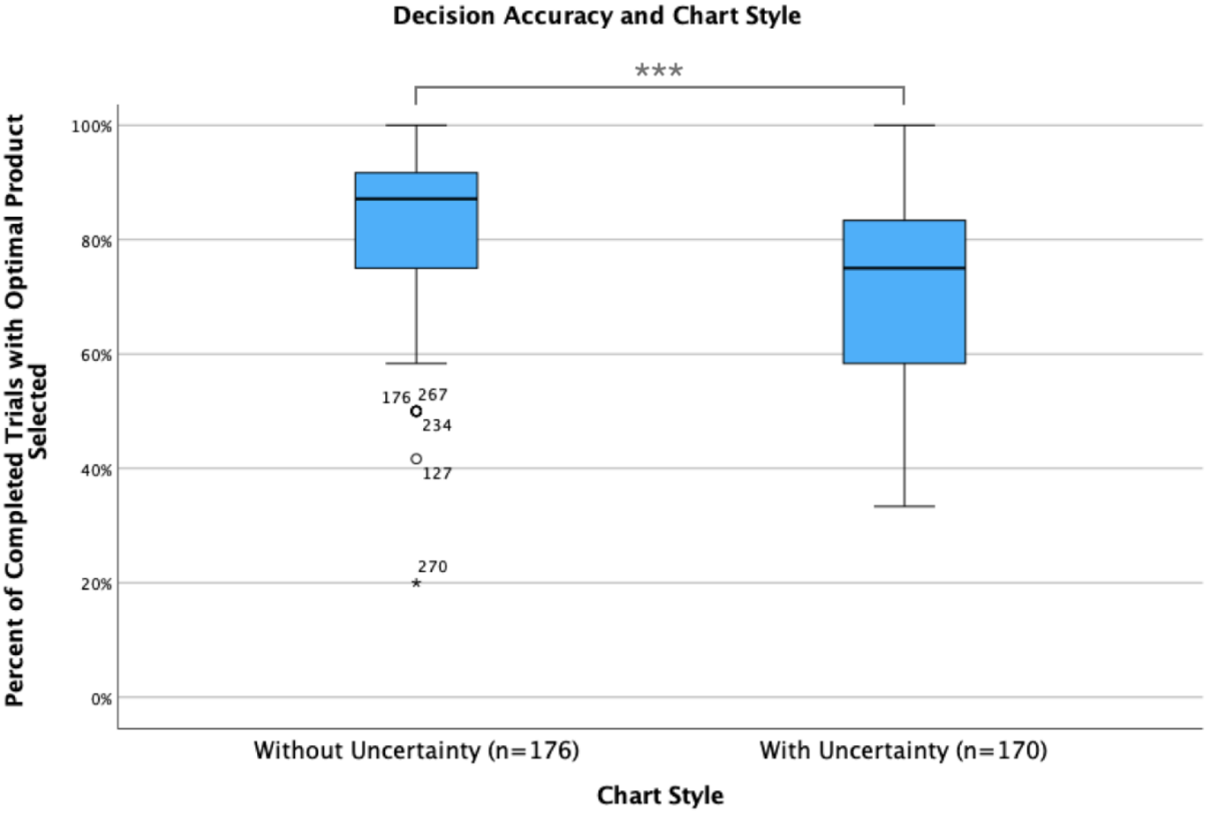


Figure 4-1: Decision accuracy was higher in charts without uncertainty than charts with uncertainty by a statistically significant amount.

**Decision Accuracy and Recommendation Style**

Descriptive statistics show that decision accuracy was higher with a detailed recommendation than without a recommendation (Figure 4-2). A one-way ANOVA was conducted to determine if decision accuracy was different with different recommendation styles. There were three recommendation styles: no recommendation (*n* = 111, Conditions 1 and 4), basic recommendation (*n* = 127, Conditions 2 and 5), and detailed recommendation (*n* = 108, Conditions 3 and 6). There was one outlier, assessed as being greater than 1.5 box-lengths from the edge of the box in a boxplot. The decision accuracy data was not normally distributed, as assessed by Shapiro-Wilk's test (p < .001). There was homogeneity of variances, as assessed by Levene's test for equality of variances (*p* = .183). Decision accuracy was statistically significantly different for different recommendation styles, *F*(2, 343) = 4.011, *p* = .019, partial $\eta^2$ = .023 (small effect size). There was an increase in decision accuracy from the no recommendation group (*M* = 73.3%, *SD* = 17.7%) to the basic recommendation group (*M* = 78.3%, *SD* = 15.1%), a mean increase of 4.95%, 95% CI [-0.0661%, 9.97%], which was not statistically significant (*p* = .054). There was an increase in decision accuracy from the basic recommendation group (*M* = 78.3%, *SD* = 15.1%), to the detailed recommendation group (*M* = 79.0%, *SD* = 16.5%), a mean increase of 0.775%, 95% CI [-4.28%, 5.83%], which was not statistically significant (*p* = .931). There was an increase in decision accuracy from the no recommendation group (M = 73.3%, *SD* = 17.7%), to the detailed recommendation group (*M* = 79.0%, *SD* = 16.5%), a mean increase of 5.7%, 95% CI [0.508%, 10.95%], which was statistically significant (*p* = .028).

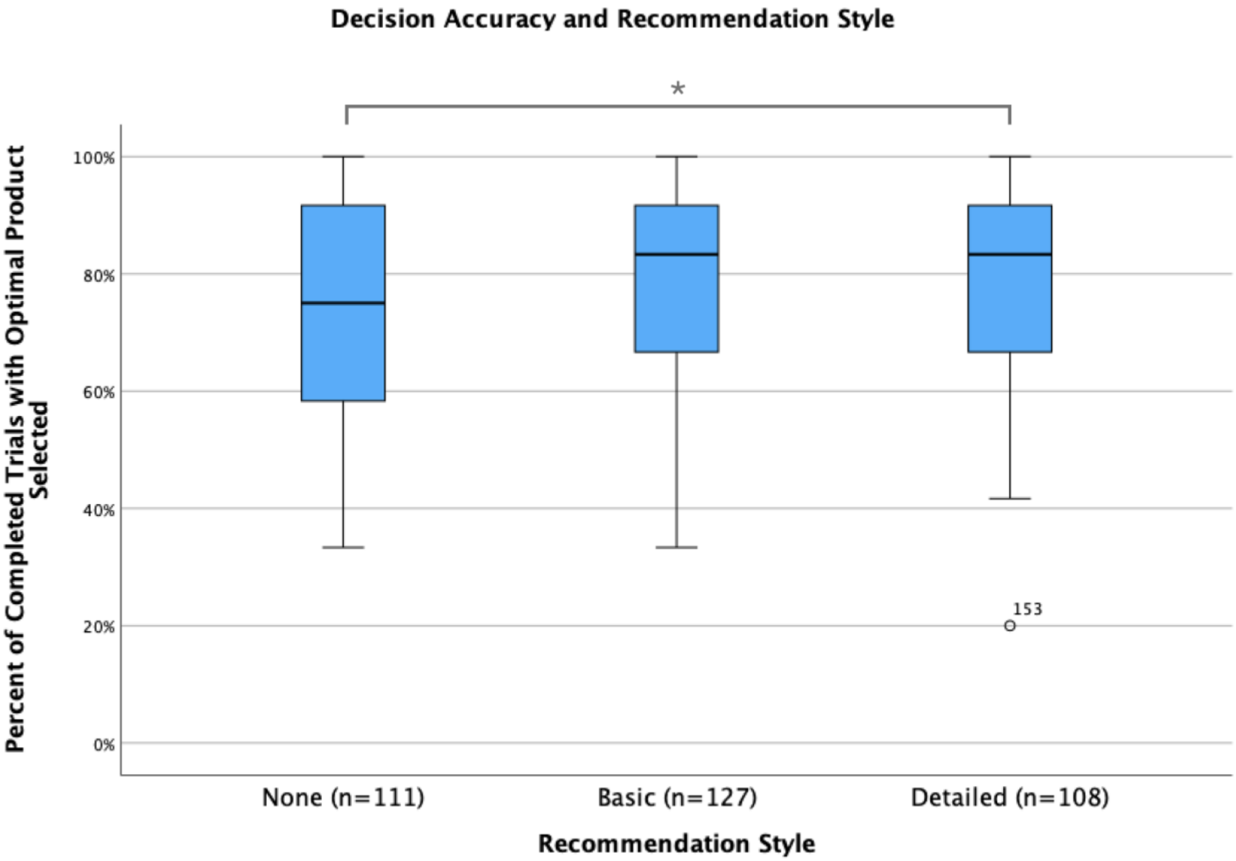


Figure 4-2: Decision accuracy was statistically significantly higher with a detailed recommendation style than with no recommendation.

### Decision Accuracy and Interaction Effects Between Chart and Recommendation Style

To check for interaction effects between chart style and recommendation style on decision accuracy, a two-way ANOVA was performed. There were five outliers, assessed as a value greater than 1.5 box-lengths from the edge of the box and one extreme outlier, assessed as a value greater than 3 box-lengths from the edge of the box. The data was not normally distributed, as assessed by Shapiro Wilk's test (*p* < .001), and there was homogeneity of variances, as assessed by Levene's test of homogeneity of variances (*p* = .484). There was not a statistically significant interaction effect of chart style and recommendation style on decision accuracy, *F*(2,340) = 1.892, *p* = .152, partial $\eta^2$ = .011 (small effect size).

The simple main effect of recommendation style on decision accuracy when charts included uncertainty was statistically significant, *F*(2,340) = 5.883, *p* = .003, partial $\eta^2$ = .033 (small effect size). All pairwise comparisons were made with a Bonferroni adjustment. When charts included uncertainty, there was an increase in decision accuracy from the no recommendation group (*M* = 65.60%, *SE* = 2.08%), to the basic recommendation group (*M* = 73.58%, *SE* = 2.02%), a mean increase of 7.99%, 95% CI [1.08%, 14.90%], which was statistically significant (*p* = .017). When charts included uncertainty, there was also an increase in decision accuracy from the no recommendation group (*M* = 65.60%, *SE* = 2.08%), to the detailed recommendation group (*M* = 74.58%, *SE* = 2.12%), a mean increase of 9.25%, 95% CI [2.048%, 16.448%], which was statistically significant (*p* = .007). When charts included uncertainty, there was a non-statistically significant mean increase of decision accuracy from the basic recommendation group to the detailed recommendation group of 1.26% (*p* = 1.00).

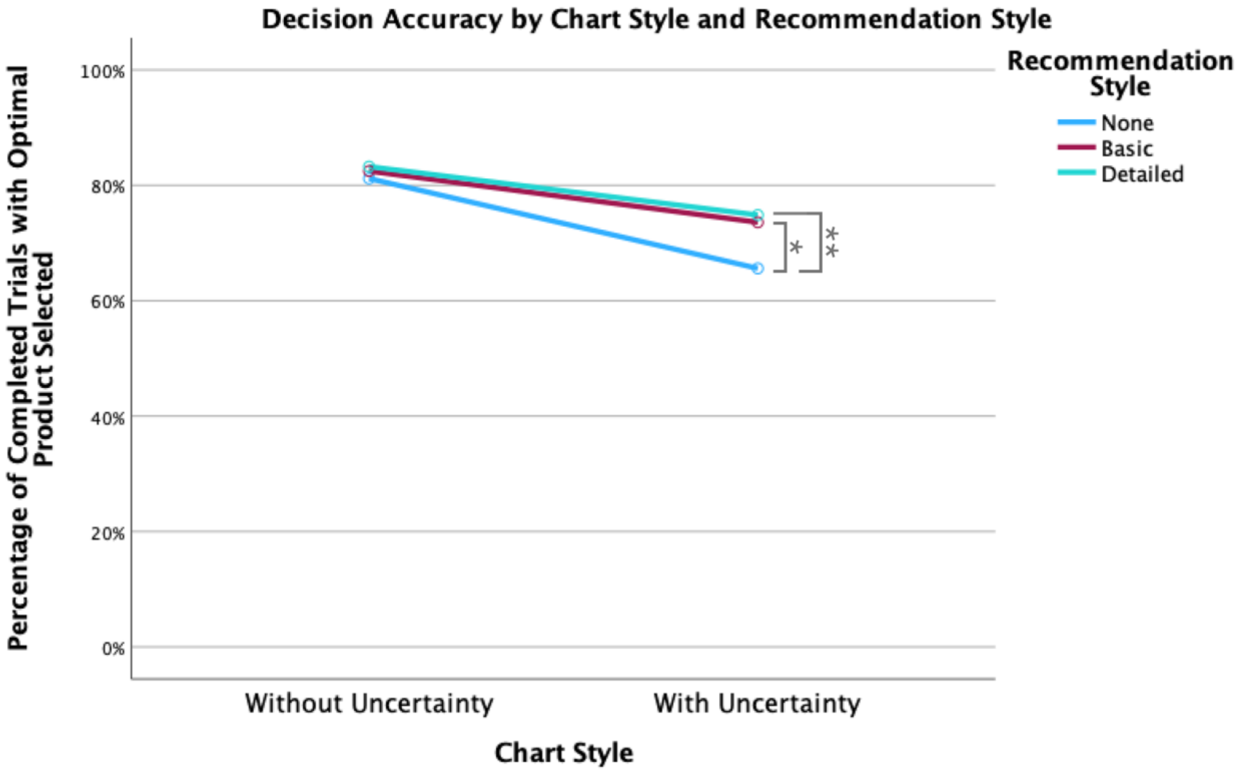


Figure 4-3: When charts did not include uncertainty, there was no statistically significant difference in decision accuracy for the different recommendation styles. When charts did include uncertainty, decision accuracy was statistically significantly higher with a basic or detailed recommendation than with no recommendation.

### Decision Accuracy and Gender

A one-way ANOVA was conducted to determine if the decision accuracy was different for the different genders. Two genders were considered: women (*n* = 243), and men (*n* = 84). There was one outlier, as assessed by boxplot; data was not normally distributed for each group, as assessed by Shapiro-Wilk test (*p* < .001); and there was homogeneity of variances, as assessed by Levene's test of homogeneity of variances (*p* = .152). Decision accuracy was lower for women (*M* = 76.25%, *SD* = 16.63%) than men (*M* = 79.81%, *SD* = 15.59%), a non-statistically significant difference, *F*1, 325) = 2.953, *p* = .087, partial $\eta^2$ = .009 (trivial effect size) (Figure 4-4).

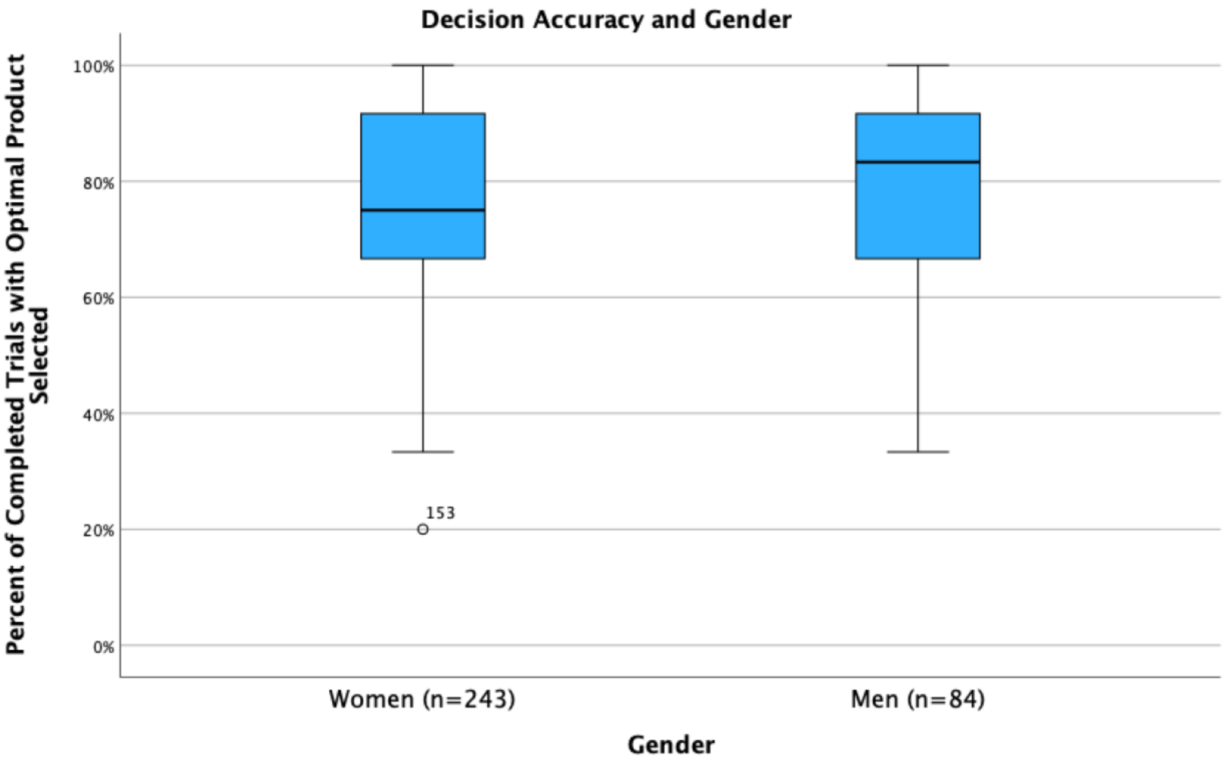


Figure 4-4: No statistically significant difference was found in decision accuracy of women vs. men.

### Decision Confidence and Knowledge about Managing Probabilistic Data

An independent-samples t-test was run to determine if there were differences in decision confidence participants with more or less knowledge about managing probabilistic data. There were ten outliers in the data, as assessed by inspection of a boxplot. Decision accuracy was not normally distributed, as assessed by Shapiro-Wilk's test, and there was homogeneity of variances, as assessed by Levene's test for equality of variances (*p* = .055). Mean decision accuracy was not statistically significantly different between participants with less knowledge about managing probabilistic data (*M* = 76.40%, *SD* = 17.20%) and participants with more knowledge about managing probabilistic data (*M* = 79.88%, *SD* = 14.15%), *t*(308) = -1.445, *p* = .053, *d* = -.209 (small effect size) (Figure 4-34).

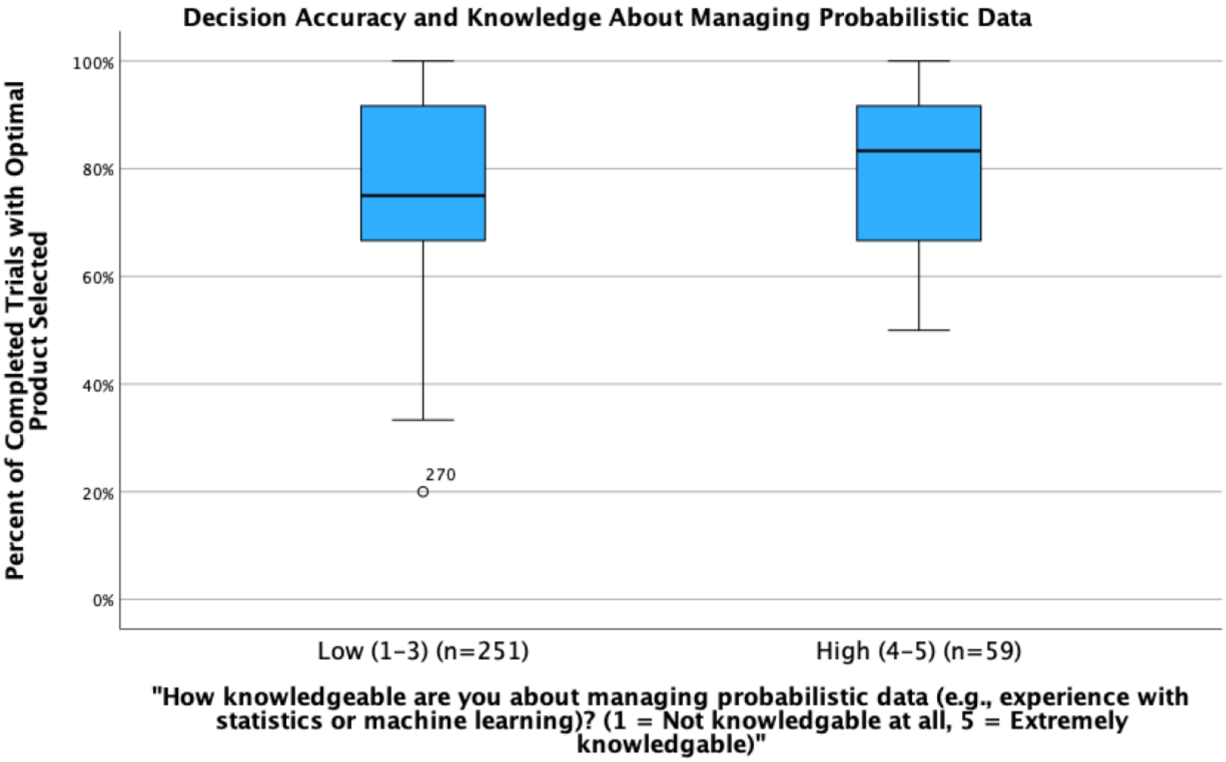


Figure 4-5: Decision accuracy was not statistically significantly different between participants with less knowledge about managing probabilistic data and participants with more knowledge about managing probabilistic data.

## Decision Response Time

Decision response time is calculated as the mean response time across all 12 trials. This includes the trials in which the page auto-advanced after 45 seconds, which occurred at least once for 16% of participants. One hypothesis included decision response time: H6: The Abi GenderMag facets will be associated with longer response times and lower decision confidence. This hypothesis was tested by performing independent t-tests, and a two-way ANOVA was performed to check for interaction effects. To understand the effects of the 45-second time limit, the

### Decision Response Time and Chart Style

A Welch t-test was run to determine if there were differences in decision response time between charts without uncertainty and charts with uncertainty due to the assumption of homogeneity of variances being violated, as assessed by Levene's test for equality of variances (*p* = .043). There were four outliers in the data, as assessed by inspection of a boxplot, and decision response times for chart style were not normally distributed, as assessed by Shapiro-Wilk's test (*p* < .001). Decision response times were shorter for charts without uncertainty (*M* = 19.26, *SD* = 4.96) than charts with uncertainty (*M* = 20.79, *SD* = 5.77), a statistically significant difference, *M* = -1.53, 95% CI [-2.67, -0.40], *t*(332.6) = -2.646, *p* = .004, *d* = -0.282 (small effect size).

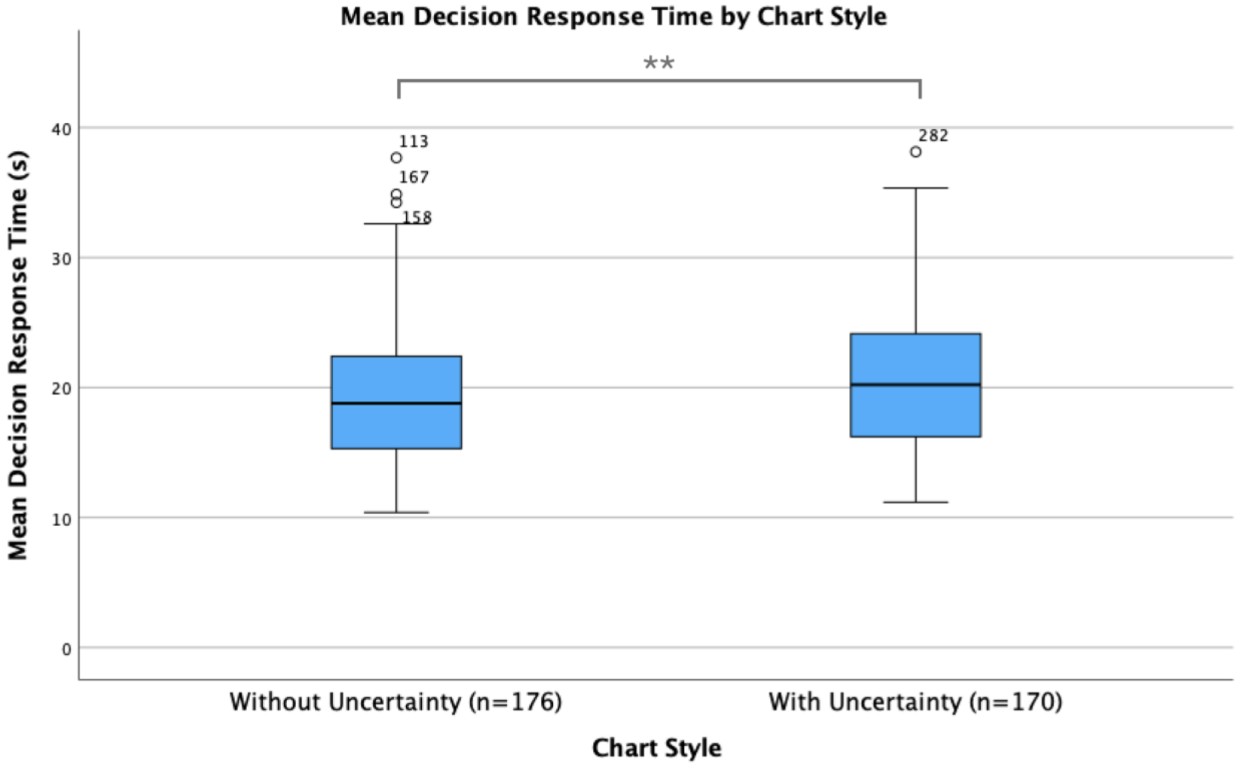


Figure 4-6: Mean decision response time was 1.42 seconds shorter for charts without uncertainty than charts with uncertainty, a statistically significant difference.

### Decision Response Time and Recommendation Style

A one-way ANOVA was conducted to determine if the mean decision response times were different for different recommendation styles. There were three decision recommendation styles: none (*n* = 111), basic (*n* = 127), and detailed (*n* = 108). There were five outliers, as assessed by boxplot; data was not normally distributed for each group, as assessed by Shapiro-Wilk test (*p* < .001); and there was homogeneity of variances, as assessed by Levene's test of homogeneity of variances (*p* = .266). Mean decision response times were statistically significantly different between different recommendation styles, *F*(2, 343) = 10.004, *p* < .001, partial $\eta^{2}$ = .055 (small effect size). Mean decision response times increased for no recommendations (*M* = 18.44, *SD* = 4.81) to basic recommendations (*M* = 20.01, *SD* = 5.30) to detailed recommendations (*M* = 21.63, *SD* = 5.71), in that order. Tukey post hoc analysis revealed that the mean increase from no recommendations to detailed recommendations (3.20, 95% CI [1.51, 4.88]) was statistically significant (*p* < .001), as well as the increase from basic recommendations to detailed recommendations (1.63, 95% CI [0.00, 3.26], *p* = .050), but the increase from no recommendations to basic recommendations was not statistically significant (1.57, 95% CI [0.020, 3.11], *p* = .060) (Figure 4-7).

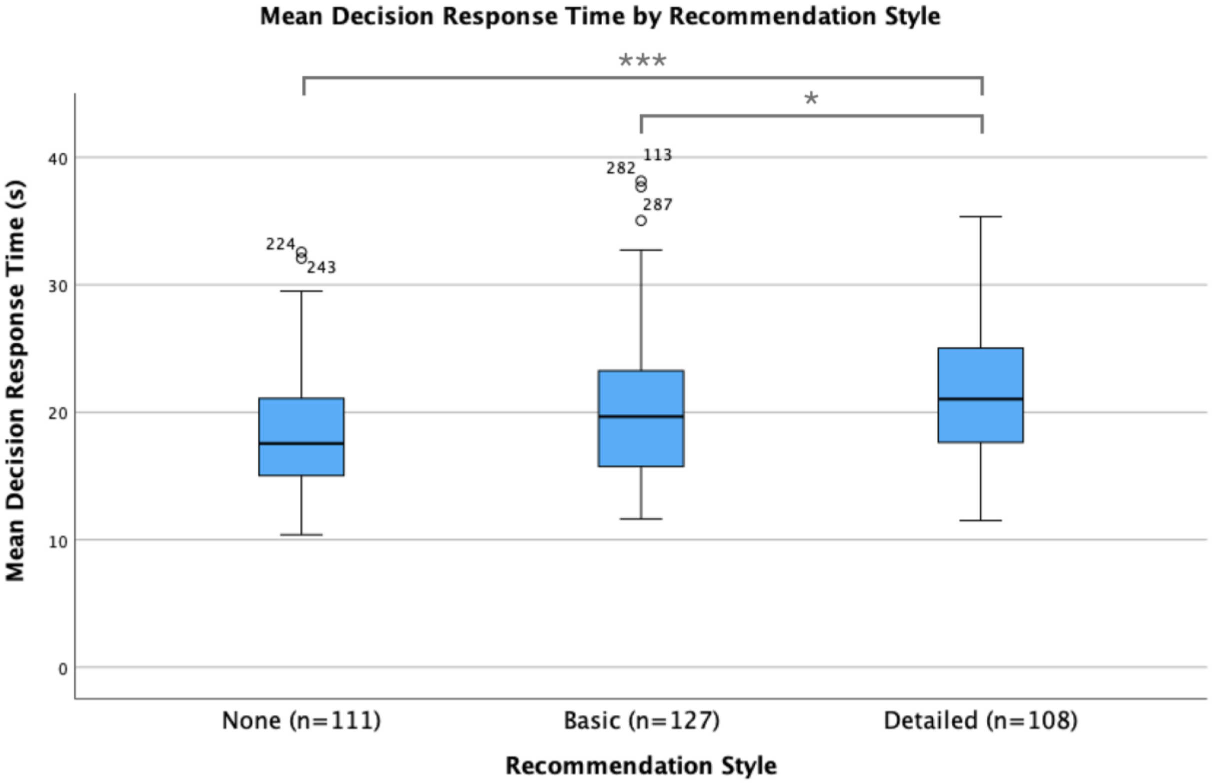


Figure 4-7: The difference between decision response times for no recommendations and detailed recommendations was statistically significant, and the difference between decision response times for basic and detailed recommendations was statistically significant.

### Decision Response Time and Gender

An independent-samples t-test was run to determine if there were differences in decision response time between women and men. There were nine outliers in the data, as assessed by inspection of a boxplot. Decision response times were not normally distributed, as assessed by Shapiro-Wilk's test (*p* < .001), and there was homogeneity of variances, as assessed by Levene's test for equality of variances (*p* = .418). Mean decision response times were not statistically significantly different between women (*M* = 19.98, *SD* = 5.19) and men (*M* = 20.37, *SD* = 6.02), *t*(325) = -0.574, *p* = .283, *d* = -.072 (medium effect size).

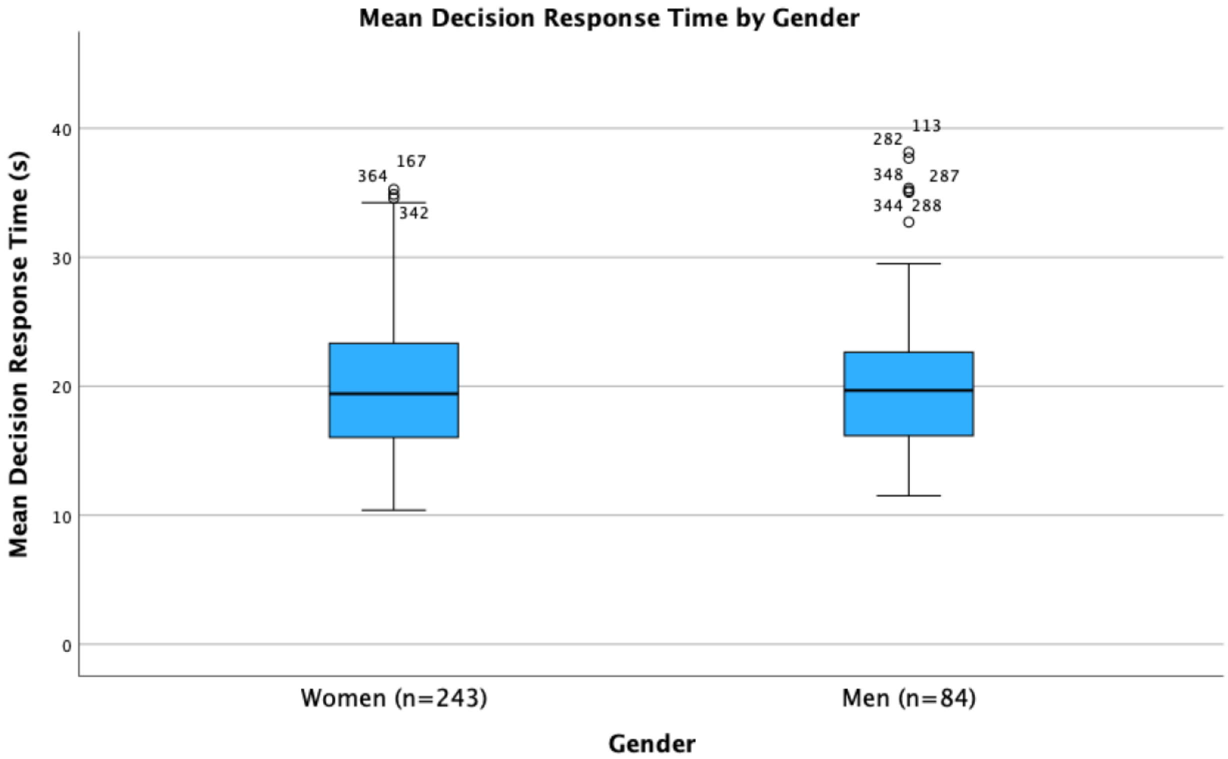


Figure 4-8: Mean decision response times were not statistically significantly different for women and men.

## Decision Response Time and GenderMag Facets

### *Attitude Toward Risk*

An independent-samples t-test was run to determine if there were differences in decision response time between participants with lower and higher tolerance of risk. There were seven outliers in the data, as assessed by inspection of a boxplot. Decision response times were not normally distributed, as assessed by Shapiro-Wilk's test ($p$ < .001), and there was homogeneity of variances, as assessed by Levene's test for equality of variances ($p$ = .606). Mean decision response times were not statistically significantly different between Abi's lower risk tolerance ($M$ = 20.66, $SD$ = 5.57) and Tim's higher risk tolerance ($M$ = 20.46, $SD$ = 5.93), $t$(168) = 0.216, $p$ = .414, $d$ = .033 (trivial effect size) (Figure 4-9).

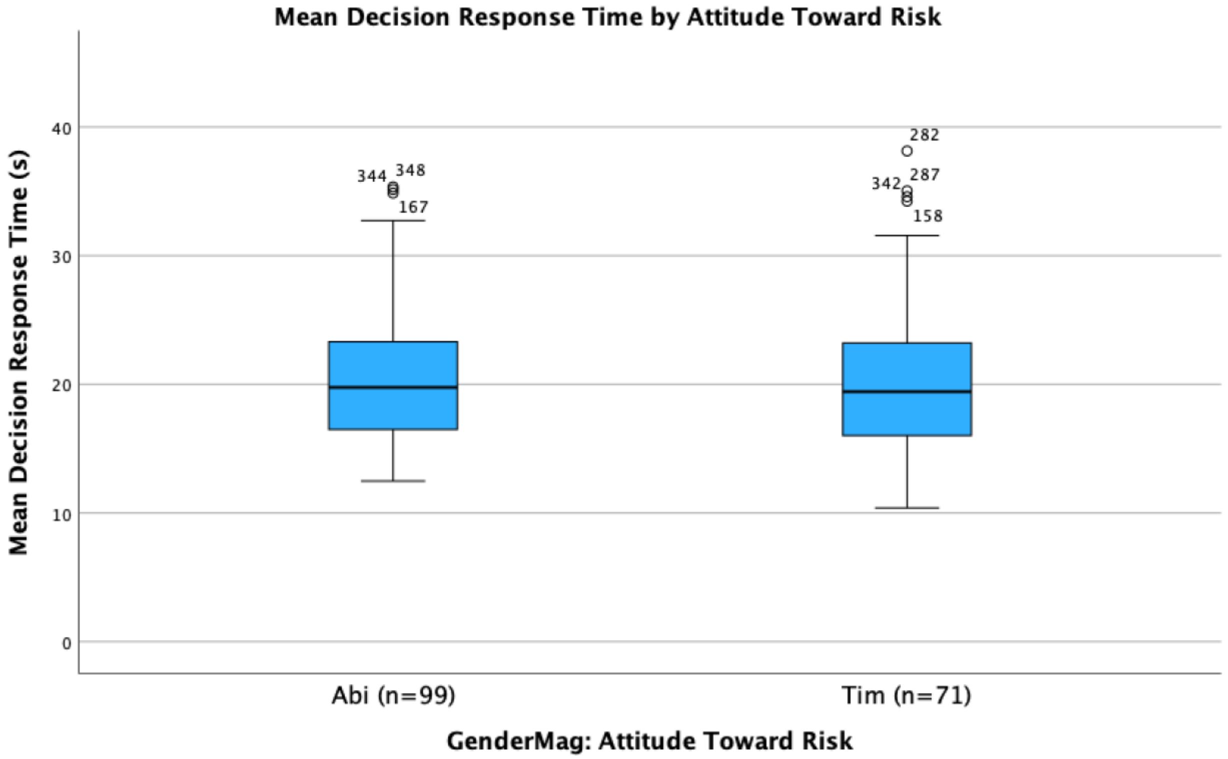


Figure 4-9: Mean decision response time was not statistically significantly different between participants with a lower tolerance for risk (Abi) and participants with a higher tolerance for risk (Tim).

*Self-Efficacy*

An independent-samples t-test was run to determine if there were differences in decision response time between participants with low and high self-efficacy. There were seven outliers in the data, as assessed by inspection of a boxplot. Decision response times were not normally distributed, as assessed by Shapiro-Wilk's test (*p* < .001), and there was homogeneity of variances, as assessed by Levene's test for equality of variances (*p* = .687). Mean decision response times were not statistically significantly different between Abi’s lower self-efficacy (*M* = 19.65, *SD* = 5.19) and Tim’s higher risk tolerance (*M* = 20.53, *SD* = 5.58), *t*(281) = -1.356, *p* = .088, *d* = -.162 (trivial effect size) (Figure 4-10).

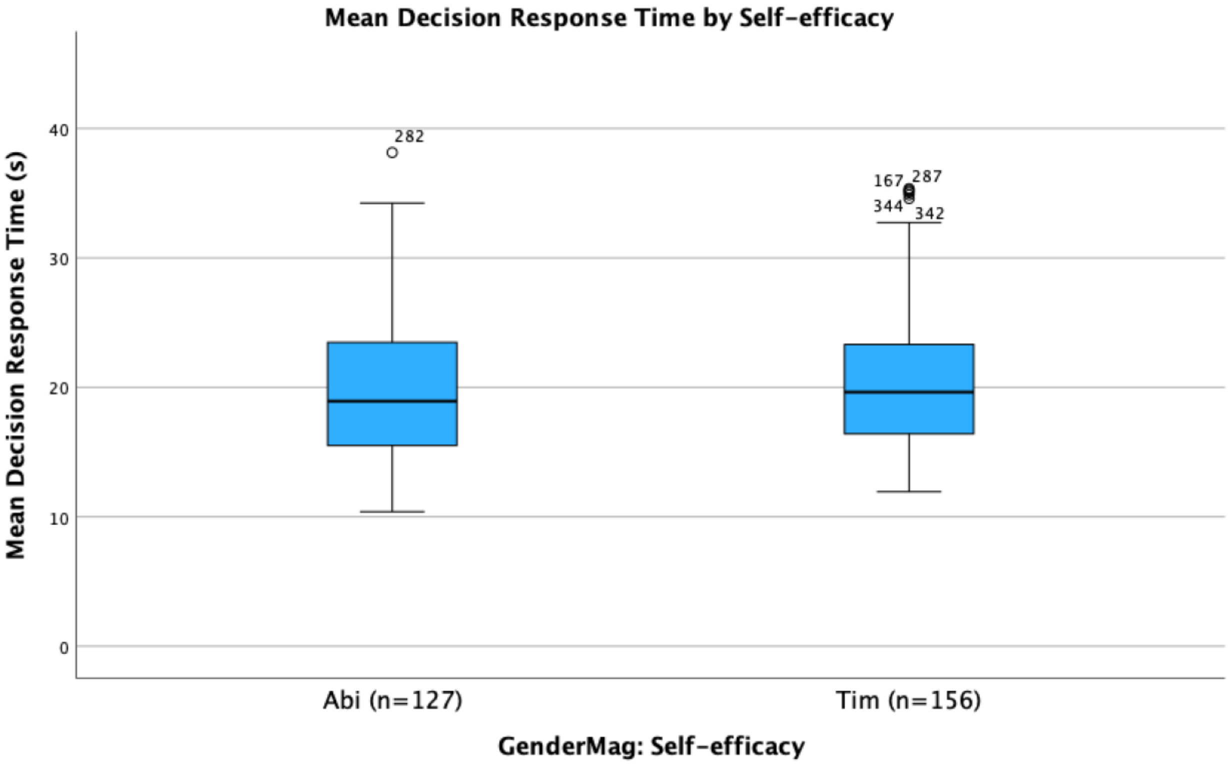


Figure 4-10: Mean decision response time was not statistically significantly different between participants with low self-efficacy (Abi) and high self-efficacy (Tim).

*Information Processing Style*

A Welch t-test was run to determine if there were differences in decision response time between information processing styles due to the assumption of homogeneity of variances being violated, as assessed by Levene's test for equality of variances (*p* = .047). There were three outliers in the data, as assessed by inspection of a boxplot, and decision response times for the two information processing styles were not normally distributed, as assessed by Shapiro-Wilk's test (*p* < .001). Mean decision response times were not statistically significantly different for participants with Abi's comprehensive information processing style (*M* = 20.40, *SD* = 5.70), *t*(296) = 1.462, *p* = .072, *d* = .159 (trivial effect size) (Figure 4-11).

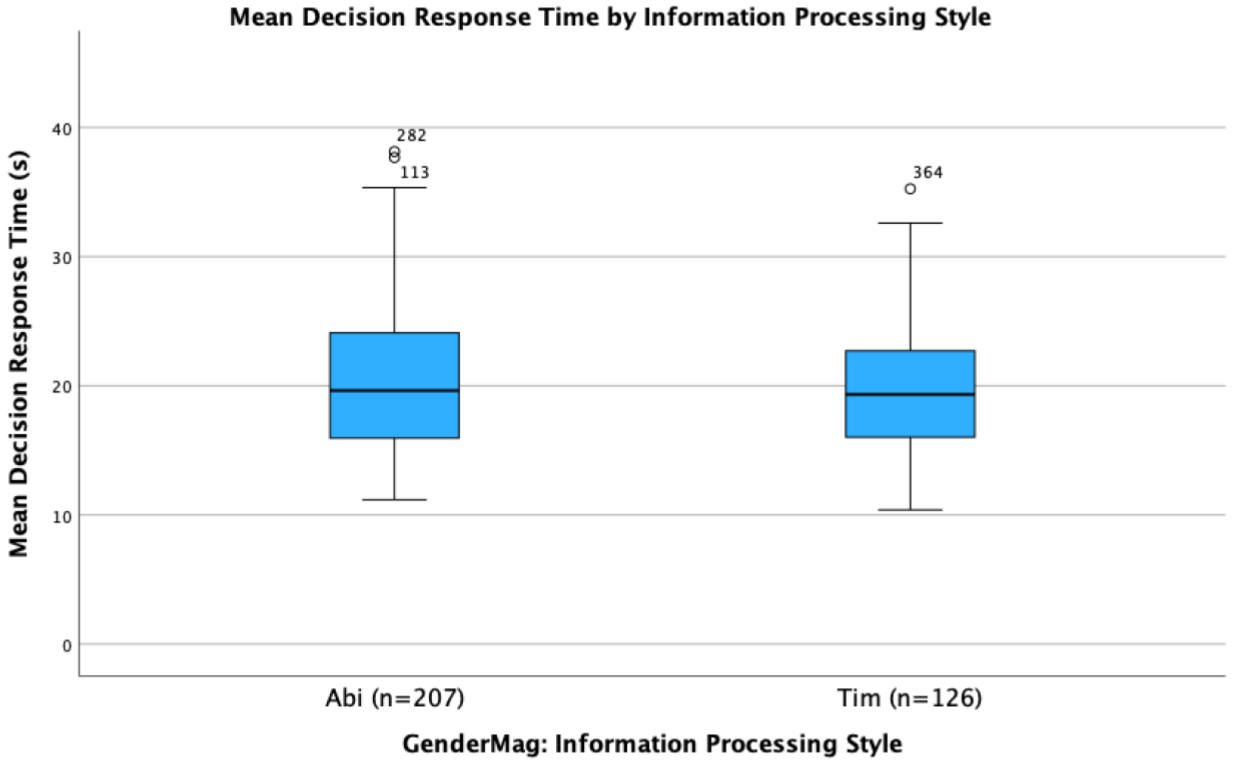


Figure 4-11: Mean decision response time was not statistically significantly different between participants with a comprehensive information processing style (Abi) and participants with a selective information processing style (Tim).

*Motivations*

An independent-samples t-test was run to determine if there were differences in decision response time between participants with different motivations for using technology. There were four outliers in the data, as assessed by inspection of a boxplot. Decision response times were not normally distributed, as assessed by Shapiro-Wilk's test (*p* < .001), and there was homogeneity of variances, as assessed by Levene's test for equality of variances (*p* = 0.390). Mean decision response times were not statistically significantly different between Abi's motivation of using technology to accomplish a goal (*M* = 20.12, *SD* = 5.88) and Tim's motivation of using technology for technology's sake (*M* = 20.41, *SD* = 5.12), *t*(223) = -0.390, *p* = .349, *d* = -.052 (trivial effect size) (Figure 4-12).

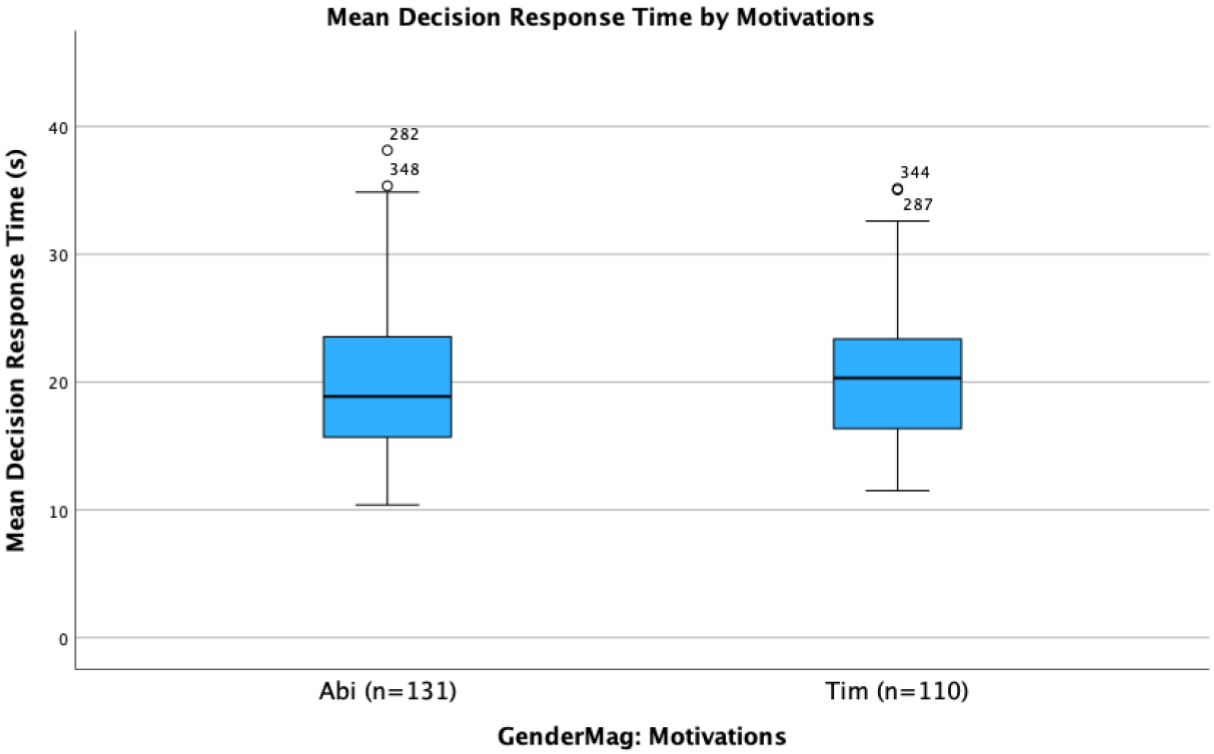


Figure 4-12: Mean decision response time was not statistically significantly different between participants who are motivated to use technology to accomplish a goal (Abi) and participants who are motivated to use technology for its own sake (Tim).

*Learning by Process or by Tinkering*

An independent-samples t-test was run to determine if there were differences in decision response time between participants who learn by process or by tinkering. There were four outliers in the data, as assessed by inspection of a boxplot. Decision response times were not normally distributed, as assessed by Shapiro-Wilk's test ($p$ < .001), and there was homogeneity of variances, as assessed by Levene's test for equality of variances ($p$ = .731). Mean decision response times were not statistically significantly different between Abi’s preference to learn by process ($M$ = 19.75, $SD$ = 5.65) and Tim’s preference to learn by tinkering ($M$ = 20.84, $SD$ = 55.30), $t$(239) = -1.535, $p$ = .063, $d$ = .198 (trivial effect size) (Figure 4-13).

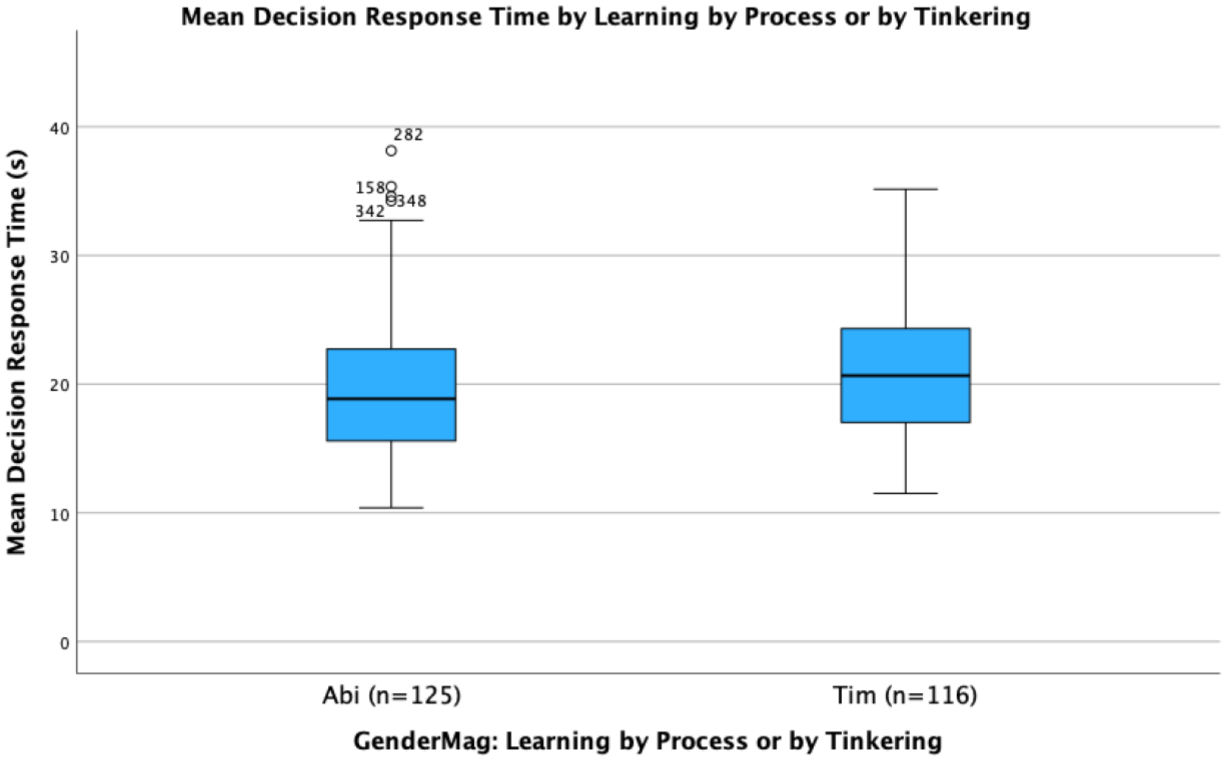


Figure 4-13: Mean decision response time was not statistically significantly different between participants with preferences for learning by process (Abi) or by tinkering (Tim).

### Decision Response Time and Interaction Effects

A three-way ANOVA was run to check for interaction effects between chart style, recommendation style, and gender. There were nine outliers, assessed as a value greater than 1.5 box-lengths from the edge of the box, and one extreme outlier, assessed as a value greater than 3 box-lengths from the edge of the box. Decision response times were normally distributed ($p > .05$) except for one group (women using charts with uncertainty and no recommendations, $p < .001$), as assessed by Shapiro-Wilk's test of normality. There was homogeneity of variances, as assessed by Levene's test for equality of variances, $p = .051$. There was no statistically significant three-way interaction between chart style, recommendation style, and gender, $F(2,315) = 1.676$, $p = .189$, partial $\eta^2 = .011$ (small effect size). There were no statistically significant interactions for decision response time between chart style and recommendation style, *F*(2, 340) = 0.084, *p* = .920, partial $\eta^2$ = .000 (no effect size), chart style and gender, *F*(1, 323) = 0.574, *p* = .449, partial $\eta^2$ = .002 (small effect size), or recommendation style and gender, *F*(2, 321) = 1.725, *p* = .180, partial $\eta^2$ = .011 (small effect size) (Figure 4-14).

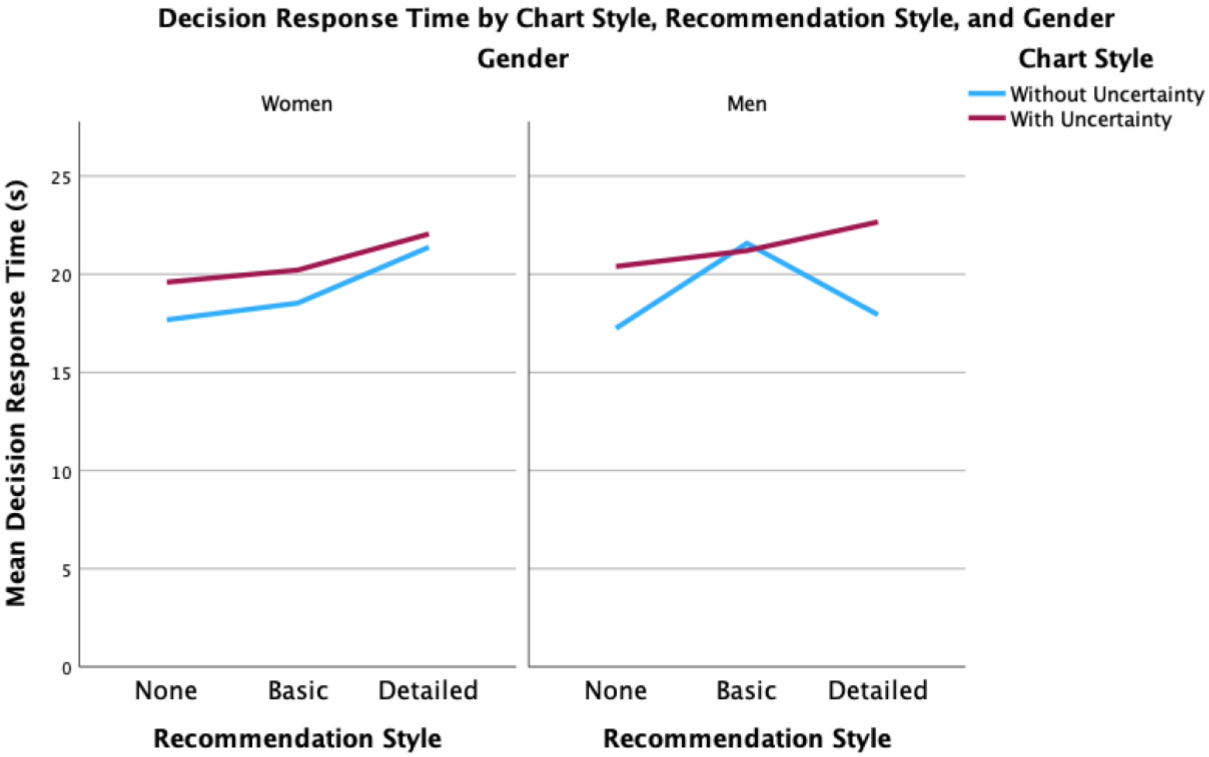


Figure 4-14: No statistically significant interactions were found between chart style, recommendation style, and gender.

## Decisions to Seek More Information

Decisions to seek more information were based on the yes/no question: “Should MegaMart seek additional information before making a final decision?” As with decision accuracy, this variable was calculated as a percentage of completed trials. A participant who answered “yes” to this question 6 times and responded to this question in all 12 trials would be described as recommending that MegaMart seek more information 50% of the time (6/12 = 50%). A participant who answered “yes” to this question 6 times and responded to this question in only 10 trials would be described as recommending that MegaMart seek more information 60% of the time (6/10 = 60%). This approach was taken to avoid treating "no" and "no response" as being equivalent.

There were three hypotheses about decisions to seek more information: H3: Bar charts with uncertainty will be associated with higher rates of deciding to seek more information than bar charts without uncertainty, H5: Detailed recommendations will be associated with fewer recommendations to seek additional information than basic recommendations, and H7: The comprehensive information processing style will be associated with more recommendations to seek additional information than the selective information processing style. To evaluate these hypotheses, a two-way ANOVA was conducted to explore the effects of chart style and recommendation style on decisions to seek more information. One-way ANOVAs were conducted to explore the effects of gender and the GenderMag facets. A two-way ANOVA was conducted to explore the effects of the predominant source of uncertainty and knowledge of managing probabilistic data on decisions to seek more information.

### Decisions to Seek More Information and Chart Style

An independent-samples t-test was run to determine if there were differences in rates of deciding to seek more information between chart styles. There were no outliers in the data, as assessed by inspection of a boxplot. Rates of deciding to seek more information for each chart style were not normally distributed, as assessed by Shapiro-Wilk's test ($p < .001$), and there was homogeneity of variances, as assessed by Levene's test for equality of variances ($p = .815$). The rates of deciding to seek more information when participants were presented with charts without uncertainty (*M* = 61.89%, *SD* = 26.76%) vs. with charts with uncertainty (*M* = 63.14%, *SD* =

27.07%) were not statistically significant different, *M* = -1.25%, 95% CI [-6.94%, 4.44%], *t*(334) = -0.432, *p* = .333 *d* = .047 (trivial effect size) (Figure 4-15).

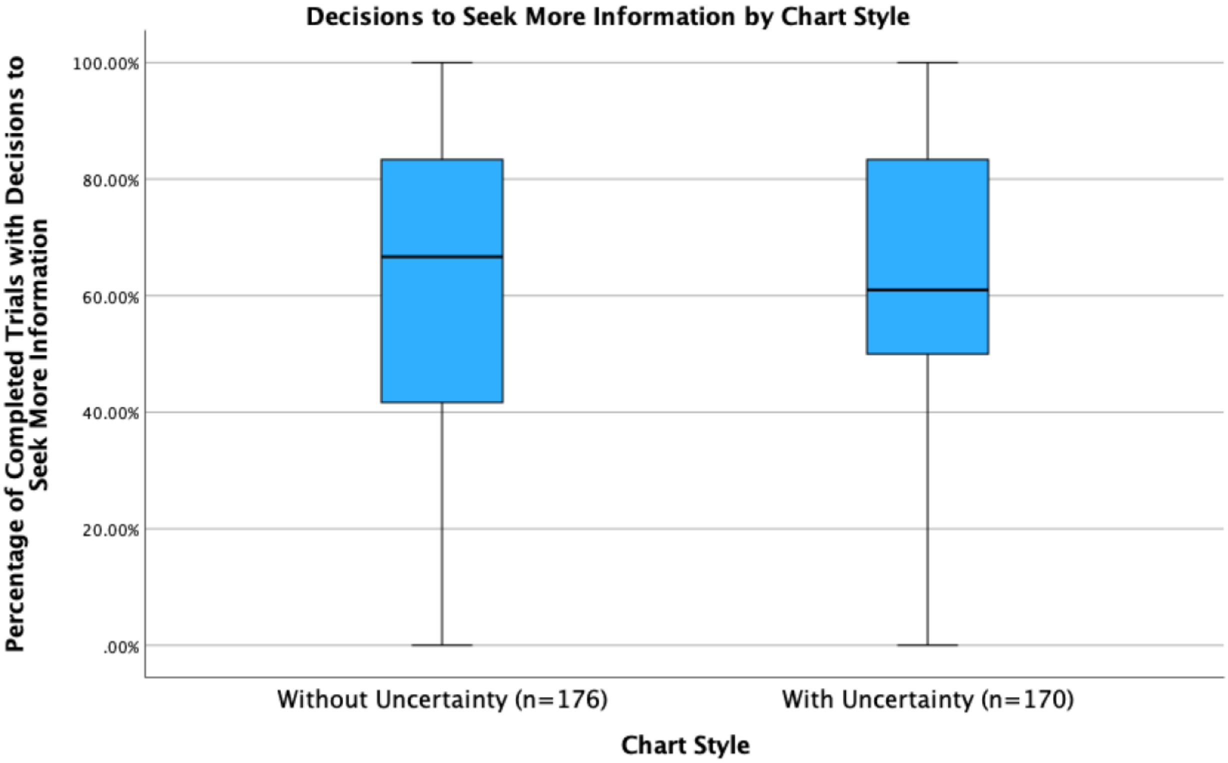


Figure 4-15: Rates of deciding to seek more information were not statistically significantly different for different chart styles.

### Decisions to Seek More Information and Recommendation Style

A one-way ANOVA was conducted to determine if there are differences in decisions to seek more information between recommendation styles. There were no outliers, as assessed by boxplot, and the data was not normally distributed, as assessed by Shapiro-Wilk's test of normality (*p* < .001). There was homogeneity of variances, as assessed by Levene's test for equality of variances, *p* = .284. Rates of deciding to seek more information differed among participants who saw no recommendations (*M* = 62.40%, *SD* = 2.64%), participants who saw basic recommendations (*M* = 64.31%, *SD* = 27.80%), and participants who saw detailed recommendations (*M* = 60.48%, *SD* = 24.84%). The differences between these recommendation style groups were not statistically significant, $F(2, 345) = 0.590$, $p = .555$, $\eta^2 = .274$ (large effect size) (Figure 4-16).

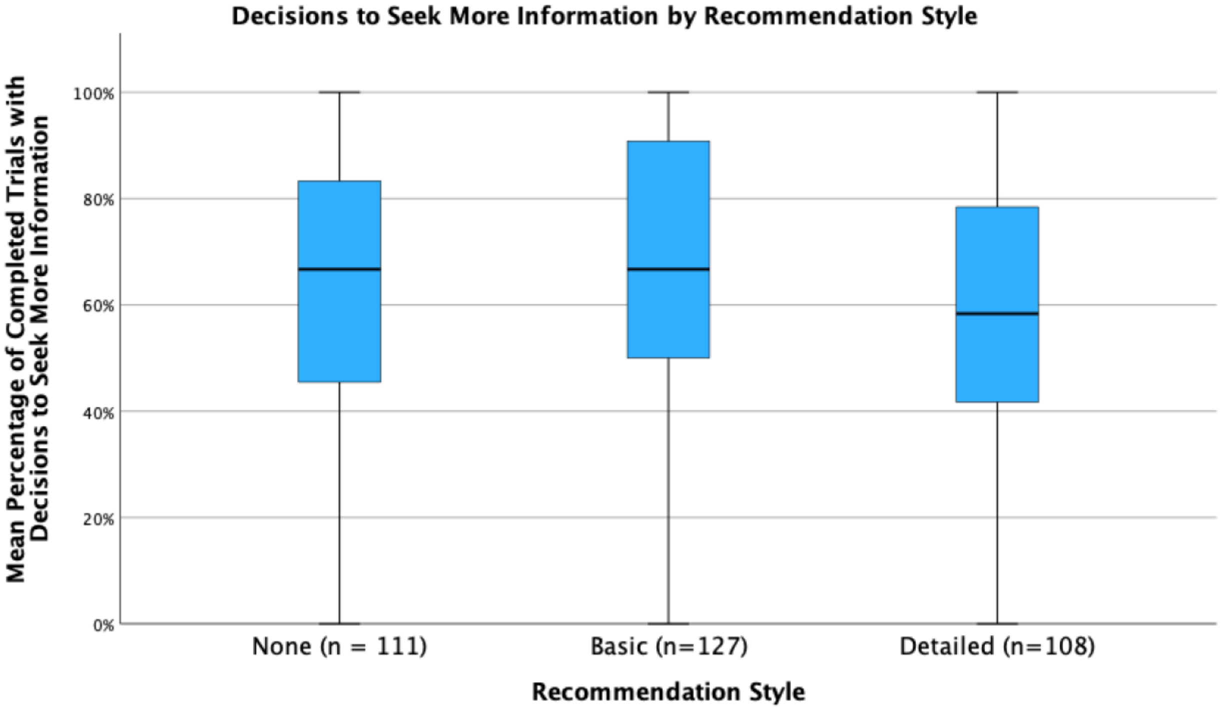


Figure 4-16: Rates of deciding to seek more information were not statistically significantly different between recommendation styles. Response ranged from recommending to seek more information on 0% of trials to 100% of trials in all three conditions.

### Decisions to Seek More Information and Gender

An independent-samples t-test was run to determine if there were differences in rates of deciding to seek more information between genders. There were no outliers in the data, as assessed by inspection of a boxplot. Rates of deciding to seek more information for each gender were not normally distributed, as assessed by Shapiro-Wilk's test ($p < .001$), and there was homogeneity of variances, as assessed by Levene's test for equality of variances ($p = .846$). There were higher rates of women deciding to seek more information ($M = 64.90\%$, $SD = 26.19\%$) than men deciding to seek more information ($M = 53.03\%$, $SD = 27.25\%$), a statistically significantly difference, *M* = 11.87%, 95% CI [5.28%, 18.46%], *t*(325) = 3.543, *p* = < .001, *d* = 0.448 (small effect size) (Figure 4-17).

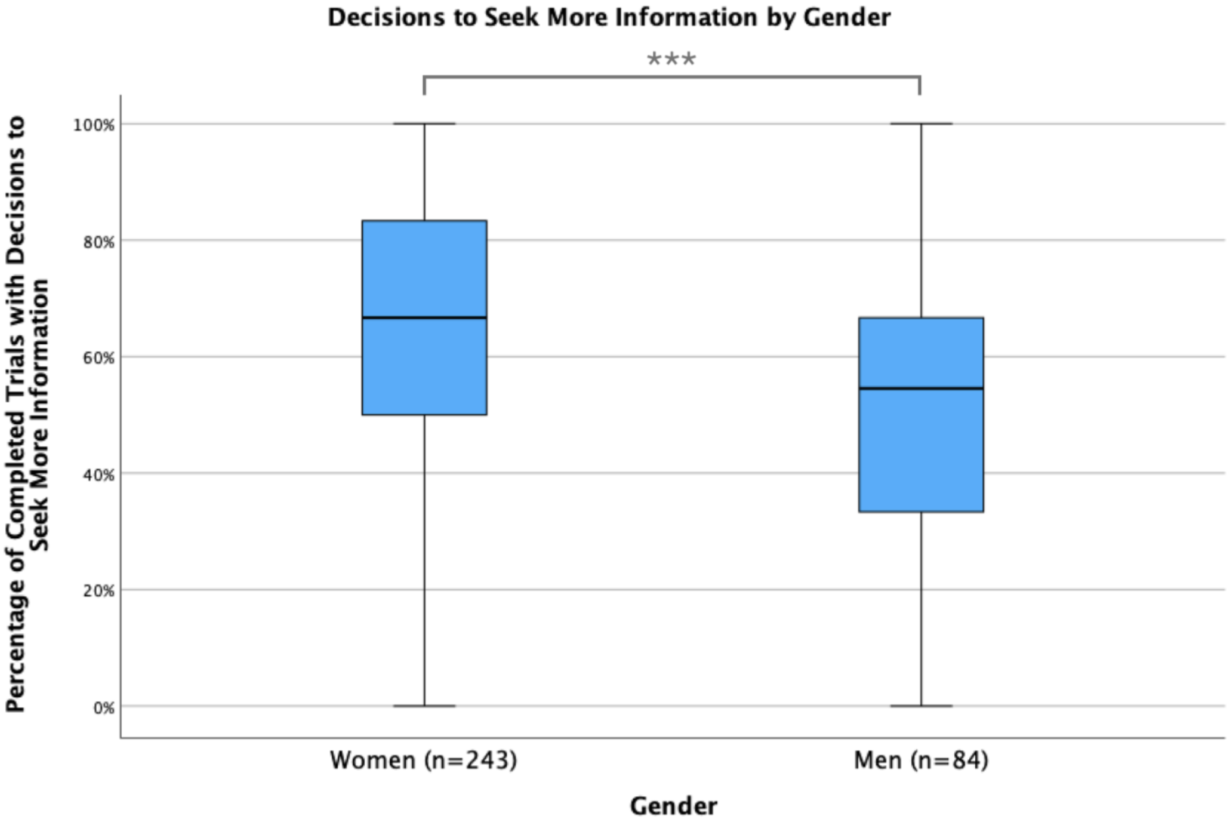


Figure 4-17: Women were more likely to make recommendations to seek more information than men, a statistically significant difference.

### Decisions to Seek More Information and GenderMag Facets

To test hypothesis H7, an independent-samples t-test was run to determine if there were differences in rates of deciding to seek more information between information processing styles. There were no outliers in the data, as assessed by inspection of a boxplot. Rates of deciding to seek more information for each information processing style were not normally distributed, as assessed by Shapiro-Wilk's test (*p* < .001), and there was homogeneity of variances, as assessed by Levene's test for equality of variances (*p* = .304). There were higher rates of deciding to seek more information in participants with Abi's comprehensive information processing style (*M* =

62.30%, *SD* = 26.06%) than in participants with Tim's selective information processing style (*M* = 60.42%, *SD* = 2%), a non-statistically significantly difference, *M* = 2.88%, 95% CI [-3.08%, 8.84%], t(331) = 0.950, *p* = .171, *d* = .107 (trivial effect size) (Figure 4-18).

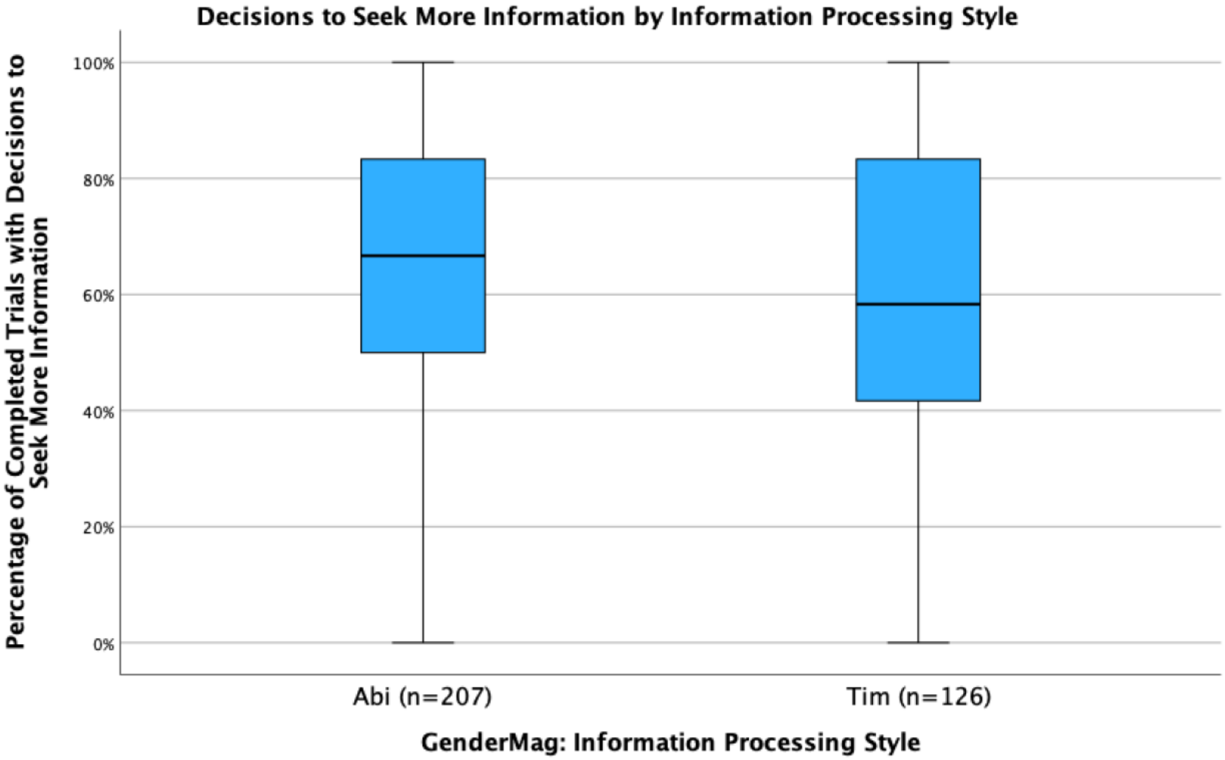


Figure 4-18: Participants with the comprehensive information processing style (Abi) and participants with the selective information processing style (Tim) did not significantly differ in their rates of deciding to seek more information.

An independent-samples t-test was run to determine if there were differences in rates of deciding to seek more information between participants with preferences for learning by process or by tinkering. There were no outliers in the data, as assessed by inspection of a boxplot. Rates of deciding to seek more information for these two groups were not normally distributed, as assessed by Shapiro-Wilk's test (*p* < .001), and there was homogeneity of variances, as assessed by Levene's test for equality of variances (*p* = .411). There were higher rates of participants who prefer to learn by process seeking more information (Abi, *M* = 66.15%, *SD* = 26.37%) than participants who prefer to learn by tinkering (Tim, *M* = 57.43%, *SD* = 26.10%), a statistically significantly difference, *M* = 8.72%, 95% CI [2.06%, 15.38%], *t*(239) = 2.578, *p* = .005, *d* = 0.332 (small effect size) (Figure 4-19). Independent-samples t-tests did not find statistically significant differences in rates of deciding to seek more information for the three other GenderMag facets: attitude toward risk, self-efficacy, and motivations.

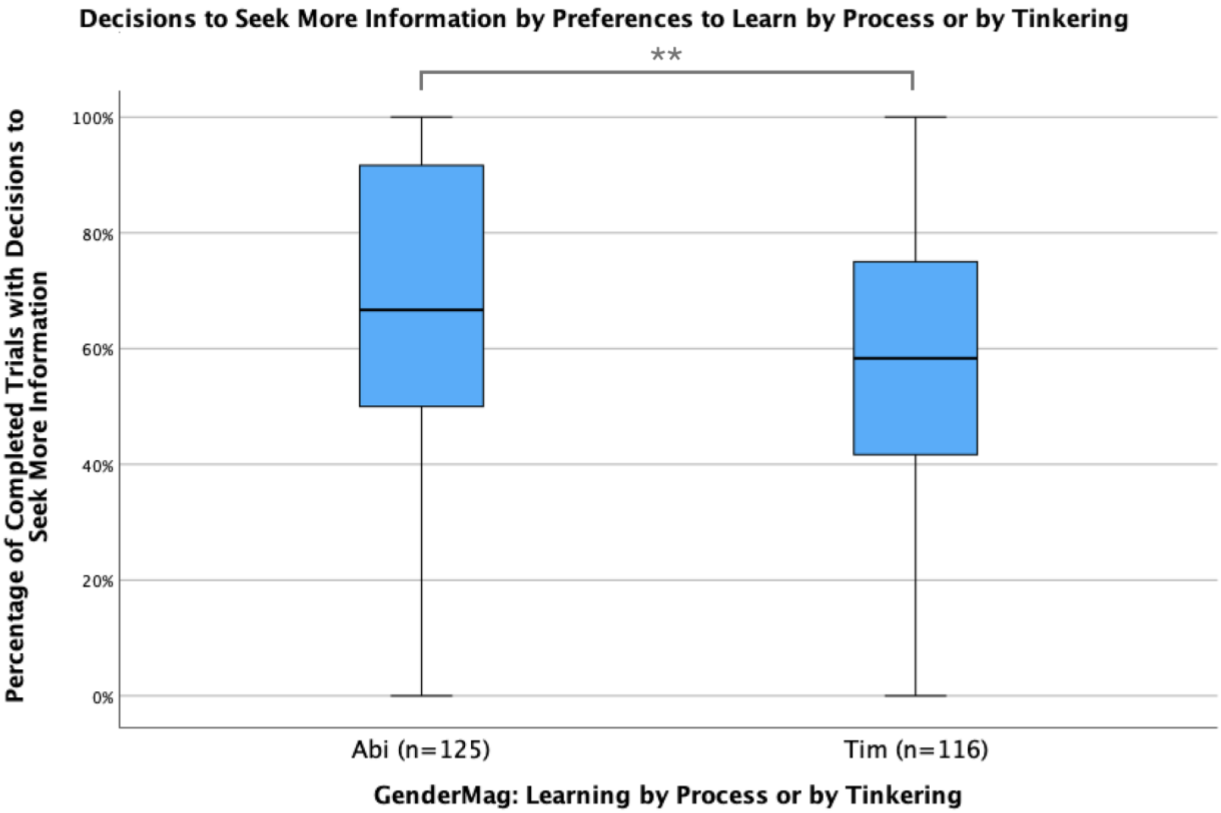


Figure 4-19: There were statistically significantly higher rates of deciding to seek more information for participants who prefer to learn by seeking more information (Abi) than participants who prefer to learn by tinkering (Tim).

### Decisions to Seek More Information and Knowledge about Managing Probabilistic Data

An independent-samples t-test was run to determine if there were differences in rates of deciding to seek more information between participants with more or less knowledge about managing probabilistic data. There were no outliers in the data, as assessed by inspection of a boxplot. Rates of deciding to seek more information for these two groups were not normally distributed, as assessed by Shapiro-Wilk's test (*p* < .001), and there was homogeneity of variances, as assessed by Levene's test for equality of variances (*p* = 0.745). There were higher rates of deciding to seek more information in participants with less knowledge about managing probabilistic data (*M* = 63.84%, *SD* = 26.41%) than participants with more knowledge about managing probabilistic data (*M* = 55.14%, *SD* = 27.29%), a statistically significantly difference, *M* = 8.70%, 95% CI [-0.143%, 17.54%], *t*(231) = 2.578, *p* = .027, *d* = .327 (small effect size) (Figure 4-20).

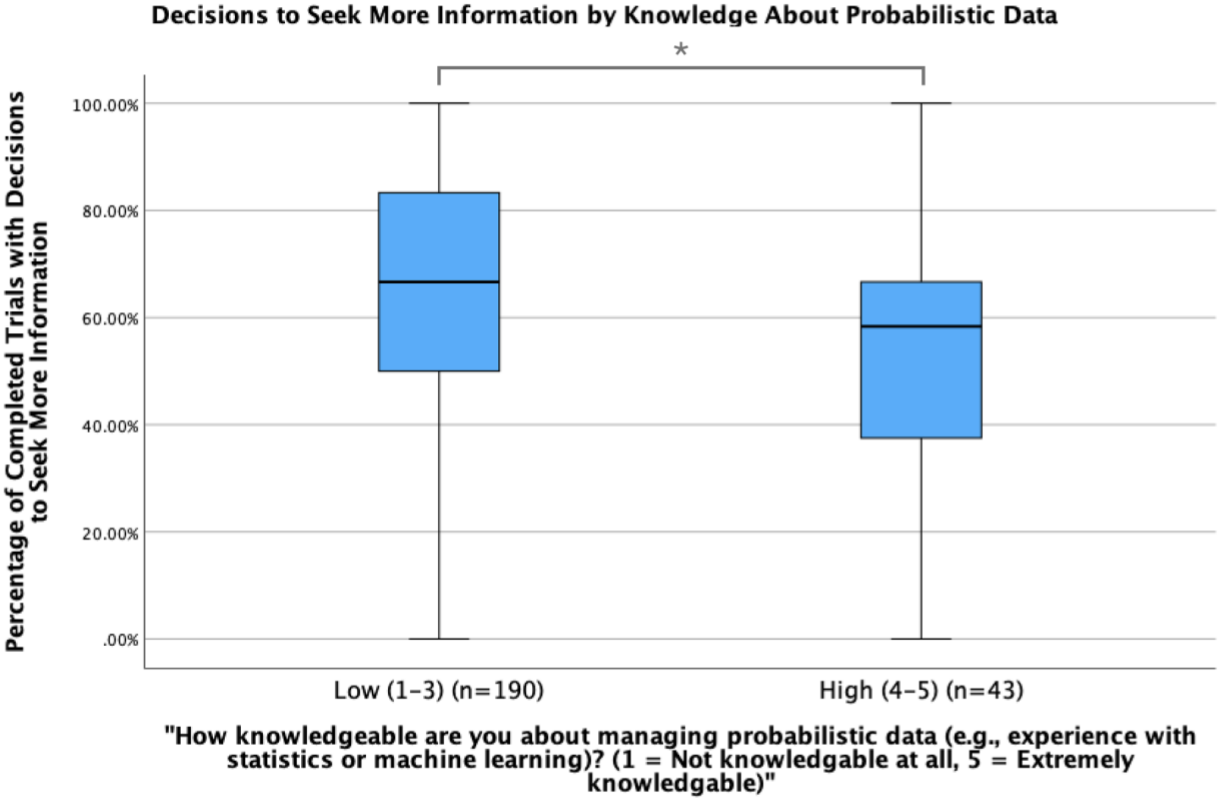


Figure 4-20: Participants with less knowledge about managing probabilistic data had statistically significantly higher rates of deciding to seek more information than participants with more knowledge about managing probabilistic data.

### Decisions to Seek More Information and Predominant Source of Uncertainty

H5 predicted that there would be a higher rate of deciding to seek more information when the uncertainty was depicted as being predominantly due to unavailable data (in half of the trials) than when uncertainty was depicted as being predominantly due to unpredictable data (in the other half of trials). A two-way mixed ANOVA was conducted to analyze the effect of predominant source of uncertainty on decisions to seek more information when the bar charts did or did not include uncertainty. Because the sources of uncertainty were only depicted in the condition with detailed ProdStar recommendations, this analysis included only data from Conditions 3 and 6. There were no outliers, as assessed by examination of studentized residuals for values greater than ±3. The data was not normally distributed, as assessed by Shapiro-Wilk's test of normality (*p* < .010). There was homogeneity of variances (*p* > .05) and covariances (*p* = .978), as assessed by Levene's test of homogeneity of variances and Box's M test, respectively. There was a statistically significant interaction between the source of information uncertainty and chart style on decisions to seek more information, *F*(1, 106) = 6.56, *p* = .012, partial $\eta^2$ = .058 (small effect size). When participants were presented with charts without uncertainty, they had higher rates of deciding to seek more information when the uncertainty was predominantly due to unavailable data (*M* = 65.99%, *SD* = 26.83%) than when the information uncertainty was predominantly due to unpredictable data (*M* = 55.86%, *SD* = 30.63%), a statistically significant difference, *M* = 10.12%, 95% CI [1.53%, 18.68%], *F*(1,53) = 343.64, *p* = .021, *d* = .866 (large effect size). When participants were presented with charts with uncertainty, they again had higher rates of deciding to seek more information when the uncertainty was predominantly due to unavailable data (*M* = 73.18%, *SD* = 27.85%) than when the information uncertainty was predominantly due to unpredictable data (*M* = 47.53%, *SD* = 32.28%), a difference that was also statistically significant, *M* = 25.65%, 95% CI [17.01%, 34.29%], *F*(1,53) = 298.78, *p* < .001, *d* = .849 (large effect size) (Figure 4-21).

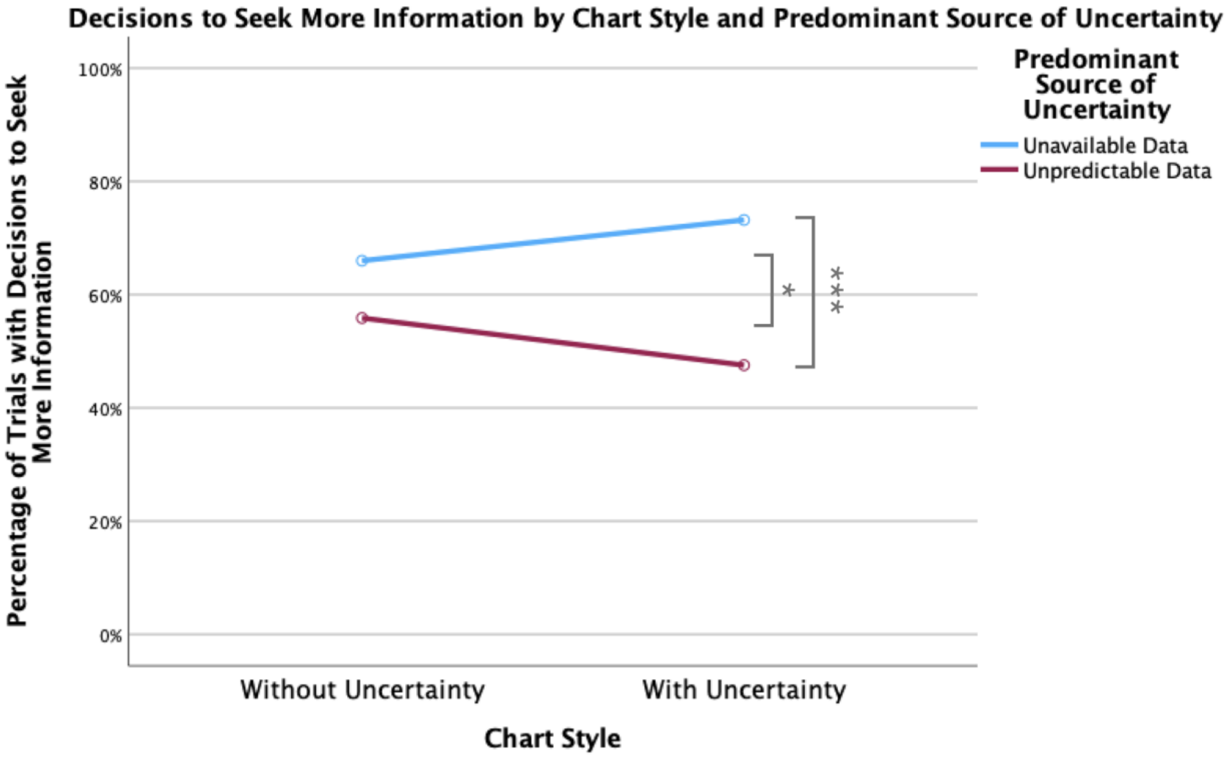


Figure 4-21: There was a statistically significant interaction between chart style and predominant source of uncertainty on decisions to seek more information. There were statistically significantly higher rates of deciding to seek more information when uncertainty was predominantly due to unavailable data than unpredictable data, for both chart styles, but to a greater extent in charts with the bars showing uncertainty.

Overall, rates of deciding to seek more information were higher when the information uncertainty was predominantly due to unavailable data (*M* = 69.58%, *SE* = 2.63%) than when it was predominantly due to unpredictable data (*M* = 51.70%, *SE* = 3.03%), a statistically significantly difference, *M* = 17.89%, 95% CI [11.88%, 23.90%], *F*(1,106) = 34.80, *p* < .001, partial $\eta^2$ = .247 (large effect size) (Figure 4-22).

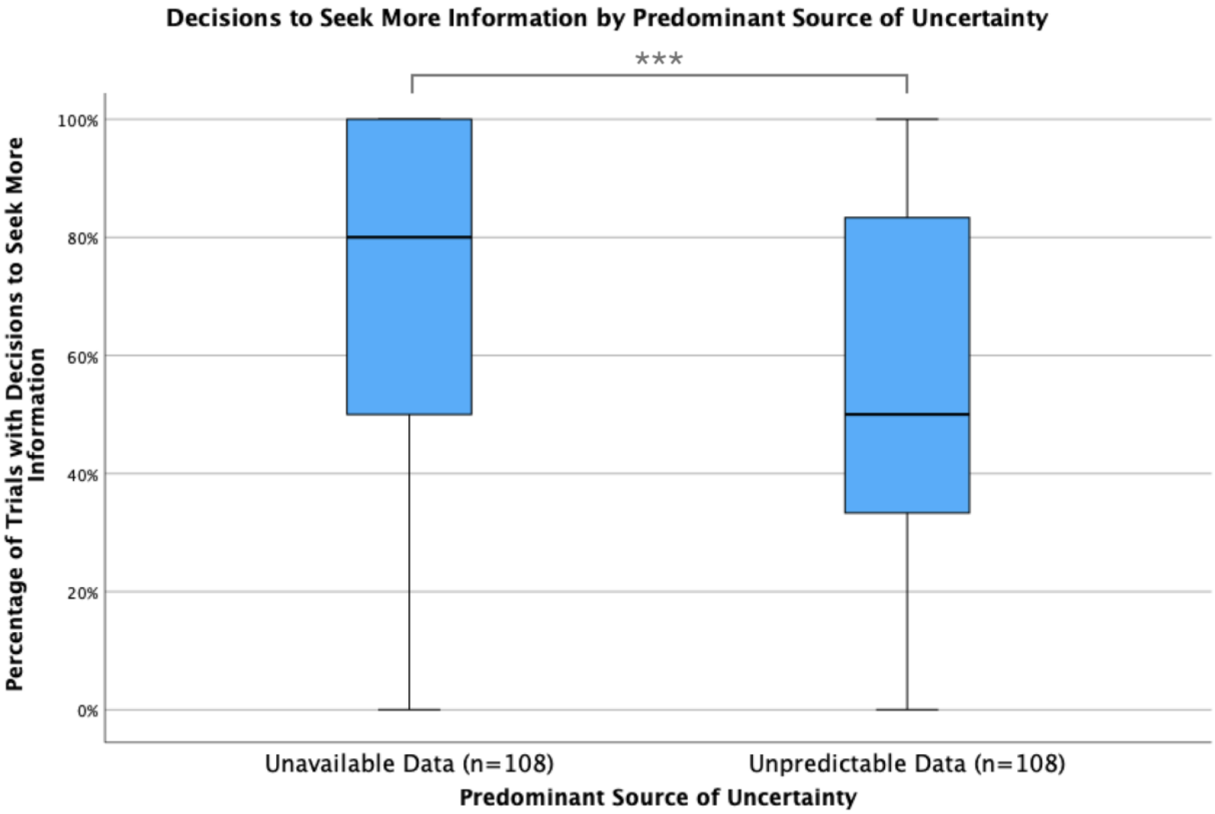


Figure 4-22: There were statistically significantly higher rates of deciding to seek more information when the uncertainty was depicted as being predominantly due to unavailable data than when it was depicted as being due to unpredictable data.

### Decisions to Seek More Information and Interaction Effects

A three-way ANOVA was run to check for interaction effects between chart style, recommendation style, and gender. There were ten outliers assessed as a value greater than 1.5 box-lengths from the edge of the box and zero extreme outliers assessed as a value greater than 3 box-lengths from the edge of the box. Mean rate of deciding to seek more information were normally distributed for seven groups (*p* > .05) and not normally distributed for eight groups (*p* < .05), as assessed by Shapiro-Wilk's test of normality. There was homogeneity of variances, as assessed by Levene's test for equality of variances, *p* = .157. There was no statistically significant three-way interaction between chart style, recommendation style, and gender *F*(2, 315) = 0.892, *p* = .411, $\eta^2$ = .006 (trivial effect size) (Figure 4-23). There was a statistically significant interaction between chart style and gender, *F*(2, 315) = 4.535, *p* = .034, $\eta^2$ = .014 (small effect size) (Figure 4-24). There was not a statistically significant interaction between chart style and recommendation style, *F*(2, 315) = 0.611, *p* = .534, $\eta^2$ = .004 (trivial effect size), or between recommendation style and gender, *F*(2, 315) = 2.043, *p* = .131, $\eta^2$ = .013 (small effect size)

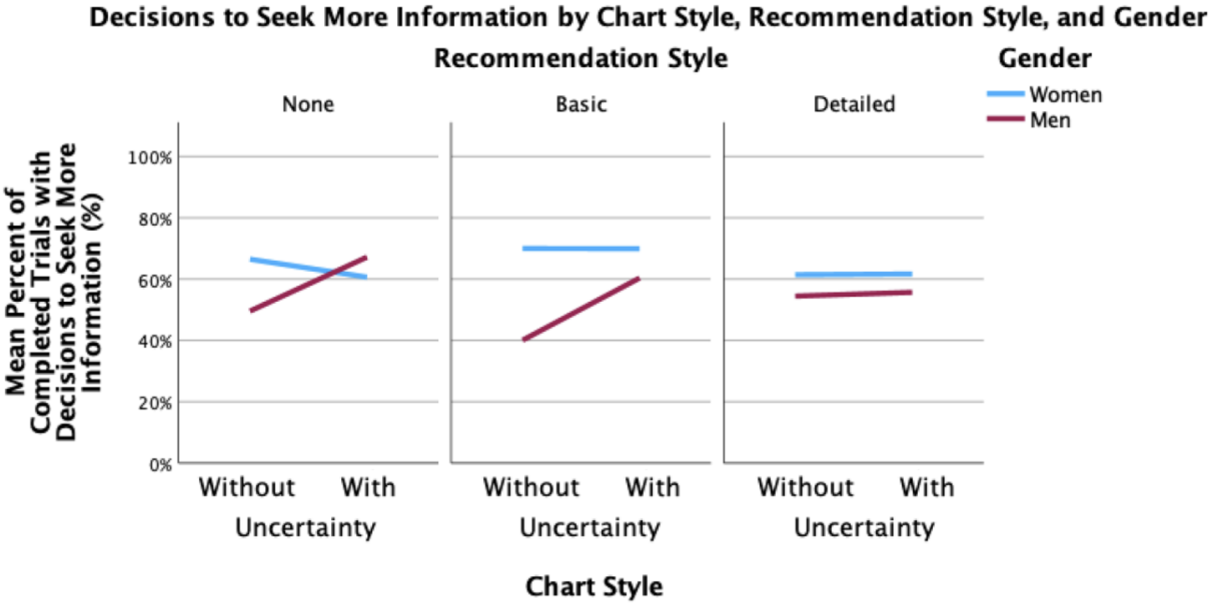


Figure 4-23: There was a statistically significant interaction between chart style and gender, but not between chart style and recommendation style or between gender and recommendation style.

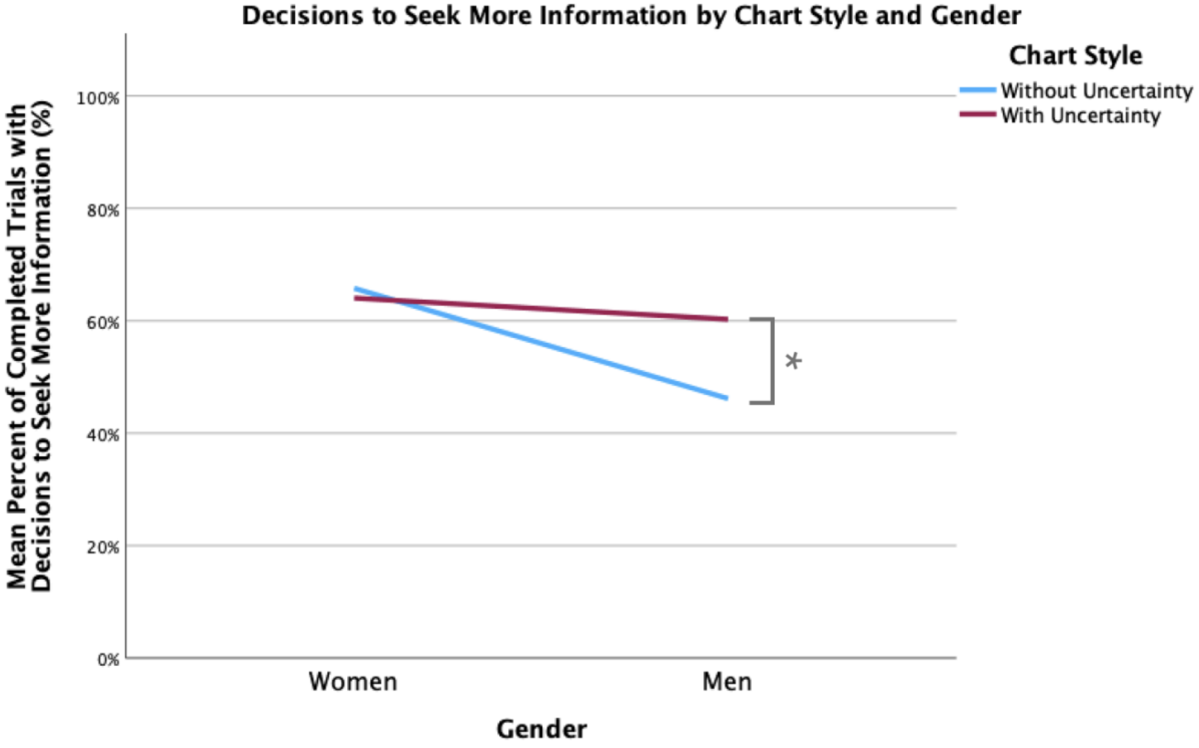


Figure 4-24: There was a statistically significantly interaction between chart style and gender on decisions to seek more information.

There was not a statistically significant simple two-way interaction between recommendation style and chart style for men, *F*(2, 78) = 0.832, *p* = .439, partial $\eta^2$ = .021 (small effect size) or for women, *F*(2,237) = 0.337, *p* = .714, $\eta^2$ = .003 (trivial effect size). All simple pairwise comparisons were made with a Bonferroni adjustment. Mean rate of deciding to seek more information was 46.15% (*SD* = 27.04%) for men who saw charts without uncertainty and 60.25% (*SD* = 25.86%) for men who saw charts with uncertainty, a statistically significant difference of 14.10%, 95% CI [2.81%, 25.39%], *p* = .015, partial $\eta^2$ = .018 (small effect size). Mean rate of deciding to seek more information was 65.77% (*SD* = 25.03%) for women who saw charts without uncertainty and 64.03% (*SD* = 27.38%) for women who saw charts with uncertainty, a non-statistically significance difference of 1.74%, 95% CI [-4,89, 8.38], *p* = .605, $\eta^2$ = .001 (trivial effect size) (Figure 4-25).

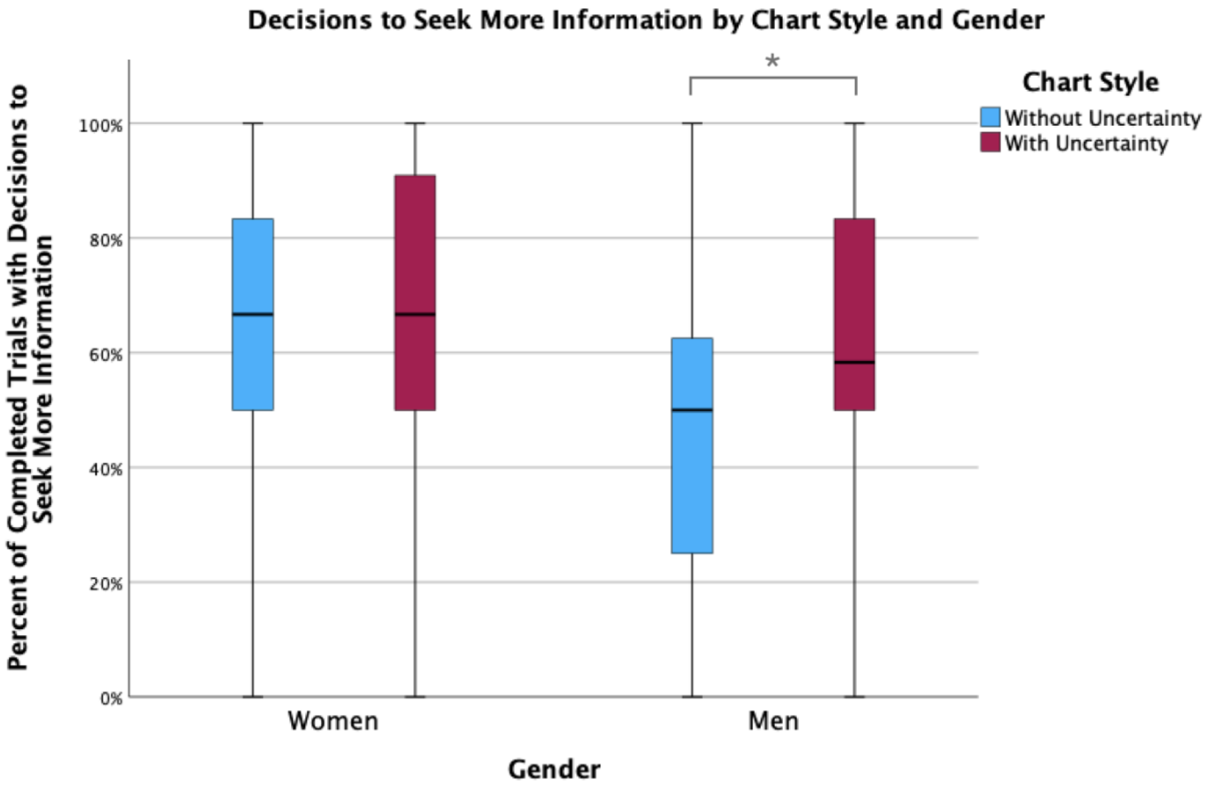


Figure 4-25: There was a statistically significantly lower rate of deciding to seek more information for when presented with charts without uncertainty than with uncertainty for men, but not for women.

Two-way mixed ANOVAs were also conducted to check for interaction effects between sources of uncertainty, gender, and knowledge of managing probabilistic data on decisions to seek more information. There was no statistically significant interaction between gender and sources of uncertainty on decisions to seek more information, $F(1,105) = 3.277$, $p = .073$, partial $\eta^2 = .031$ (small effect size). There was no statistically significant interaction between knowledge of managing probabilistic data and sources of uncertainty, $F(1,92) = 0.312$, $p = .578$, partial $\eta^2 = .003$ (trivial effect size).

### Decision Confidence

For each of the twelve product decision trials, participants were asked to rate their decision: “You selected Product [A, B]. How confident are you that you selected the optimal product?” This was a Likert style question, with possible responses ranging from 1 (“Not confident at all”) to 5 (“Very confident”). Decision confidence was calculated as the mean confidence across the twelve trials. The differences in decision confidence between chart style, recommendation style, gender, GenderMag facets, and knowledge of probabilistic data were analyzed with independent-samples t-tests (if two levels of independent variable) or one-way ANOVA (if three levels of independent variable).

### Decision Confidence and Chart Style

An independent-samples t-test was run to determine if there were differences in mean decision confidence between charts without uncertainty and charts with uncertainty. There were eleven outliers, assessed as a value greater than 1.5 box-lengths from the edge of the box, and one extreme outlier, assessed as a value greater than 3 box-lengths from the edge of the box. The data was not normally distributed, as assessed by Shapiro-Wilk’s test (*p* < .001), and there was homogeneity of variances, as assessed by Levene's test for equality of variances (*p* = .550). There was higher mean decision confidence with charts without uncertainty (*M* = 3.60 *SD* = 0.67) than charts with uncertainty (*M* = 3.42, *SD* = 0.63), a difference that was statistically significantly different, *M* = 0.174, 95% CI [0.0361, 0.312], *t*(344) = 2.482, *p* = .007, *d* = .267 (small effect size) (Figure 4-26).

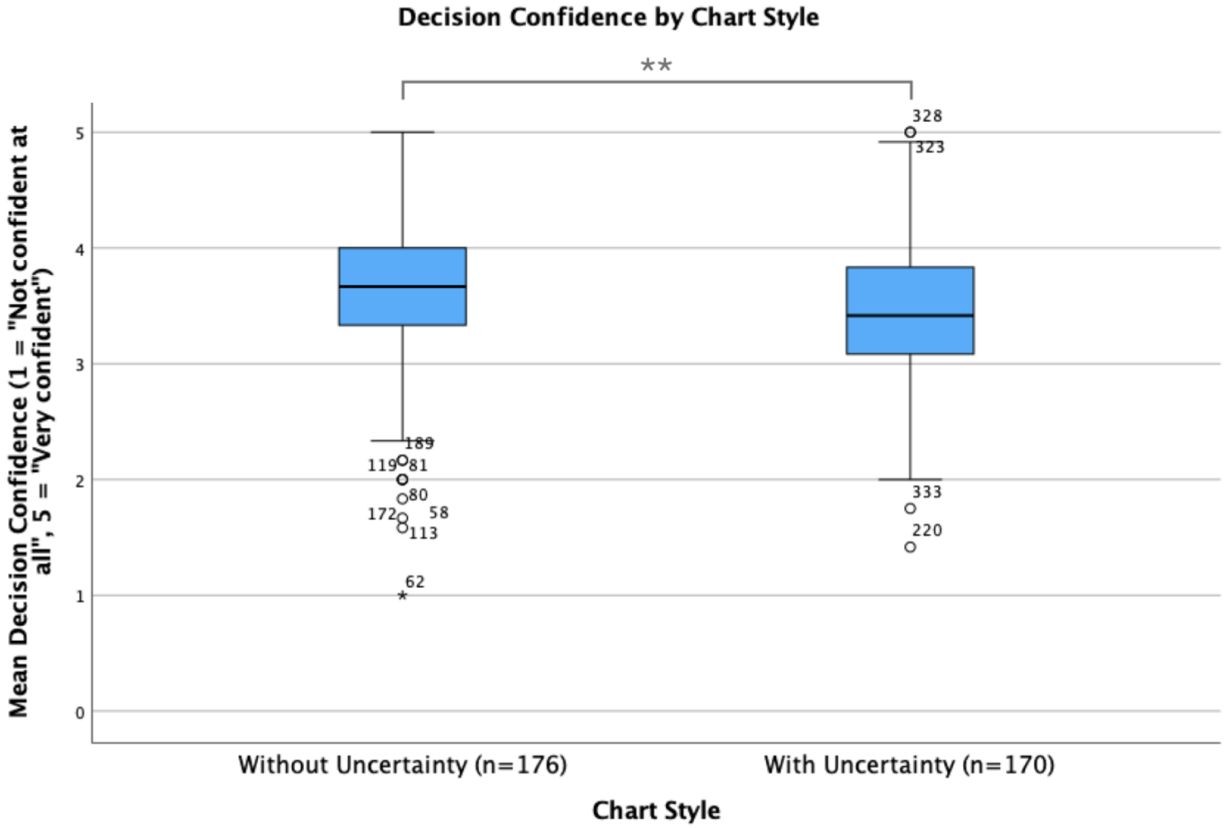


Figure 4-26: Participants reported statistically significantly higher decision confidence when provided with charts without uncertainty than charts with uncertainty.

### Decision Confidence and Recommendation Style

A one-way ANOVA was run to determine if there were differences in mean decision confidence between recommendation styles. There were seven outliers, assessed as a value greater than 1.5 box-lengths from the edge of the box. Decision confidence was normally distributed for the basic and detailed recommendation styles ($p$ > .05), but not for the group with no recommendations ($p$ < .001). There was not homogeneity of variances, as assessed by Levene’s test for equality of variances ($p$ = .027). Decision confidence increased from the group with no recommendations ($M$ = 3.50, $SD$ = 0.70), to the group with detailed recommendations ($M$ = 3.51, $SD$ = 0.63), to the group with basic recommendations ($M$ = 3.52, $SD$ = 0.64), but the differences between recommendation styles was not statistically significant, *F*(2, 343) = 0.00., *p* = .978, partial $\eta^2$ = .000 (no effect size) (Figure 4-27).

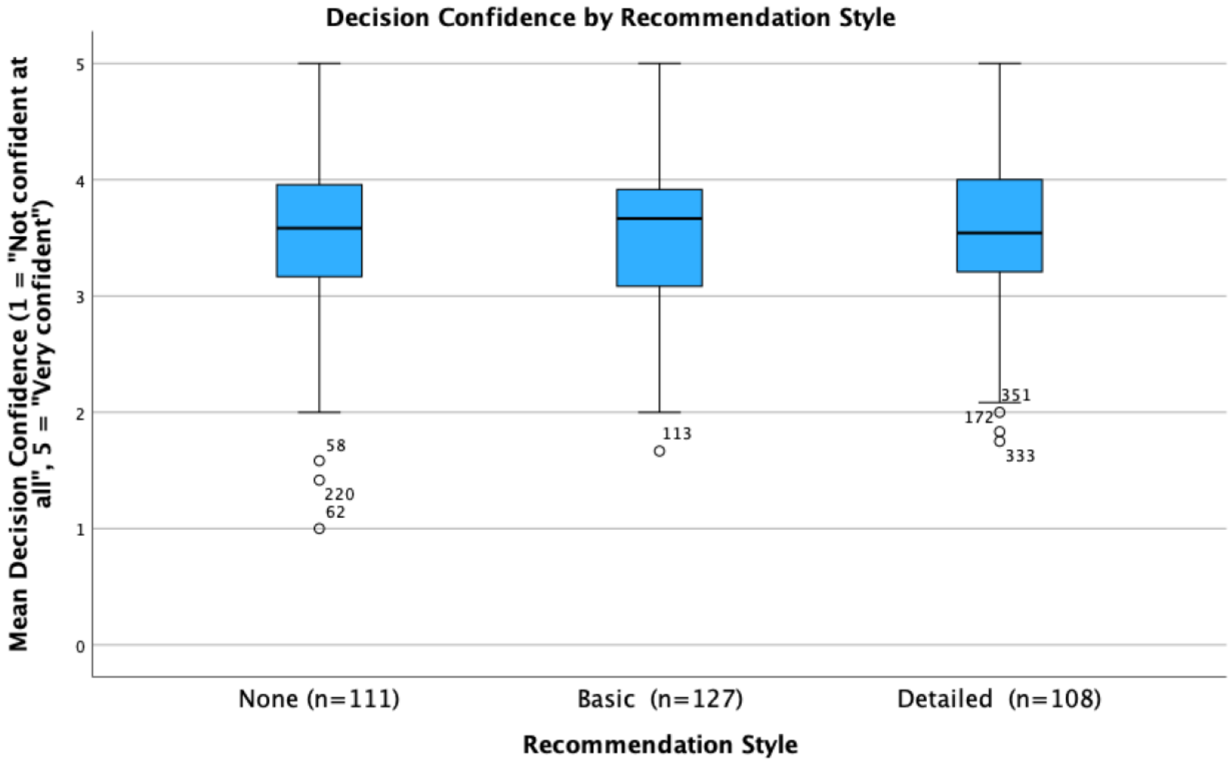


Figure 4-27: No statistically significant differences in decision confidence were found between the recommendation styles.

### Decision Confidence and Gender

An independent-samples t-test was run to determine if there were differences in mean decision confidence between women and men. There were eight outliers, assessed as a value greater than 1.5 box-lengths from the edge of the box, and one extreme outlier. The data was not normally distributed, as assessed by Shapiro-Wilk’s test (*p* < .001), and there was homogeneity of variances, as assessed by Levene's test for equality of variances (*p* = .131). There was lower mean decision confidence for women (*M* = 3.43 *SD* = 0.66) than men (*M* = 3.76, *SD* = 0.57), a difference that was statistically significantly different, *M* = -0.413, 95% CI [-0.594, -0.233], t(325) = -3.976, *p* < .001, *d* = -.503 (medium effect size) (Figure 4-28).

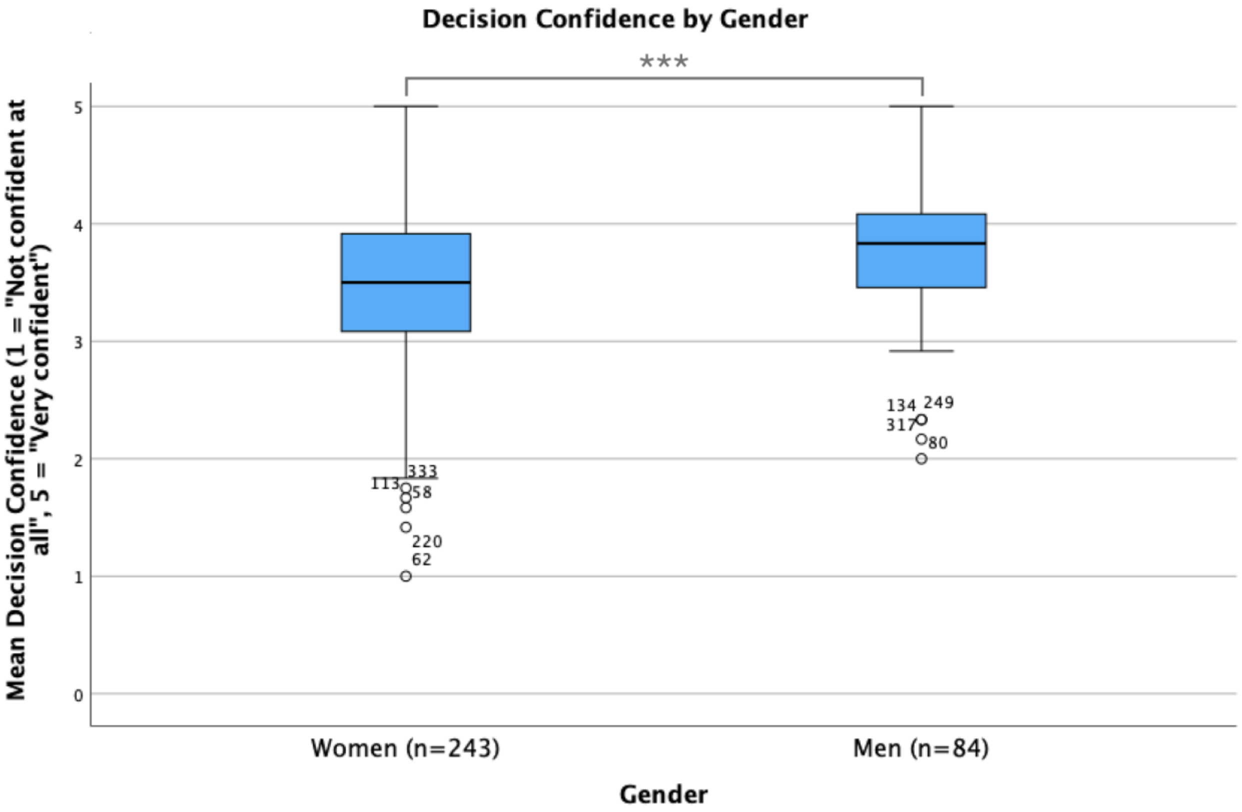


Figure 4-28: Mean decision confidence was lower for women than men, a statistically significant difference.

## Decision Confidence and GenderMag Facets

Independent-samples t-tests were performed to check for group differences in decision confidence for any of the five GenderMag facets. The number of participants in each group varied based on participants' completion of GenderMag questions (e.g., fewer participants completed the questions about attitude toward risk than questions about self-efficacy).

### *Attitude Toward Risk*

An independent-samples t-test was run to determine if there were differences in decision confidence between participants with lower and higher tolerance of risk. There were three outliers in the data, as assessed by inspection of a boxplot. Decision confidence was normally distributed for participants with the Tim facet, as assessed by Shapiro-Wilk's test ($p$ = .060), but not for participants with the Abi facet (*p* = .001). There was homogeneity of variances, as assessed by Levene's test for equality of variances (*p* = .200). Mean decision confidence not statistically significantly different between Abi’s lower risk tolerance (*M* = 3.50, *SD* = 0.622) and Tim’s higher risk tolerance (*M* = 3.39, *SD* = 0.728), *t*(168) = 0.984, *p* = .163, *d* = .153 (trivial effect size) (Figure 4-29).

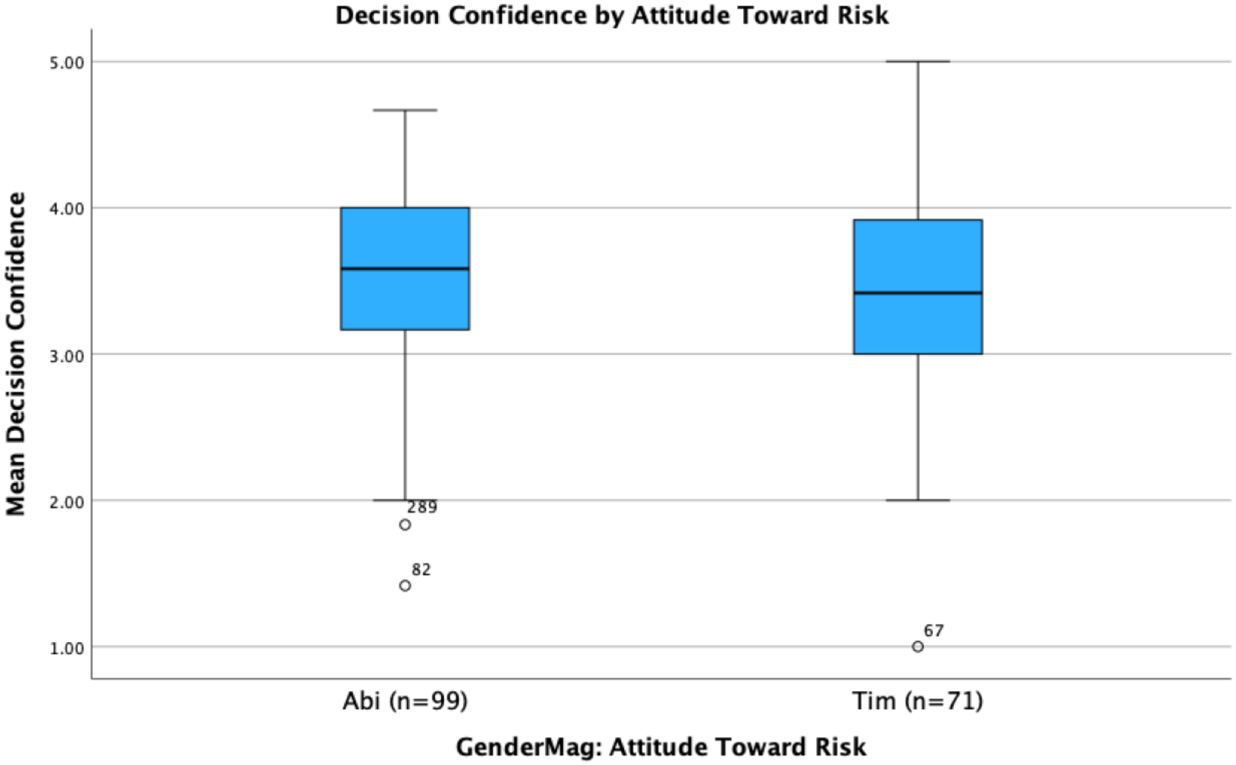


Figure 4-29: Decision confidence was not statistically significantly different between participants with a lower tolerance for risk (Abi) and participants with a higher tolerance for risk (Tim).

*Self-Efficacy*

An independent-samples t-test was run to determine if there were differences in decision confidence between participants with low and high self-efficacy. There were four outliers in the data, as assessed by inspection of a boxplot. Decision confidence was not normally distributed, as assessed by Shapiro-Wilk's test (*p* < .05), and there was homogeneity of variances, as assessed by Levene's test for equality of variances (*p* = .966). Mean decision confidence was statistically significantly lower for participants with Abi's lower risk self-efficacy (*M* = 3.37, *SD* = 0.642) than participants with Tim's higher self-efficacy (*M* = 3.52, *SD* = 0.639), *t*(281) = -1.968, *p* = .025, *d* = .235 (small effect size) (Figure 4-30).

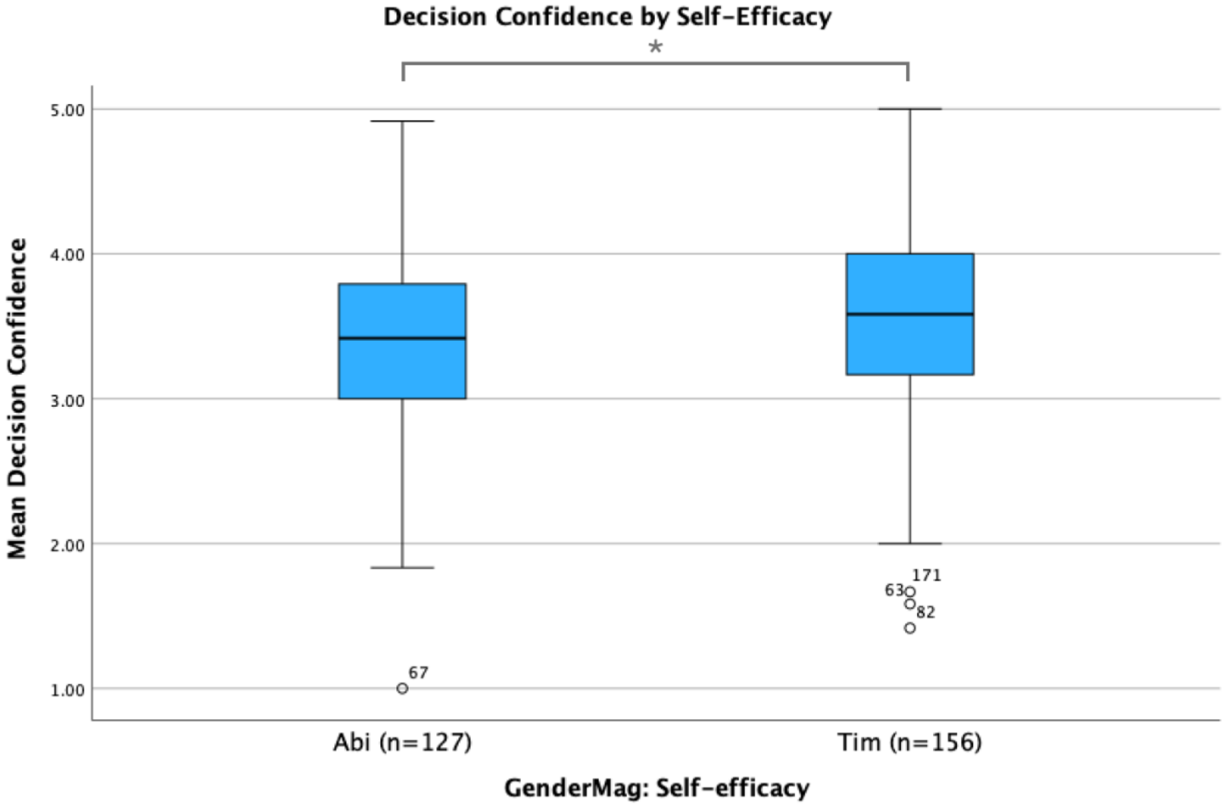


Figure 4-30: Mean decision confidence was statistically significantly lower for participants with low self-efficacy (Abi) than high self-efficacy (Tim).

*Information Processing Style*

An independent-samples t-test was run to determine if there were differences in decision confidence between information processing styles. There were six outliers in the data, as assessed by inspection of a boxplot. Decision confidence was not normally distributed, as assessed by Shapiro-Wilk's test (*p* < .05), and there was homogeneity of variances, as assessed by Levene's test for equality of variances (*p* = .350). Mean decision confidence was statistically significantly higher for participants with Abi's comprehensive information processing style (*M* =

3.57, *SD* = 0.620) than participants with Tim’s selective information processing style (*M* = 3.45, *SD* = 0.684), *t*(331) = 1.708, *p* = .044, *d* = .196 (trivial effect size) (Figure 4-31).

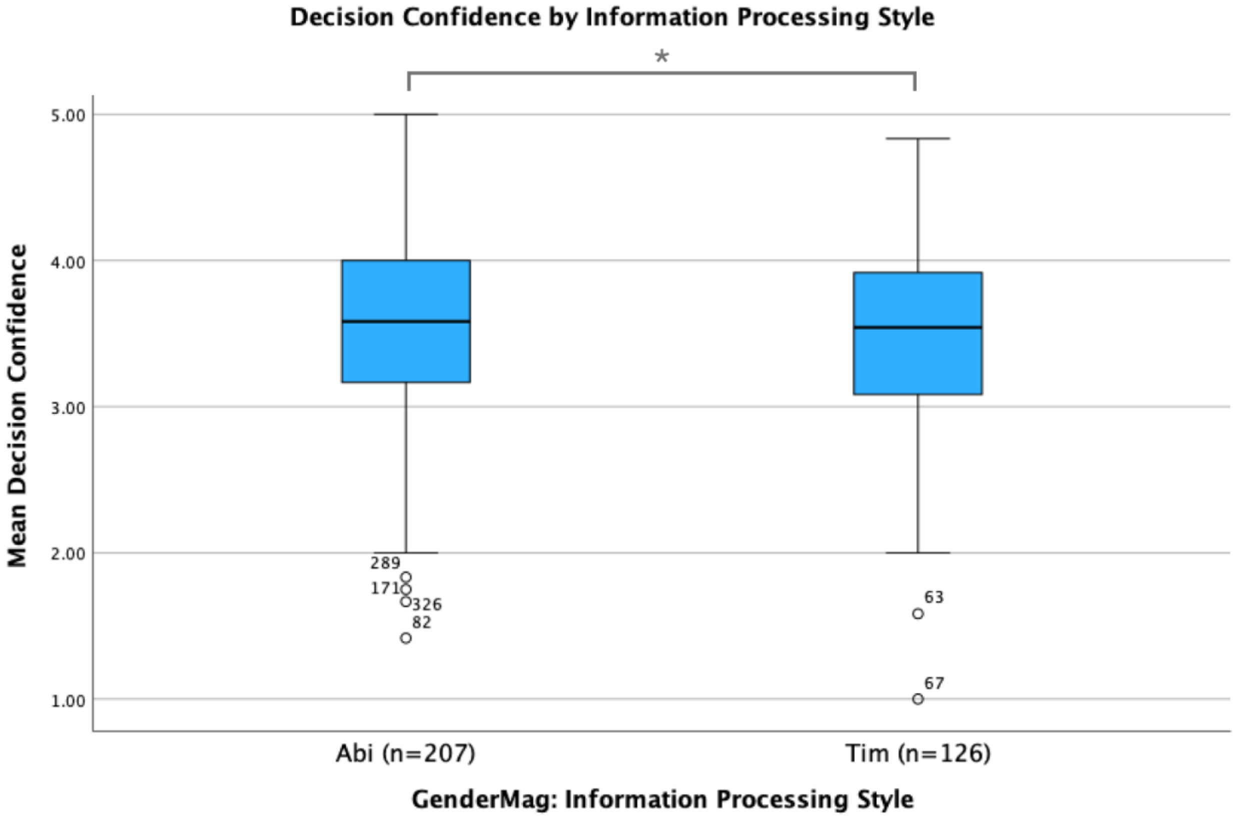


Figure 4-31: Mean decision confidence was statistically significantly higher for participants with a comprehensive information processing style (Abi) than participants with a selective information processing style (Tim).

*Motivations*

An independent-samples t-test was run to determine if there were differences in decision confidence between participants with different motivations for using technology. There were nine outliers in the data, as assessed by inspection of a boxplot. Decision confidence was not normally distributed, as assessed by Shapiro-Wilk's test (*p* < .05), and there was homogeneity of variances, as assessed by Levene's test for equality of variances (*p* = .759). Mean decision confidence was statistically significantly lower for participants with Abi’s motivation of using technology to accomplish a goal (*M* = 3.41, *SD* = 0.638) than with participants with Tim’s motivation of using technology for technology's sake (*M* = 3.61, *SD* = 0.642), *t*(239) = -2.401, *p* = .009, *d* = -.310 (small effect size) (Figure 4-32).

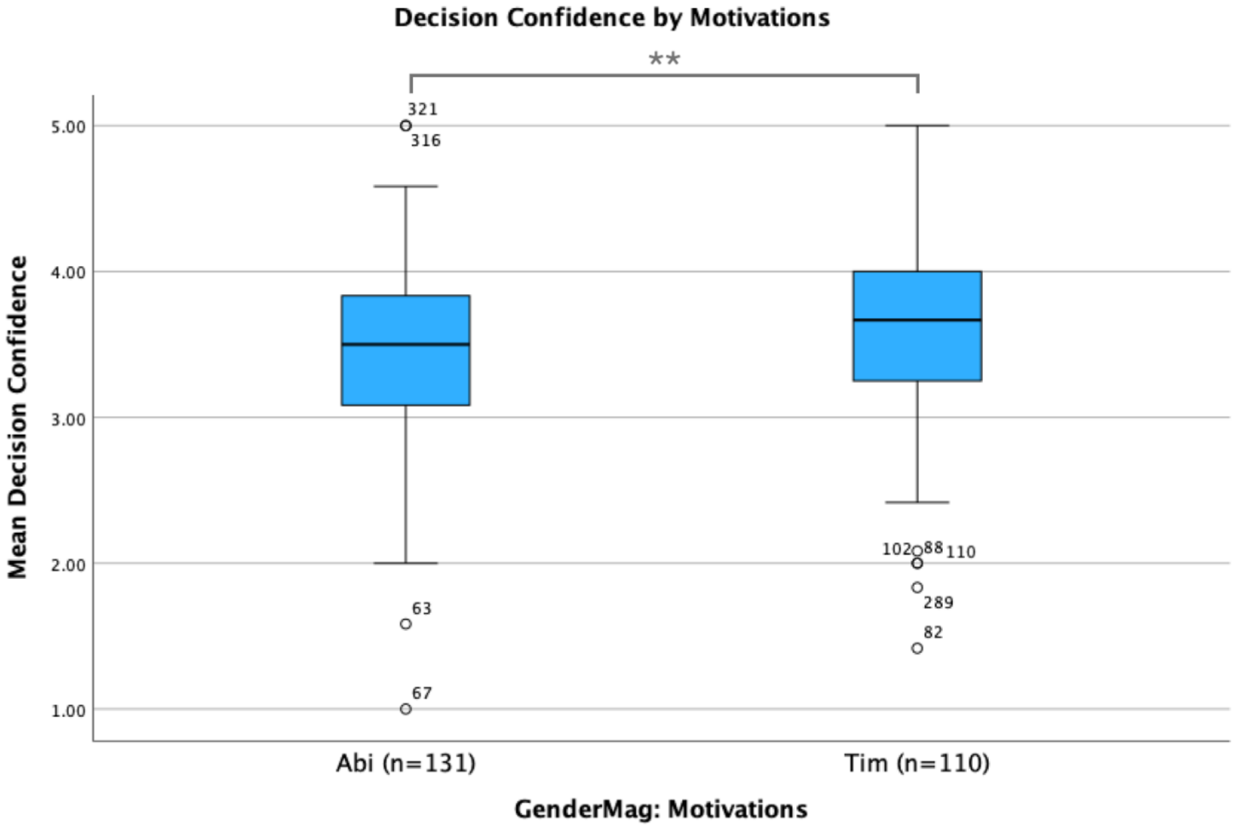


Figure 4-32: Mean decision confidence was statistically significantly lower for participants who are motivated to use technology to accomplish a goal (Abi) than participants who are motivated to use technology for its own sake (Tim).

*Learning by Process or by Tinkering*

An independent-samples t-test was run to determine if there were differences in decision confidence between participants who learn by process or by tinkering. There were seven outliers in the data, as assessed by inspection of a boxplot. Decision confidence was not normally distributed, as assessed by Shapiro-Wilk's test (*p* < .05), and there was homogeneity of variances, as assessed by Levene's test for equality of variances (*p* = .626). Mean decision confidence was statistically significantly lower for participants with Abi's preference to learn by process (*M* =

3.42, *SD* = 0.634) than with participants with and Tim's preference to learn by tinkering (*M* = 3.60, *SD* = 0.650), *t*(239) = -2.130, *p* = .017, *d* = -.275 (small effect size) (Figure 4-33).

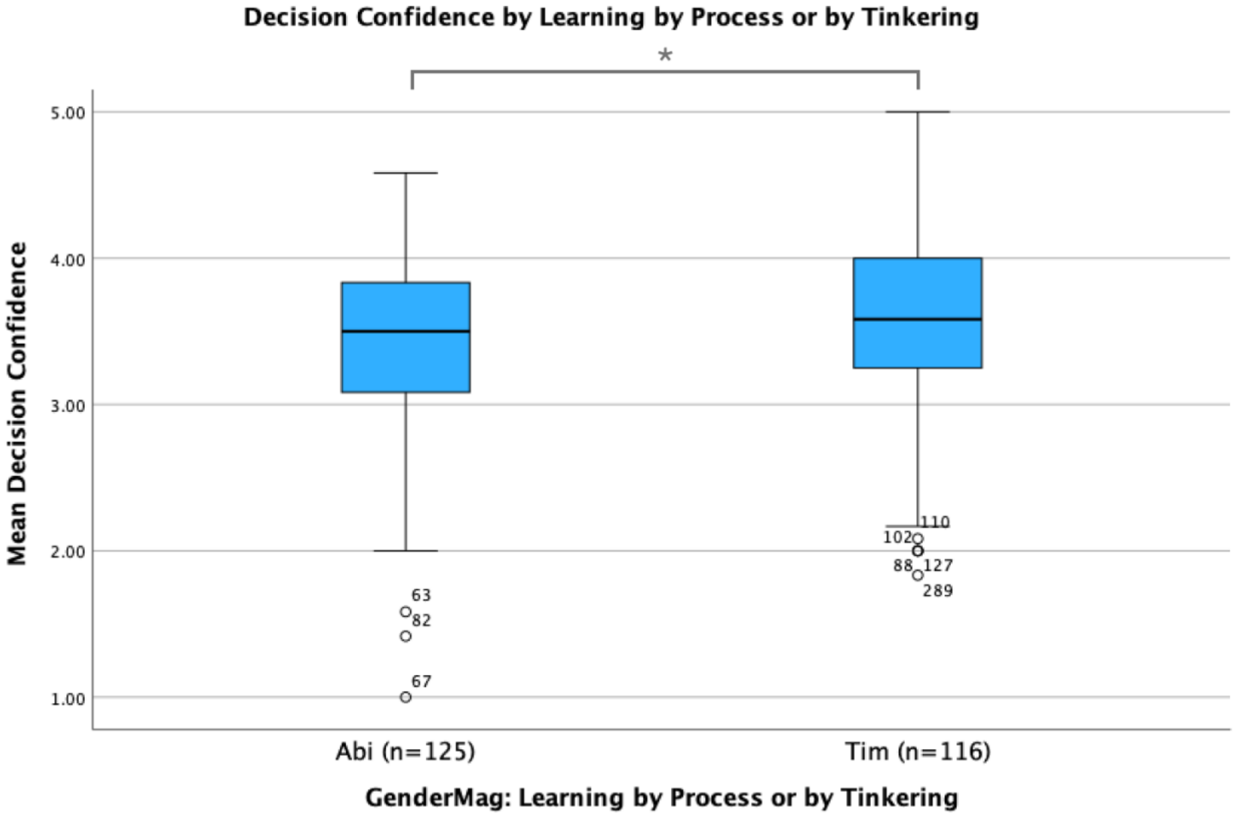


Figure 4-33: Mean decision confidence was statistically significantly lower for participants with a preference for learning by process (Abi) than participants with a preference for learning by tinkering (Tim).

### Decision Confidence and Knowledge about Managing Probabilistic Data

An independent-samples t-test was run to determine if there were differences in decision confidence participants with more or less knowledge about managing probabilistic data. There were ten outliers in the data, as assessed by inspection of a boxplot. Decision confidence was not normally distributed, as assessed by Shapiro-Wilk's test for participants with less knowledge about managing probabilistic data (*p* < .001), but it was normally distributed for participants with more knowledge about managing probabilistic data (*p* = .086). There was homogeneity of variances, as assessed by Levene's test for equality of variances (*p* = .382). Mean decision confidence was statistically significantly lower for participants with less knowledge about managing probabilistic data (*M* = 3.49, *SD* = 0.631) than for participants with more knowledge about managing probabilistic data (*M* = 3.68, *SD* = 0.691), *t*(308) = -2.059, *p* = .020, *d* = -.298 (small effect size) (Figure 4-34).

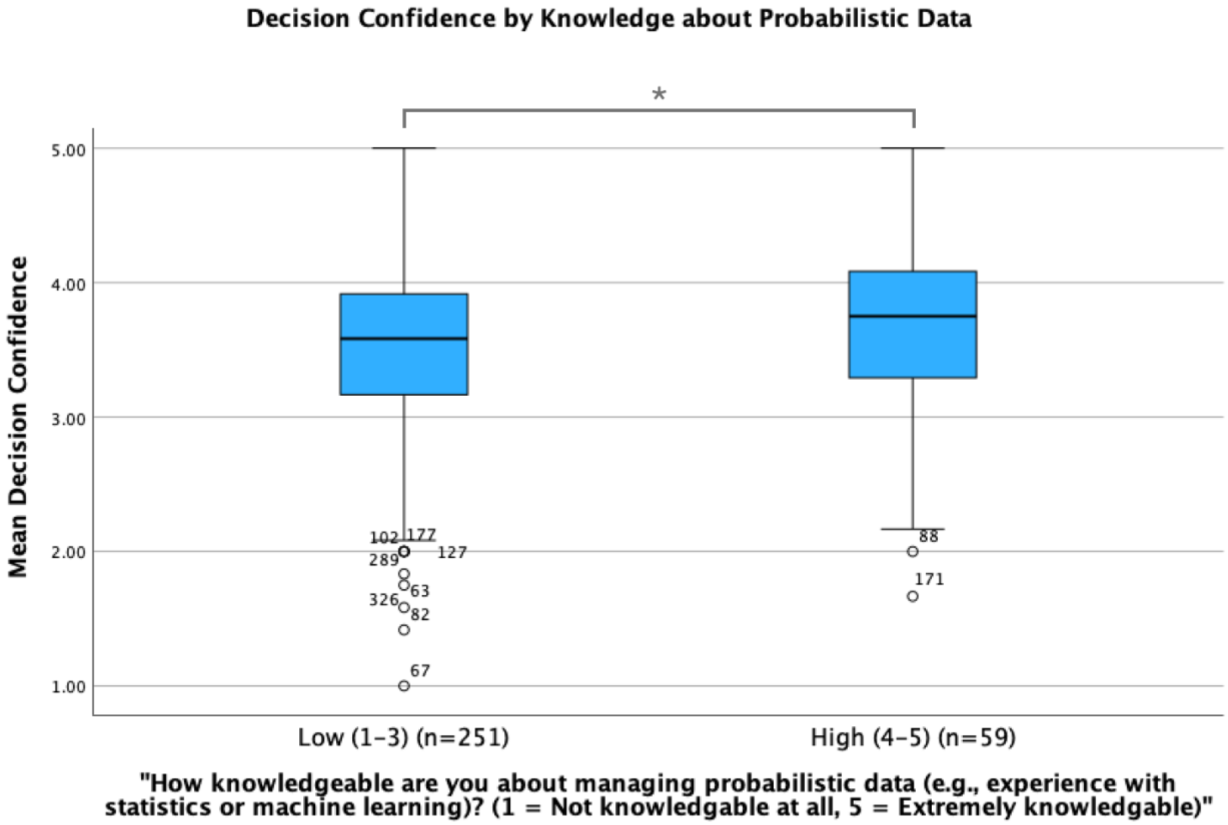


Figure 4-34: Participants with less knowledge about managing probabilistic data had statistically significantly lower mean decision confidence than participants with more knowledge about managing probabilistic data.

### Trust of Recommendations

Trust of recommendation accuracy was measured by responses to a 1-5 Likert-style question in the questionnaire: “How much did you trust the recommendations to be accurate?” This dependent variable was treated as an ordinal variable, so it was evaluated by Mann-Whitney U tests. This analysis tested H2: The detailed recommendation will be associated with higher trust than the basic recommendation. Because this question was only presented to participants who saw a ProdStar recommendation, this analysis excludes participants in Conditions 1 and 4, in which no recommendation was provided.  Results are presented below for Mann-Whitney U tests comparing reported trust of recommendations and recommendation style, chart style, gender, and knowledge of probabilistic data. No statistically significant relationships between the GenderMag facets and reported trust of recommendations were identified.

### Trust of Recommendations and Chart Style

A Mann-Whitney U test was run to determine if there were differences in trust of recommendations between charts without uncertainty displayed and charts with uncertainty displayed. Distributions of the reported trust of recommendations with charts that did or did not include uncertainty were similar, as assessed by visual inspection. Median reported trust was not statistically significantly different between charts without uncertainty (*Mdn* = 3.00) and with uncertainty (*Mdn* = 3.00), *U* = 6610, *z* = 0.862, *p* = .389, *r* = -.0557 (trivial effect size) (Figure 4-35).

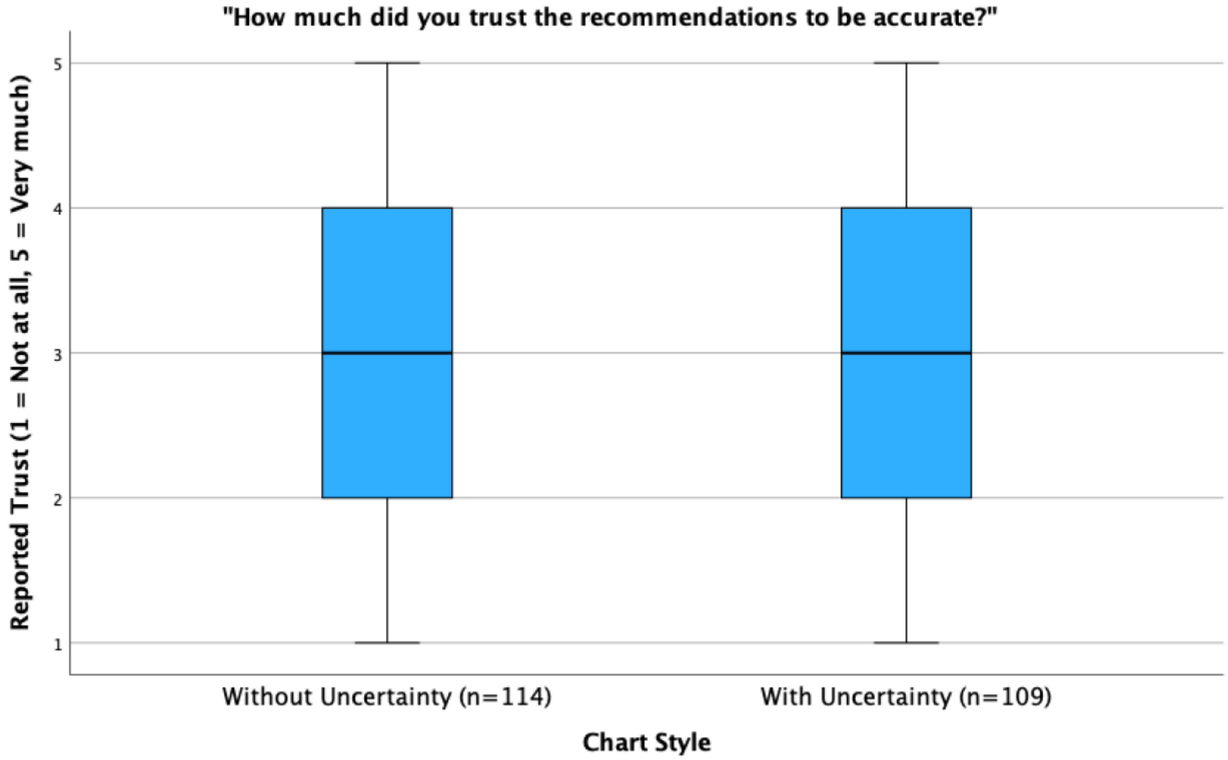


Figure 4-35: No statistically significant difference was found between reported trust of the recommendations was found between charts without uncertainty and charts with uncertainty.

### Trust of Recommendations and Recommendation Style

A Mann-Whitney U test was run to determine if there were differences in reported trust of the recommendations between the two recommendation styles. Distributions of the reported trust of the recommendations for the basic recommendation and the detailed recommendation were not similar, as assessed by visual inspection. Reported trust of the detailed recommendation (mean rank = 124.68) was statistically significantly higher than trust of the basic recommendation (mean rank = 100.51), *U* = 7545, *z* = 2.921, *p* = .003, *r* = .196 (small effect size) (Figure 4-36).

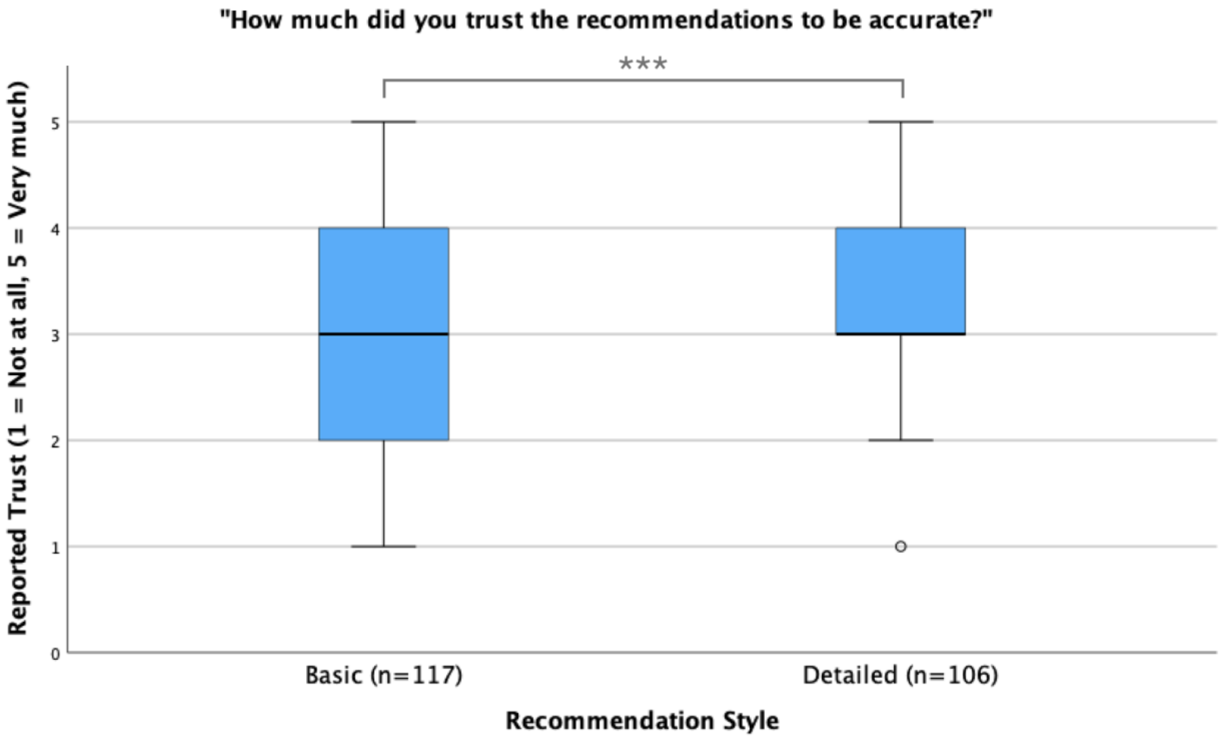


Figure 4-36: Reported trust of the detailed recommendation was statistically significantly higher than trust of the basic recommendation.

### Trust of Recommendations and Gender

A Mann-Whitney U test was run to determine if there were differences in trust of recommendations between men and women. Distributions of the reported trust of recommendations for men and women were similar, as assessed by visual inspection. Median reported trust was not statistically significantly different between men (*Mdn* = 3.00) and women (*Mdn* = 3.00), *U* = 4304, *z* = -0.411, *p* = .967, *r* = -.003 (trivial effect size) (Figure 4-37).

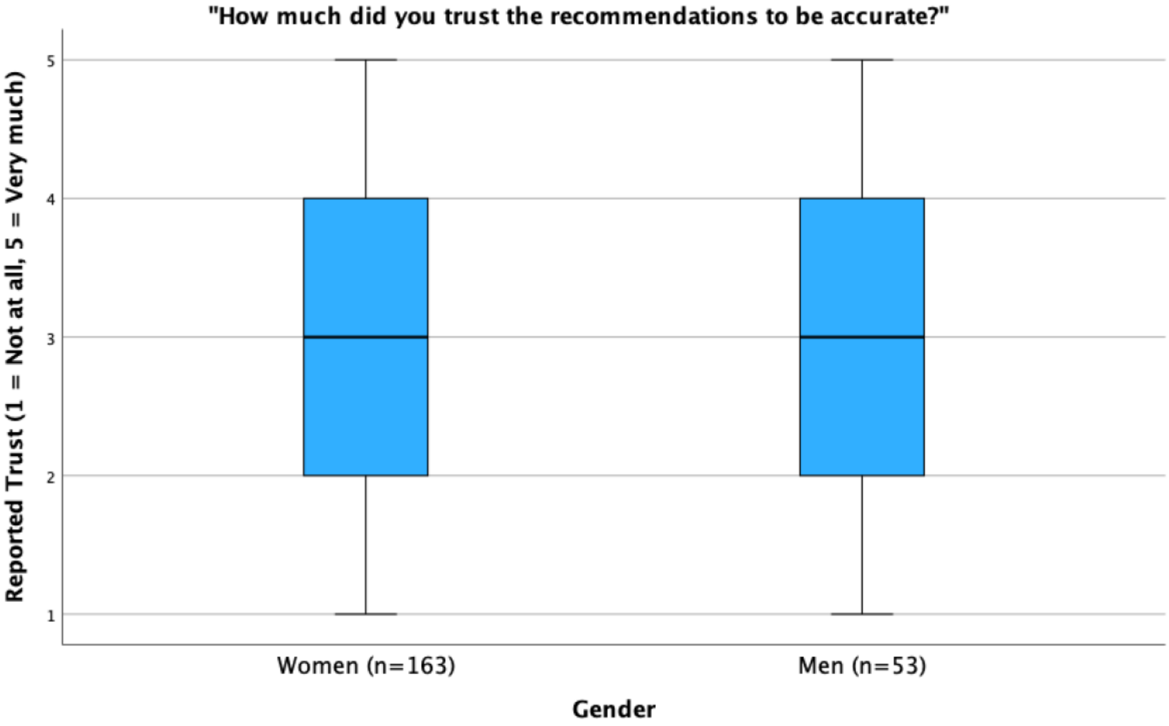


Figure 4-37: No statistically significant differences in reported trust of the recommendations were found between women and men.

### Trust of Recommendations and Knowledge About Managing Probabilistic Data

A Mann-Whitney U test was run to determine if there were differences in reported trust of the recommendations between participants with low (1-3 on Likert scale) and high (4-5 on Likert scale) knowledge of managing probabilistic data. Distributions of the reported trust of recommendations for these two groups were not similar, as assessed by visual inspection. Reported trust of the recommendations for participants with low knowledge of probabilistic data (mean rank = 103.31) were statistically significantly higher than for participants with high knowledge of probabilistic data (mean rank = 81.10), $U$ = 2162.500, $z$ = -2.150, $p$ = .032, $r$ = .153 (a small effect size) (Figure 4-38).

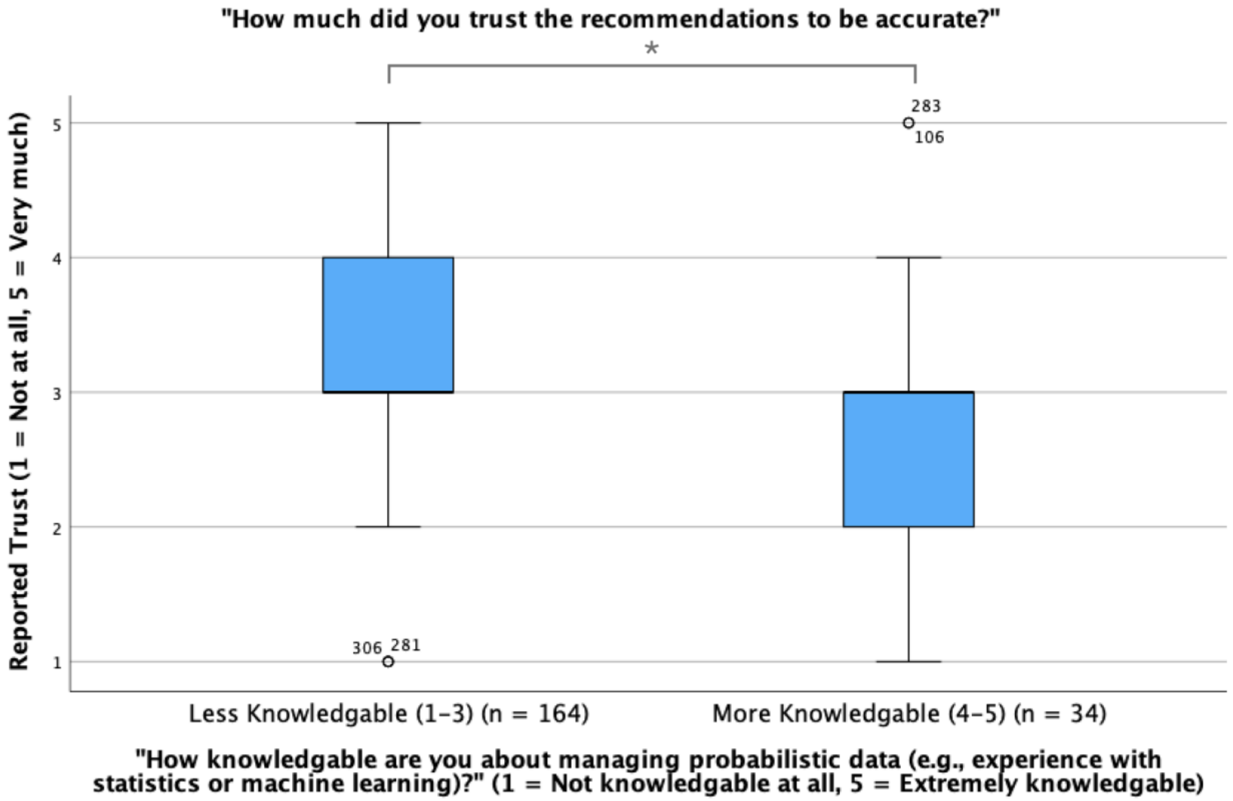


Figure 4-38: Reported trust of the recommendations was statistically significantly higher for participants with less knowledge about managing probabilistic data.

## Reliance on Recommendations

Reliance on recommendations was measured by a single 1-5 Likert-style survey question: “How much did you rely on the ProdStar recommendations to make your decisions?” Only participants in Conditions 2, 3, 5, and 6 were included in this analysis, since Conditions 1 and 3 did not have access to a ProdStar recommendation. Reliance on recommendation was treated as an ordinal variable and analyzed with Mann-Whitney U tests.

### Reliance on Recommendations and Chart Style

A Mann-Whitney U test was run to determine if there were differences in reliance on recommendations between charts without uncertainty displayed and charts with uncertainty displayed. Distributions of the reported reliance on recommendations with charts that did or did not include uncertainty were similar, as assessed by visual inspection. Median reported reliance on the recommendations was not statistically significantly different between charts without uncertainty (*Mdn* = 3.03) and charts with uncertainty (*Mdn* = 3.19), *U* = 6517.5, *z* = 1.033, *p* = .301, *r* = -.070 (trivial effect size) (Figure 4-39).

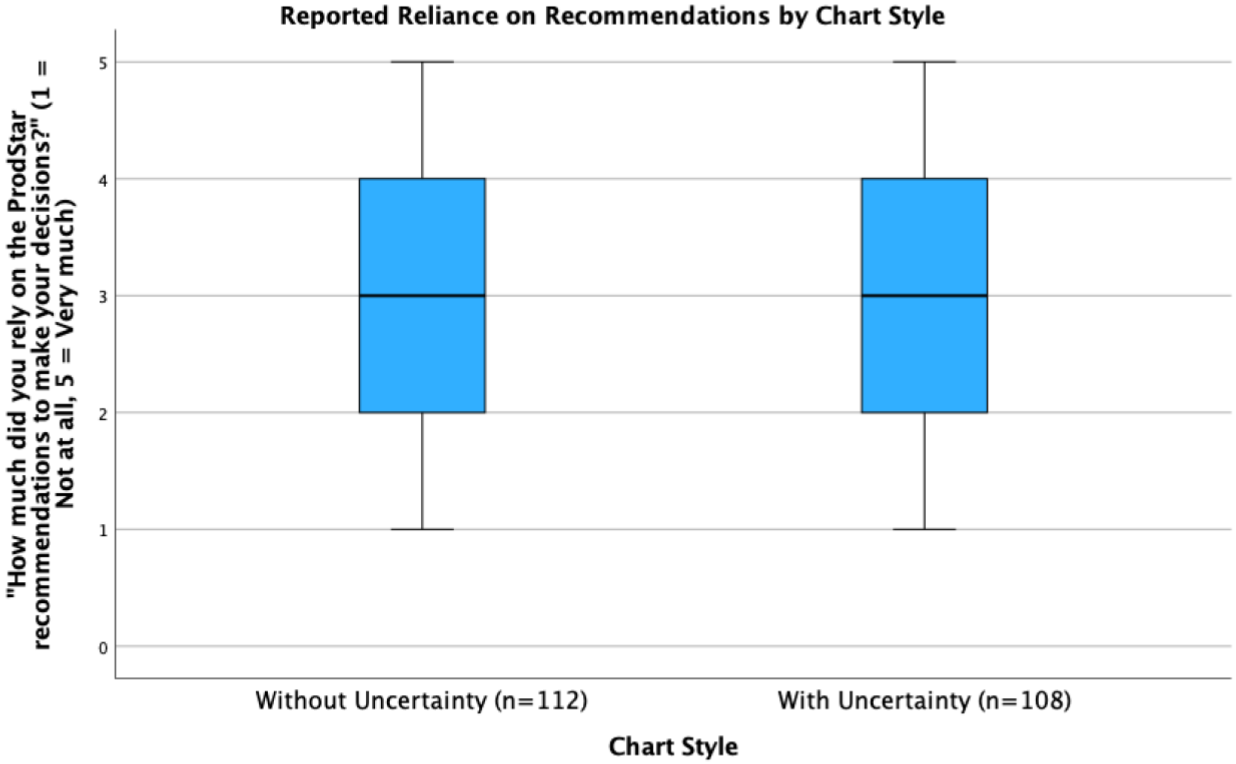


Figure 4-39: There was no statistically significant difference in reliance on recommendations between charts with and without uncertainty.

### Reliance on Recommendations and Recommendation Style

A Mann-Whitney U test was run to determine if there were differences in reliance on recommendations between basic and detailed recommendations. Distributions of the reported reliance on recommendations with basic and detailed recommendations were similar, as assessed by visual inspection. Median reported reliance on the recommendations was lower with basic recommendations (*Mdn* = 2.90) than detailed recommendations (*Mdn* = 3.34), a statistically significantly difference, *U* = 7361, *z* = 2.929, *p* = .003, *r* = -.197 (small effect size) (Figure 4-40).

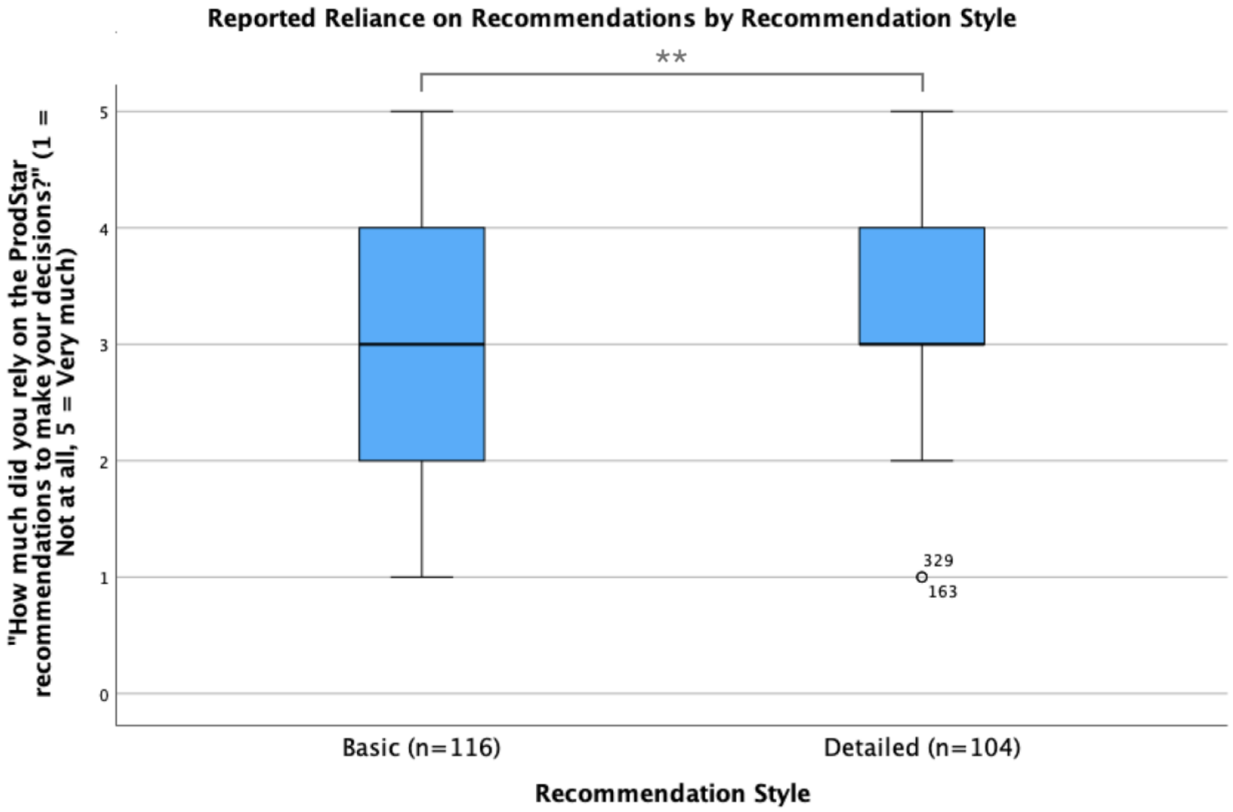


Figure 4-40: Reported reliance on the ProdStar recommendations was statistically significantly higher when detailed recommendations were provided than when basic recommendations were provided.

### Reliance on Recommendations and Gender

A Mann-Whitney U test was run to determine if there were differences in reliance on recommendations between women and men. Distributions of the reported reliance on recommendations for women and men were similar, as assessed by visual inspection. Median reported reliance on the recommendations for women (*Mdn* = 3.15) and men (*Mdn* = 2.91), was not statistically significantly different, (*U* = 3800, *z* = -1.310, *p* = .190, *r* = -.090 (trivial effect size) (Figure 4-41).

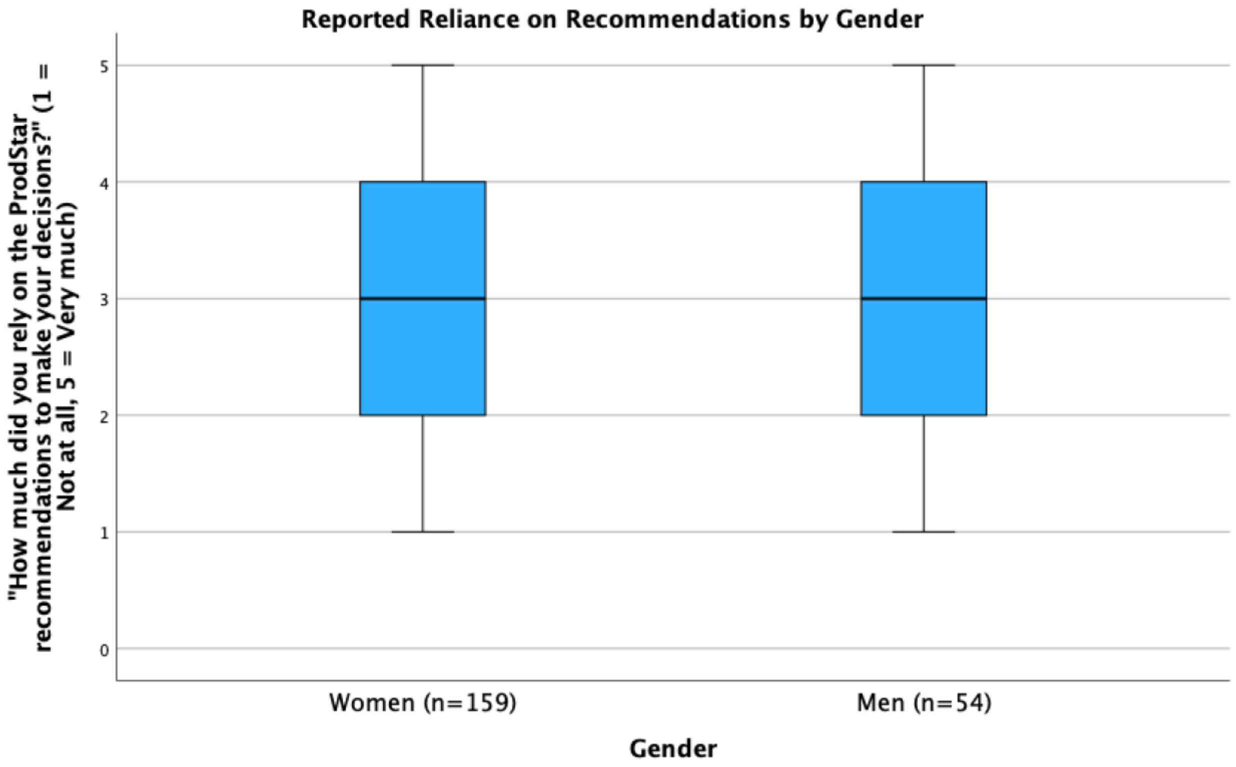


Figure 4-41: There was not a statistically significant difference in reported reliance on the ProdStar recommendations between women and men.

**Reliance on Recommendations and Information Processing Style**

A Mann-Whitney U test was run to determine if there were differences in reliance on recommendations between comprehensive and selective information processing styles. Distributions of the reported reliance on recommendations for both information processing styles were similar, as assessed by visual inspection. Median reported reliance on the recommendations for participants with the comprehensive information processing style (Abi, *Mdn* = 3.16) and selective information processing style (Tim, *Mdn* = 3.02), was not statistically significantly different, *U* = 5166.5, *z* = -.834, *p* = .404, *r* = -.057 (trivial effect size) (Figure 4-42).

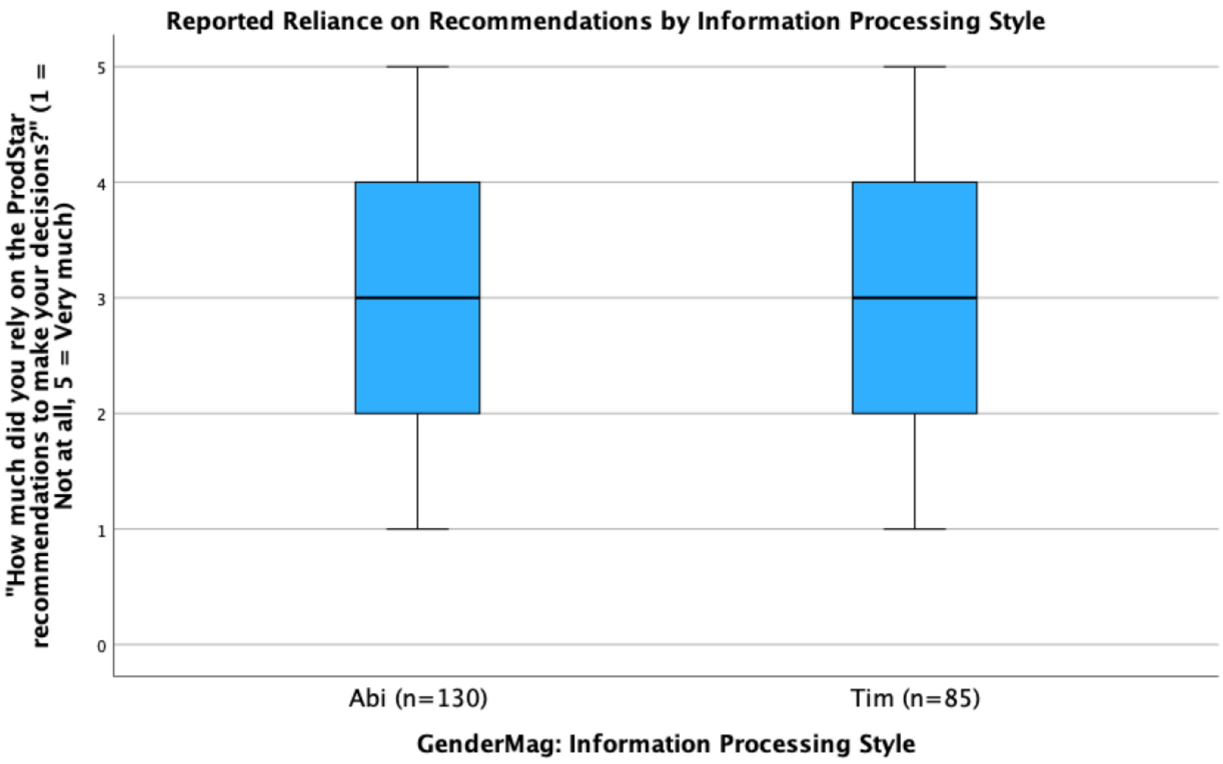


Figure 4-42: There was not a statistically significant difference in reliance on the ProdStar recommendations between participants with the comprehensive information processing style (Abi) and participants with the selective information processing style (Tim).

**Summary of Results**

The overall results are reported in Table 4-1. Overall, chart style had a statistically significant effect on decision accuracy, decision response time, and decision confidence. Chart style also had statistically significant interaction effects with recommendation style on decision accuracy, and with predominant source of uncertainty and gender on decisions to seek more information. Recommendation style had a statistically significant effect on decision accuracy, decision response time, trust of recommendations, and reliance on recommendations. Recommendation style also had statistically significant interaction effect with chart style on decision accuracy. The predominant source of uncertainty had a statistically significant effect, as well as interaction effects with chart style, on decisions to seek more information. The GenderMag facets had a statistically significant effect on decisions to seek more information and decision confidence. Gender had a statistically significant effect, as well as interaction effects with chart style, on decisions to seek more information. Knowledge of probabilistic data had a statistically significant effect on decisions to seek more information, decision confidence, and trust of recommendations.

Table 4-1: In this table, effects of each independent variable on each dependent variable are reported. Statistically significant results are indicate with an "X," statistically significant interaction effects are indicated with an "I," and analyses that were not performed because they were not relevant to the hypotheses are indicated with an "N/A."

| | Task Performance | | | Judgments of Information and Visualizations | | |
|---|---|---|---|---|---|---|
| Variable | Decision Accuracy | Decision Response Time | Decisions to Seek More Information | Decision Confidence | Trust of Recommen-dations | Reliance on Recommen-dations |
| Chart Style | X, I | X | I | X | - | - |
| Recommen-dation Style | X, I | X | - | - | X | X |
| Predominant Source of Uncertainty | - | - | X, I | - | - | - |
| GenderMag facets | - | - | X | X | N/A | N/A |
| Gender | - | - | X, I | X | - | - |
| Knowledge of Probabilistic Data | - | N/A | X | X | X | N/A |

# CHAPTER 5. DISCUSSION

## Discussion of Results

This section includes a discussion of the results, starting with a review of significant findings for each of the independent variables. Next, the main conclusions of this study and opportunities for future work are reviewed, including additional analyses that could be run with the data from this experiment. This section concludes with a brief overview of limitations.

Overall, four of the eight hypotheses were supported, two were partially supported, and two were not supported (Table 5-1).

Table 5-1: Of the eight hypotheses, four were supported, two were partially supported, and two were not supported.

| Hypothesis | Results |
|---|---|
| H1: Participants will make recommendations to seek more information in a larger percentage of trials when provided with bar charts with uncertainty displayed than when provided bar charts without uncertainty displayed. | Partially supported |
| H2: Participants will report lower confidence in their decisions when provided with bar charts with uncertainty displayed than when provided bar charts without uncertainty displayed. | Supported |
| H3: Participants will achieve higher decision accuracy when a ProdStar decision recommendation is provided than when not. | Supported |
| H4: Participants will report higher trust of the recommendations when a detailed recommendation is provided than when a basic recommendation is provided | Supported |
| H5: Participants will make recommendations to seek more information in a larger percentage of trials when the uncertainty is shown as being predominantly due to unavailable data than when the uncertainty is predominantly due to unpredictable data. | Supported |
| H6: Participants with the Abi GenderMag facets will have longer response times and report lower confidence in their decisions. | Partially supported |
| H7: Participants with the comprehensive information processing style will make recommendations to seek more information in a larger percentage of trials than participants with the selective information processing style. | Not supported |
| H8: Participants with the selective information processing style will report higher reliance on decision recommendations than participants with the comprehensive information processing style. | Not supported |

### Chart Style: The Impact of Showing Uncertainty on the Bars

The charts without uncertainty were associated with significantly higher decision accuracy, shorter response times, and higher decision confidence. In some circumstances, charts without uncertainty were associated with more decisions to seek additional information.

It was hypothesized (H1) that participants would more frequently recommend seeking more information when viewing bar charts with uncertainty displayed than when provided bar charts without uncertainty displayed because the extra information would inform them about knowledge gaps (Dong & Hayes, 2012) and encourage them to reflect on the data (Kaur et al., 2020). While there was not a statistically significant main effect of chart style on decisions to seek more information, there were two statistically significant interactions between chart style and other variables.

First, men were more likely to decide to seek more information when they were provided charts with uncertainty (*M* = 60%) than without uncertainty (*M* = 46%), a statistically significant difference with a small effect size that was not seen in women (*M* = 66% for charts without uncertainty, *M* = 64% for charts without uncertainty). The gender differences are discussed in a later section, but this could be a result of either gender-trending differences in decision-making or of the demographics of participants who completed the full survey, who were predominantly women.

Second, the chart style and the predominant source of uncertainty had a significant interaction effect on decisions to seek more information ($p$ = .058, partial $\eta^2$ = .058, a small effect size). The analysis of this interaction effect only included data from Conditions 3 and 6, in which participants were provided a detailed recommendation depicting the predominant source of uncertainty. Participants were more likely to decide to seek more information when the uncertainty was predominantly due to unavailable data (discussed below). This difference was seen with both chart styles, but the difference was particularly strong when charts included uncertainty (a mean difference of 10.1% for charts without uncertainty, and a mean difference of 25.7% for charts with uncertainty). Previous papers have suggested that providing more information about uncertainty can prompt users to reflect more before making a decision (Aerts et al., 2003; Dong & Hayes, 2012; Kaur et al., 2020), and in this study, the uncertainty information in the chart may have prompted participants to not only reflect more on the charts, but also to reflect more on information in the detailed ProdStar recommendation.

Based on these two interaction effects, H1 was partially supported. Men were more likely to make recommendations to seek more information when provided charts with uncertainty, but women were not. Participants were more likely to account for the predominant source of uncertainty in their decisions to seek more information when provided charts with uncertainty.

It was also hypothesized that participants would report lower decision confidence when provided charts with uncertainty than without uncertainty (H2), a trend that was also noted in other studies about uncertainty (Arshad et al., 2015; Correll & Gleicher, 2014; Dong & Hayes, 2012). While the difference in decision confidence for charts with and without uncertainty was statistically significant ($p$ = .007, $d$ = .267, a small effect size), the difference in scores between groups was relatively small. Mean decision confidence was 3.42 / 5 for charts with uncertainty and 3.60 / 5 for charts without uncertainty (based on a 1-5 Likert-style question that was presented after each of the 12 trials). The small difference in decision confidence between chart styles suggests that while the depiction of uncertainty in the charts may have influenced decision confidence, other factors may have played a larger role.

Previous work has shown that multiple factors affect confidence judgments (Harvey, 1997), Individual differences play a role in confidence judgments (Blais et al., 2005; Klayman et al., 1999; Lundeberg et al., 1994; Pallier et al., 2002), possibly to a greater extent than the decision domain or task (Klayman et al., 1999). Heuristics and biases can also affect decision accuracy. People use heuristics and biases in decision-making with uncertain information, such as the anchoring bias (making judgments based on the first input they encounter) and the representativeness heuristic (making judgments based on how well the input aligns with their existing knowledge), which can result in over-confidence (Tversky & Kahneman, 1974). In this study, participants who saw a ProdStar recommendation (Conditions 2, 3, 5, and 6) may have anchored on the recommendation, which was at the top of the page, before studying the charts. Since the recommendations were perfectly reliable, those participants may have also noticed that the recommendations generally aligned with the chart data and their building mental model of ProdStar, boosting their decision confidence. However, while these heuristics and biases, combined with ProdStar's reliability, may have increased decision confidence, this study also included time pressure, which is associated with reduced decision confidence (Zakay & Wooler, 1984). Overall, H2 was supported, and the small margin in decision confidence between chart styles may be due to multiple competing factors that affect confidence judgments.

Charts without uncertainty were also associated with higher decision accuracy (a medium effect size) and faster response times (a small effect size), results that are consistent with some previous work  (Correll & Gleicher, 2014; Lem et al., 2013; Pfaff et al., 2013). In this study, it is possible that the relatively small amount of uncertainty in the 12 scenarios limited the value of displaying uncertainty on the charts. The scenarios were designed to only have up to 20% uncertainty per decision criterion, which was not enough to determine whether Product A or B had the highest score. In other words, removing all sources of uncertainty in the 12 scenarios would never change which product was most optimal. Therefore, an expert user analyzing both chart styles should always come to the same conclusion and have equivalent decision accuracy. For the laypeople in this study, however, the ranges of uncertainty on the chart may have simply provided more ways to interpret the data, ultimately leading to greater variability in responses and cognitive load.

Previous studies found that users vary in their strategies when interpreting uncertainty data. In one study, decision-making strategies varied with the magnitude of uncertainty, and participants either ignored uncertainty information or preferred options with less uncertainty (Padilla et al., 2015). Another study described three common strategies: choosing whichever option a) had the best best-case scenario, b) had the best worst-case scenario, or c) had a significantly higher best-case scenario or significantly lower worst-case scenario (Dilla & Steinbart, 2005). They varied the uncertainty visualization (graph vs. table, minimum and maximum values vs. midpoint value and range) and found that most participants used whichever method required the least cognitive effort. Participants employed more strategies using the range, or error bar width, when presented with data in charts, but they employed more strategies using the minimum and maximum values when presented with data in tables (Dilla & Steinbart, 2005). In this study, participants may have also varied in their strategies for factoring the uncertainty minimum, maximum, midpoint, and range into their decision-making. In another study, decision accuracy decreased and response times increased when making decisions using visualizations under high cognitive load (Allen et al., 2014).

The slower response times when charts include uncertainty, a mean difference of 1.53 seconds, may simply be a result of participants taking more time to reflect about the data. In a previous study, decision response times were longer when an interaction decision analysis tool was added but other factors remained constant (Pfaff et al., 2013). A review of uncertainty visualizations concluded that a high quality, detailed visualization can also help people make decisions more quickly (Eberhard, 2021), but this advantage may depend on the extent to which the uncertainty affects potential outcomes and the decision stakes (e.g., in a high-stakes decision with significant uncertainty, a detailed visualization may help users make robust decisions quickly, but in a low-stakes decision with less uncertainty, a detailed visualization adds unnecessary complexity). The design implication is that more information is sometimes useful, but the amount of information needs to be optimized for the specific context.

## Recommendation Style: The Impact of Providing Different Styles of Recommendation

Group differences between the three different recommendation styles (none, basic, and advanced) were statistically significant for decision accuracy, decision speed, trust of the recommendations, and reliance on the recommendations. Recommendation style was hypothesized to be related to decision accuracy (H3), trust of the recommendations (H4), and decisions to seek more information (H5).

Providing a ProdStar recommendation was predicted to be associated with higher decision accuracy (defined as the percentage of trials in which participants selected the optimal product) than not providing a ProdStar recommendation (H3). The results partially support this hypothesis. There was a statistically significant effect of recommendation style on decision recommendation ($p = .019$, partial $\eta^2 = .023$, a small effect size). However, while the mean decision accuracy for both the basic and detailed recommendations was higher than the mean decision accuracy with no recommendation in this data sample, only the difference between detailed recommendations and no recommendations was statistically significant ($p = .028$). All three recommendation conditions had considerable variation in the data: no ProdStar recommendations (*M* = 73.3%, *SD* = 17.7%), basic ProdStar recommendations (*M* = 78.3%, *SD* = 15.1%), and detailed ProdStar recommendations (*M* = 79.0%, *SD* = 16.5%).

Although only the detailed recommendation was associated with statistically significantly higher decision accuracy than the no recommendation (*p* = .028), decision accuracy in the basic recommendation condition was almost statistically significantly higher than the no recommendation condition (*p* = .054). The information about sources of uncertainty in the detailed recommendation may have motivated participants to take time to reflect on the data before making a decision, ultimately leading to higher decision accuracy.

This explanation aligns with mean decision response times, which were also statistically significantly different between recommendation styles (*p* = .004, $\eta^2$ = .055, a small effect size). Response times were statistically significantly longer for detailed recommendations than no recommendations by 3.2 seconds (*p* < .001), and longer for basic recommendations than for no recommendations by 1.6 seconds (*p* = .050). While this increase is too small to have an appreciable effect on results in many real-world decision-making scenarios, it is possible that the difference would scale up in more complex or higher stakes decisions. However, it is encouraging that participants were not slowed down to the extent that could not perform the task; most participants did not approach the 45-second time limit. Of all participants, 16% reached the 45-second limit in at least one trial, and the maximum number of trials in which any participant reached the time limit was 3. The small increase in decision response time is not out of place in HAT literature, where increased transparency has been associated with positive, negative, or neutral effects on response times (van de Merwe et al., 2024). If the increased decision response time with recommendations scales to real-world tasks, the practical implication is that designers must consider the potential trade-off between decision response time and accuracy.

The increase in decision accuracy with detailed recommendations also aligns with previous work about transparency in HAT and XAI. Two literature reviews of empirical HAT studies report a positive relationship between high agent transparency and high user performance (O'Neill et al., 2022; van de Merwe et al., 2024). These results also support recommendations from the XAI community to include more information about sources of uncertainty as a way to improve human decision-making by developing interpretable AI-generated model outputs and decision support (Barredo Arrieta et al., 2020; Kaur et al., 2020; Tomsett et al., 2020).

One confounding factor that may have affected the decision accuracy results is the extent to which participants followed the task instructions. In the questionnaire, participants were asked to briefly describe their strategy for deciding between Products A and B. While many participants described a process of utilizing the criteria weights and scores according to the task training, some instead described strategies that were inconsistent with the training. Example quotations include: "*I would look at the cost and quality mainly. If the price was similar but the quality was bad for one, I would choose the one with better quality*" and "*Demand and cost because if demand isn't there the product won't be successful*." Participants who used their own criteria weights, rather than those depicted in each MegaMart scenario, would arrive at different decision outcomes. That different approach may account for many of the outliers and low decision accuracy scores. Future work could include coding these responses, then comparing the performance of participants whose strategies were or were not compliant with the task instructions.

However, even with outliers, the higher decision accuracy with detailed recommendations supports the hypothesis. Because of the small scoring margin between Products A and B (5% difference in scores), the ProdStar recommendation was expected to help users make decisions with nearly ambiguous data. As discussed in the above section, some participants may have anchored on the perfectly reliable ProdStar recommendations at the top of the page before inspecting the chart. This may have led participants to develop greater trust of ProdStar and reliance on the recommendations since they did not obviously conflict with the data on the charts.

This explanation is supported by the data; user-reported reliance on the recommendations was statistically significantly higher for detailed recommendations than basic recommendations ($p$ = .003, $r$ = -.197, a small effect size). In a perfect system, high reliance on recommendations and recommendation reliability could combine to create a virtuous cycle. Greater reliance on the recommendation system would improve participants' decision accuracy, and the high decision accuracy could result in higher reliance. This pattern could be particularly applicable in a real-world scenario with a feedback loop about decision outcomes. However, it should also be noted that high reliance on the recommendations has the potential to invoke automation bias, in which users rely upon automation instead of using their own critical thinking skills and fail to notice when the system does make an error (Parasuraman & Manzey, 2010).

It was also predicted that participants would report higher trust of the detailed recommendations than the basic recommendations (H4). This was measured by a single 1-5 Likert scale question: "How much did you trust the recommendations to be accurate" (1 = Not at all, 5 = Very much). This hypothesis was also based on previous work in the HAT and XAI, in which high agent transparency and interpretable are associated with high trust of the agent or model output (Barredo Arrieta et al., 2020; J. Y. C. Chen et al., 2014; Kaur et al., 2020; O'Neill et al., 2022; Tomsett et al., 2020; van de Merwe et al., 2024).

This hypothesis was supported; reported trust of the recommendations was statistically significantly higher with detailed recommendations than basic recommendations (*p* = .003, *r* = .196, a small effect size). This supports previous findings that trust of agents is a function of both reliability and transparency (J. Y. C. Chen et al., 2014). However, while many of the HAT studies reviewed in this research used imperfect agent-provided recommendations (e.g., the agent's recommendation was only correct 70% of the time), this study used recommendations that were 100% reliable. This is interesting because it shows that even with a perfectly reliable system, higher transparency still results in higher trust of the system. This result suggests that transparency may be a part of users' mental model of recommendation agents, regardless of the agent's reliability. Perhaps users still expect a recommendation agent to make an error at some point, and the high transparency gives them confidence that they will be able to catch this error. It is worth noting that in this study, participants did not receive feedback about their performance, so their perception of ProdStar's accuracy was a result of how well its recommendations aligned with their interpretation of the MCDA charts.

The higher trust of the detailed recommendation in this study also contrasts with one previous study, in which the high transparency condition, which added uncertainty information, was associated with decreased trust of recommendations (Stowers et al., 2020). Uncertainty in the Sowers et al. study was depicted by highlighting uncertain decision factors in a scatter plot, changing the opacity of elements with uncertainty on a map, and describing the sources of uncertainty in a textual description. The authors suggested that the decreased trust with high transparency in their study may have been due to the large amount of information in that version of the data dashboard.

However, comparing the Stowers et al. study to this study reveals two other possible explanations. First, in the Stowers et al. study, the depictions of uncertainty described the presence of uncertainty, but not the amount of uncertainty, while this study made the amount of uncertainty more explicit (to some extent with the detailed recommendations, and to a greater extent when paired with charts with uncertainty). Describing the presence of uncertainty as a binary yes/no value may make it difficult for users to understand how the uncertainty factored into decision recommendations or MCDA scoring, ultimately reducing trust in the recommendations (e.g., participants using a route planning recommendation might see text stating that vehicle speed may be affected by adverse weather conditions, but no information about the likelihood or extent of delay and no notice of whether this information was factored into the recommendation).

Second, the agent recommendations were 63% reliable in the Stowers et al. study, but 100% reliable in this study. Properly calibrated trust of agent recommendations is a result of both reliability and transparency; low reliability with high transparency results in appropriately low trust (J. Y. C. Chen et al., 2014). In the Stowers et al. study, the higher transparency condition with depictions of uncertainty may have helped participants recognize the agent’s shortcomings and correctly assume it to be less trustworthy. In this study, the higher transparency in the detailed recommendations helped participants recognize that the recommendations were both reliable and trustworthy.

### Predominant Source of Uncertainty: The Impact of Describing Sources of Uncertainty

It was hypothesized (H5) that participants would more frequently make recommendations to seek more information when the source of uncertainty was predominantly due to unavailable data (epistemic uncertainty) than when it was predominantly due to unpredictable data (aleatoric uncertainty). This hypothesis stemmed from previous research in the XAI, ML and, HCI communities, which suggests that helping users understand the amount and type of uncertainty in AI-generated model outputs will result in greater trust of the model and better human decision-making (Barredo Arrieta et al., 2020; Kaur et al., 2020; Tomsett et al., 2020). That research suggests that users who have an accurate understanding of the sources of uncertainty can calibrate themselves on when to seek more information. If there is epistemic uncertainty, they can gather more data to make a more sound decision, and if there is aleatoric uncertainty, they can recognize that gathering more data is futile and instead make a decision with the available information.

This hypothesis was strongly supported by the results of this study. Overall, decisions to seek more information were statistically significantly higher in trials when the uncertainty was predominantly due to unavailable data than predominantly due to unpredictable data ($p < .001$, partial $\eta^2 = .247$, a large effect size). This effect was even more pronounced when the interaction of chart style was considered (Table 5-2). For charts without uncertainty, there was a statistically significant difference in decisions to seek more information depending on the source of uncertainty (a mean difference of 10.12%, $p = .021$, $d = .866$, a large effect size). For charts with uncertainty, there was a larger difference based on the predominant source of uncertainty (a mean difference of 25.65%, $p < .001$, $d = .849$, a large effect size).

Table 5-2: Participants made significantly more recommendations to seek more information when the uncertainty was predominantly due to unavailable data than when it was predominantly due to unpredictable data. This difference was even more pronounced when the detailed recommendations were paired with charts with uncertainty.

| **Chart Style** | **Predominant Source of Uncertainty** | **Mean Percentage of Trials with Recommendations to Seek More Information *(SD)*** | |
|---|---|---|---|
| Without Uncertainty | Unavailable Data | 65.99% *(26.83%)* | * |
| | Unpredictable Data | 55.86% *(30.63%)* | |
| With Uncertainty | Unavailable Data | 73.18% *(27.85%)* | *** |
| | Unpredictable Data | 47.53% *(27.85%)* | |

The participants in this study did not receive any special training about sources of uncertainty, beyond a single sentence in the training video about the detailed ProdStar recommendations: “It also tells them if uncertainty in the scores is due to unavailable data, like when their marketing team hasn’t gathered enough information, or due to unpredictable data, like volatility in market demand for their products.” The participants in this study were not experts in interpreting probabilistic data; only 17% of participants described their knowledge about managing probabilistic data as being higher than 3 (“Somewhat knowledgeable”) on a 1-5 Likert scale. These results are encouraging for developing interpretable AI-generated model outputs, because they suggest that even with minimal task training, layperson users can correctly interpret and integrate information about epistemic and aleatoric uncertainty in their decision-making processes. These results also served as an experiment manipulation check, indicating that participants understood generally understood the information provided by the detailed recommendations.

**GenderMag Facets: The Impact of Gender-Trending Characteristics**

The remaining hypotheses were based on the GenderMag facets. First, it was hypothesized that participants with the Abi facets would have longer decision response times and lower decision confidence than participants with the Tim facets (H6). These predictions were based on the Abi facets' potential impact on decision-making. The comprehensive information processing style is characterized by taking time to review all the relevant data before making a decision, so participants with this facet were expected to have longer response times. Abi's low self-efficacy and reduced risk tolerance were expected to be associated with reduced decision confidence. However, this hypothesis was only partially supported. No statistically significant differences in decision response time were found, but decision confidence was statistically significantly lower for participants with three of the Abi facets.

Overall, decision confidence was statistically significantly lower for participants with four of the facets: Abi's low self-efficacy (*p* = .025, *d* = .235, a small effect size), Tim's selective information processing style (*p* = .044, *d* = .196, a negligible effect size), Abi's motivations to use technology to accomplish a goal *p* = .009, *d* = -.310, a small effect size), and Abi's preference to learn by process (*p* = .017, *d* = -.275, a small effect size). For these four facets, the difference in decision confidence was small, aligning with the above discussion about many competing factors affecting confidence judgments (Blais et al., 2005; Klayman et al., 1999). Decision confidence was not statistically significantly different for participants with different attitudes toward risk.

These results partially support the hypothesis that Abi's facets would be associated with lower decision confidence. It seems reasonable to find that low self-efficacy is related to low decision confidence, given the close ties between self-efficacy and self-confidence (Cramer et al., 2009). However, Abi's motivation to use technology as a means to accomplish a goal (rather than Tim's motivation to use technology for technology's sake) and Abi's preference to learn by process (rather than Tim's preference for tinkering) were expected to be less relevant to decision-making. It is possible that participants with preferences for tinkering and exploring technology (Tim facets) do so out of confidence (i.e., they feel safe to explore and make mistakes), and this confidence also affected their decision-making in this experiment.

While these results partially support the hypothesis that Abi's facets would be associated with lower decision confidence, it is worthwhile to also consider these results in the context of the gender analysis. Women had statistically significantly lower decision confidence than men (a medium effect size), but only three of the Abi facets (typically more common in women) were associated with lower decision confidence than the Tim facets. In the distribution of GenderMag facets in women and men (Figure 3-7), more men identified with Abi than Tim in the attitude toward risk and the information processing style facets. These were also the only facets in which the Abi style was associated with higher decision confidence, suggesting that there may be an interaction between gender and GenderMag facets on decision confidence. However, regardless of potential interactions with gender, the design implication is that some users may have different needs (e.g., more information, different information, or alternate presentations of the information) to feel confident in their decision-making.

No statistically significant differences in decision response time were found among the GenderMag facets. This again aligns with the gender analysis, in which no statistically significant difference was found between women and men. While it is possible that none of the facets are relevant to decision-making with MCDA dashboards and information uncertainty, another possibility is that the population of participants did not represent a typical GenderMag facet distribution.

The population in this experiment already had an imbalanced gender representation (243 women, 83 men, 13 other genders, and 3 with no gender specified), indicating that men were less likely to open and/or complete this survey. Potential participants with more of the Tim facets may have also decided against participating in the full survey. If this were the case, it would also affect the facet scoring since the threshold between labeling a participant as an Abi or a Tim is based on the population median, not a fixed value (Vorvoreanu et al., 2019).

This explanation is supported by the distribution of the GenderMag facets in the participants of this study. Participants were grouped by how many Abi vs. Tim facets they have, and this data was plotted by gender (Figure 5-1). While it is typical for participants to have a mix of Abi and Tim facets (e.g., few people are a “Pure Abi,” identifying with all five of the Abi facets), in previous research it was typical for most women to have 3-5 Abi facets and for most men to have 3-5 Tim facets (Burnett, 2020; Vorvoreanu et al., 2019). The trendlines in Figure 5-1 for this study show a relatively normal distribution for women and men, whereas the trendlines in previous studies were more heavily skewed to the left for women (toward more Abi facets) and to the right for men (toward more Tim facets) (Figure 5-2) (Burnett, 2020; Vorvoreanu et al., 2019). This shows that something about the distribution of Abi and Tim facets among women and men in this study differed from previous studies, which may partially explain why many of the GenderMag analyses did not yield statistically significant results.

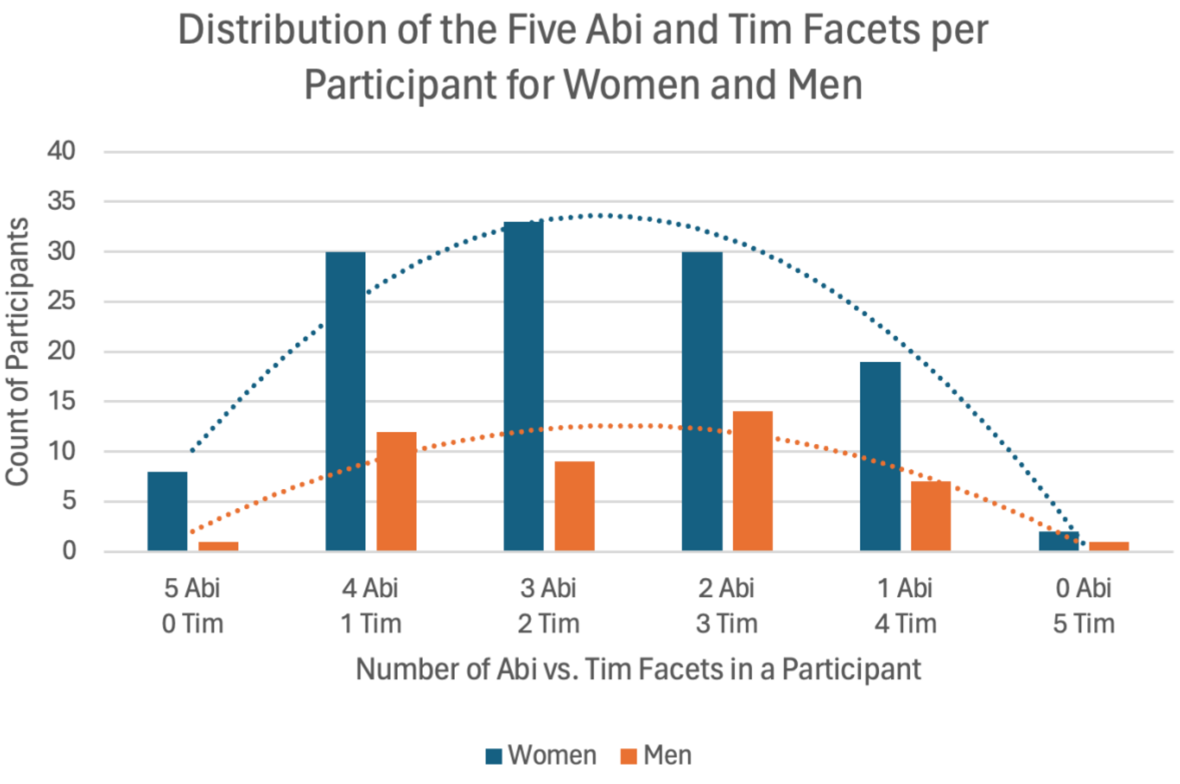


Figure 5-1: There was a nearly normal distribution of the number of the five Abi and Tim facets in participants for women and a men, with a slight skew toward predominantly Abi facets for women and toward predominantly Tim facets for men.

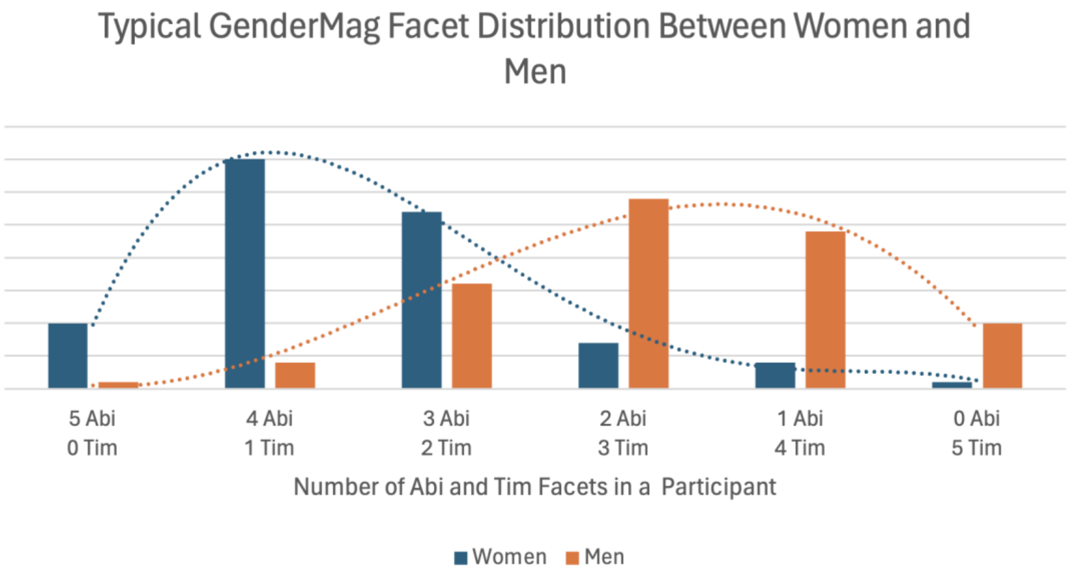


Figure 5-2: Approximation of a more typical distribution of Abi vs. Tim facets in previous GenderMag studies. Adapted from (Vorvoreanu et al., 2019).

To understand the differences between the distribution of Abi vs. Tim facets in this study and their distribution in other studies, it is also useful to consider the median scores in this population for each facet, which were used to determine the Abi vs. Tim threshold. The median scores were near the midpoint of the 1-9 scale for three of the five facets (Figure 5-3). The median scores were skewed toward Abi's comprehensive information processing style and Tim's high self-efficacy. The practical implication of this is that some participants may be labeled differently in another population sample, limiting the power of the statistical analysis in this study. Overall, this suggests that the GenderMag facet survey may be a more useful instrument for describing trends in a population than for analyzing individual differences.

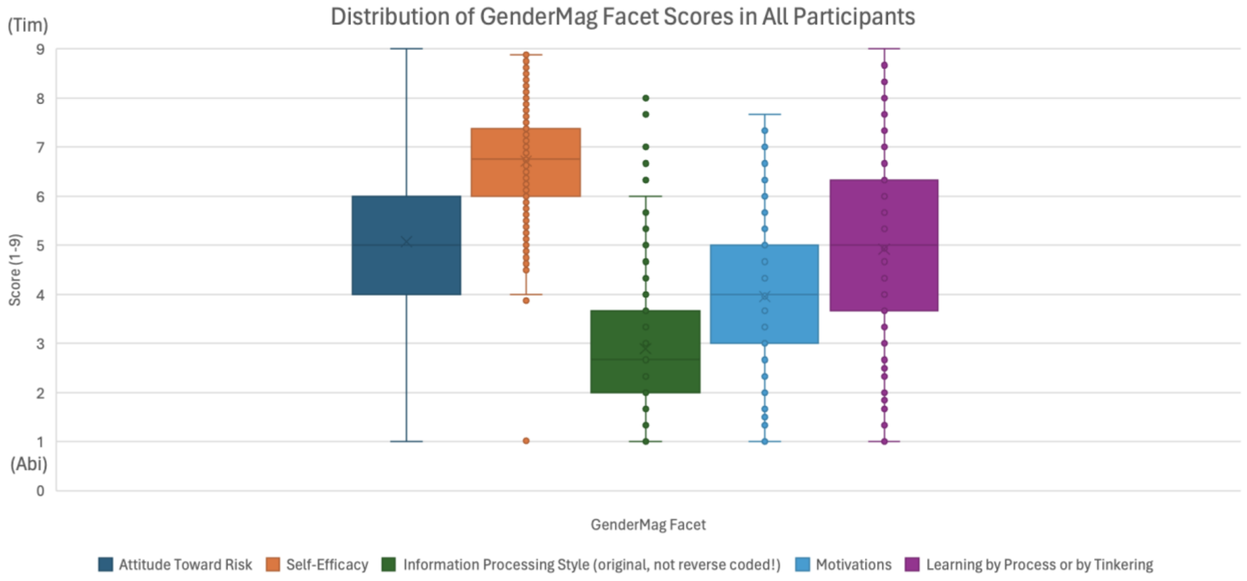


Figure 5-3: Median scores were close to the midpoint (4.5) of the range for three facets. Median score was closer to the Abi endpoint for Information Processing Style. The Information Processing Style data were reverse coded for this chart so that low scores always correspond to the Abi facet.

There were also two hypotheses related to the Information Processing Style facet. It was predicted that participants with the comprehensive information processing style would make more decisions to seek additional information than participants with the selective information processing style, due to their preferences for gathering all relevant data before taking an action (H7). It was also predicted that participants with the selective information processing style would report higher reliance on the ProdStar recommendations than participants with the comprehensive information processing style due to their preference for using just the most relevant or important information before taking an action (H8).

Neither of these hypotheses was supported; there were no statistically significant differences in decisions to seek more information or reliance on recommendations between information processing style. This could be due to the distribution of GenderMag facets in this population, as discussed above; this population was skewed toward Abi's comprehensive information processing style. Since the Abi vs. Tim threshold was based on the median score in this population, some of the "Tims" in this study might be "Abis" in another study. This may have masked any true differences between information processing styles in the dependent variables. However, another possibility is that the relative simplicity of the data dashboard in this study was not enough to inspire differences based on this facet. In other words, perhaps decision-making only differs between information processing styles when users are provided a more complex dashboard with more data to consider or under different circumstances (e.g., higher time pressure or workload).

Of the GenderMag-related analyses conducted in this study, one other statistically significant result was found. Participants with a preference for learning by process were more likely to make decisions to seek more information than participants with a preference for learning by tinkering (a mean difference of 8.72%, $p$ = .005, $d$ = .332, a small effect size). This result was unexpected, so it warrants further analysis. This facet was scored based on 1-9 Likert scale agree/disagree responses to three statements:

1. “I enjoy finding the lesser-known features and capabilities of new software and technology.”
2. “I explore areas of new software and technology before it is time for me to use it.”
3. “I’m never satisfied with the default settings for new software and technology; I customize them in some way.”

Participants who more strongly disagreed with these statements were more like to make decisions to seek more information.

Participants with a preference for learning by process typically approach problems by following a specified procedure, as opposed to an open-ended, tinkering approach (Burnett et al., 2016). Previous research about STEM education contrasted a strict, scientific process to problem-solving from a tinkering-based approach in more intuitive learning environments (Wagh et al., 2017). Other researchers made a connection between open-ended problem-solving approaches and the concept of “engineering intuition” (Miskioglu & Martin, 2019). This suggests that there may be an association between the ideas of tinkering and intuition-based problem-solving. In this study, the participants who prefer to learn by tinkering may have considered it sufficient to make decisions based on intuition or gut feel, leading to lower rates of decision to seek more information. Participants who prefer to learn by process, on the other hand, may have seen the recommendation to seek more information as an appropriate, procedural next step to managing uncertainty in the data.

## Gender: The Impact of Gender on MCDA Dashboard Decision-Making

While there were no hypotheses about the effects of gender on any of the dependent variables, it is worthwhile to discuss gender differences in the result in the broader context of inclusive HCI design, particularly given the lack of significant results in the analysis of

GenderMag facets. Group difference between women and men were compared for all six dependent variables. Statistically significant differences were found between these two genders for decisions to seek more information and decision confidence.

Women made more recommendations to seek additional information than men, a mean difference of 11.87% (*p* < .001, *d* = .448, a small effect size). There also was a statistically significant interaction effect between gender and chart style on decisions to seek more information, which was briefly discussed above. There were similar rates of deciding to seek more information among women for both chart styles and men for charts with uncertainty. However, when charts did not include uncertainty, men less frequently decided to seek more information.

The previous studies reviewed in this research did not explicitly measure decisions to seek more information in the context of gender, but an inclination for deciding to analyze all available information before making decision was seen to be more common in women (Meyers-Levy, 1988; Meyers-Levy & Loken, 2015) and is the basis of the information processing style facet (Burnett et al., 2016). In some of, there was an interaction between gender and information congruency on decision-making (Meyers-Levy & Maheswaran, 1991; Meyers-Levy & Zhu, 2010). Women and men made similarly robust decisions when input data was moderately or highly incongruent. When the input data was minimal incongruent, however, women's decision-making was unaffected, while men made more errors due to overlooking the subtle differences in input data (Meyers-Levy & Maheswaran, 1991). These researchers attribute their results to differences in the selective and comprehensive information processing style. A similar effect may have occurred in this study, with ambiguity due to uncertainty in the charts playing a similar role as moderately or highly incongruent input data. In other words, participants with the selective information processing style (which typically are predominantly men) may decide the data on the dashboard is sufficient unless the uncertainty on the carts creates ambiguity or incongruency, in which case they were more likely to decide to seek more information.

However, the explanations for of these significant gender-related results are based on the same research as the GenderMag facets, which did not have similarly significant results. As discussed above, this could be due to an uneven distribution of gender-trending characteristics in the study participants (which were predominantly women). It could be due to the GenderMag facet survey methods relying too heavily on the population score distribution to be a valid method of the facets. It could also suggest that further research is needed to determine whether previous gender studies research still holds true in younger generations. The GenderMag process was published in 2016, so all the references used to develop the facets predate 2016 (Burnett et al., 2016).

There was one more statistically significant result related to gender. Women reported lower confidence in their decisions than men, a mean difference of -0.413 on a 1-5 Likert scale ($p$ < .001, $d$ = -.503, medium effect size). This aligns with previous studies in which women displayed less confidence than men (Bleidorn et al., 2016; Chiesi & Primi, 2015; Estes & Hosseini, 1988; Lundeberg et al., 1994) and expands on it by demonstrating a gender confidence gap in MCDA-based decision making with uncertainty. More generally, this builds on previous gender-related research, in which women’s strategies and tolerance of risk in decision-making differed from men’s (Apesteguia et al., 2012; Atkinson et al., 2003; Fehr-Duda et al., 2006; Jianakoplos & Bernasek, 1998).

### Knowledge of Probabilistic Data Influences Perceptions of ProdStar

Although there were no hypotheses about knowledge of probabilistic data as an individual difference, previous research showed that it may play a role in decision making with information uncertainty (Arshad et al., 2015). Trust of the recommendations seemed to be moderated by knowledge of probabilistic data. Participants with more knowledge of probabilistic data reported statistically significantly lower trust of the ProdStar recommendations than participants with less knowledge of probabilistic data ($p$ = .032, $r$ = .153, a small effect size). However, the participants with more knowledge of probabilistic data also reported higher decision confidence ($p$ = .020, $d$ = -.298, a small effect size). The findings about decision confidence align with previous research, in which ML researchers were more confident making decisions using charts with uncertainty than non-ML researchers, possibly because of their greater knowledge about probabilistic data (Arshad et al., 2015).

The reduced trust in ProdStar recommendations for participants with more knowledge of probabilistic data could be because these participants disagreed with the recommendations. However, decision accuracy was not statistically significantly different between participants with more or less knowledge of probabilistic data ($p$ = .053, $d$ = -.209, a small effect size). Their reduced trust could also be related to transparency; participants with greater knowledge of probabilistic data may want more (or different) information in an MCDA decision dashboard with uncertainty. These participants may have formed a more complex mental model of ProdStar that included factors like the scoring algorithm, which was not fully explained in this experiment. The design implication is that there may be different user interface needs to promote trust in recommendations for users with more knowledge of probabilistic data.

## Conclusions and Future Work

Finally, several conclusions are discussed in the context of the original research questions.

- RQ1: When there is uncertainty in the input data, how does the method of displaying decision comparisons and recommendations on a data dashboard affect decision-making?
- RQ2: How does decision-making vary with gender-trending characteristics when using a data dashboard that includes decisions comparisons and recommendations with uncertainty in the input data?

First, while including uncertainty on charts was associated with reduced decision accuracy and confidence in this study, it had the desirable effect of prompting users to think more deeply about the input data before making a decision. The reduced decision accuracy may have been influenced by noise in the data from a subset of participants who did not perform the task according to the instructions (i.e., by ignoring MegaMart's decision criteria weights). Additional analyses could attempt to sort these outliers and remove them or treat them as a covariate, either through qualitative analysis of the open-field text responses or through the survey question about reliance on the criteria weights, to re-evaluate decision accuracy. The increased response times, while significant, were small. This result could be viewed in a positive light; adding another layer of the information to the charts increased mean response time by only 1.53 seconds (out of the 45-second time limit), and a small increase in response time may be an acceptable trade-off if it means users are reflecting on the data to make more sound decisions. Future work could explore a scaled-up version of this interface in a more complex task to see if the difference in response times also scales up. Finally, the reduced decision confidence when provided charts with uncertainty is not desirable, but future work could explore moderating factors such as more layers of information (e.g., a textual description to provide greater context about the uncertainty) or providing users with an explanation of the MCDA recommendation algorithm. Furthermore, including uncertainty on charts may instead have a positive effect on decision confidence in highly ambiguous, high-stakes decisions.

Second, providing detailed information in the recommendations was associated with desirable results, including higher decision accuracy (particularly when paired with charts that include uncertainty), higher trust of the recommendation, and higher reliance on the recommendation. Providing detail about whether the uncertainty was predominantly due to epistemic or aleatoric uncertainty was particularly useful, in that layperson participants accurately integrated this into their decisions about whether to seek more information. As when charts included uncertainty, detailed recommendations were associated with longer response times, but the 3.2 second increase (relative to providing no recommendation) may be an acceptable trade-off if people are using that time to make more robust decisions. The higher trust in detailed recommendations was a particularly interesting finding because indicates that transparency is still an important aspect of forming trust, even with a perfectly reliable system.

Third, this research showed that there are some gender-trending differences in decision-making in an MCDA context with uncertainty in the input data. Women made more decisions to seek additional information and reported lower decision confidence. Men were less likely than women to decide to seek more information when charts did not include uncertainty, but equally likely when charts did include uncertainty. The causality of these results is not yet fully understood, given that the most obvious research basis for gender-trending differences was already used to develop the GenderMag facets, which yielded few significant results in this study. However, that may have been due to limitations of the GenderMag facet survey as a way to measure individual differences.

Fourth, this research indicated that individuals with greater knowledge about managing probabilistic data may have different user needs in MCDA decision dashboards with uncertainty. These participants reported higher decision confidence but lower trust in the recommendations. Future work could analyze the open-ended response from these participants to attempt explain their reduced trust in the recommendations.

## Future Work

Overall, this research expanded on existing uncertainty visualization, decision recommendation styles, and individual differences on MCDA dashboards. Participants achieved higher decision accuracy and reported both higher trust of and higher reliance on the detailed ProdStar recommendations, and they appropriately calibrated their decisions to seek more information based on predominant sources of uncertainty (epistemic vs. aleatoric). Future work could explore similarly detailed recommendations in different decision-making domains with uncertainty, such as geospatial information or medical imaging and diagnoses. This could determine whether depicting the sources of uncertainty is similarly interpretable and beneficial in other applications and check for interaction effects with other types of charts. Future research should include users with varying degrees of knowledge about managing probabilistic data, since they may have differing trust and usage of decision dashboards with probabilistic data.

This research also expanded on research about transparency, reliability, and trust of recommendation tools, finding that higher transparency is associated with higher trust, even with a perfectly reliable recommendation. Future work could explore varying degrees of transparency with other types of highly reliable agents, such as agents that could be assigned to carry out missions or provide other types of decision support.

Finally, this research built on existing gender research by studying gender-trending differences with decision support dashboards. Women more frequently decided to seek more information than men, a finding that has HCI implications. This decision-making style may be preferable when decision accuracy is the highest priority but less preferable when decision speed is the highest priority. Future work could explore interfaces that enable or encourage either decision-making style based on the user or task. Future research could also evaluate decisions to seek more information in other decision domains and with other decision support tools to see if this result is reproduceable outside of this MCDA scenario. Finally, future work could also analyze the use of validated scales, such as measures of self-efficacy and attitudes toward risk, to study gender-trending characteristics instead of the GenderMag facet survey,

### Additional Analyses

Finally, additional analyses could be performed on the data collected in this experiment. Several dependent variables were not yet analyzed. The SUS (Brooke, 1996) and Net Promoter Scores (Reichheld, 2003) could be analyzed to learn more about the overall usability of the dashboard. The reported reliance on the charts, decision criteria, and sources of uncertainty, as well as perceived accuracy of the recommendations, could be analyzed to further explore participants' perception and usage of the data dashboard. Differences between responses of each of the 12 trials could be analyzed, including the potential of order effects. Although all 12 scenarios were designed to equally difficult, it is also possible that certain questions were disproportionately challenging for charts with or without uncertainty.

More analyses about individual differences could be performed to gain a broader view of individual differences. The GenderMag facets and participants' reported knowledge of managing probabilistic data could also be further analyzed. These individual differences were included only when relevant to a hypothesis, but they may have affected other dependent variables or had significant interaction effects with the independent variables. This survey also collected data about age, college, and student classification, all of which may play a role in decision-making.

The survey also included open-text field questions, asking participants to describe their strategy for deciding between Products A and B and about the pros and cons of this data dashboard. A qualitative analysis of these results could provide a richer understanding of participants' decision-making process and mental models of the recommendations and uncertainty provided in this dashboard. This coded data could also be used to identify outliers who did not follow the task instructions, potentially leading to a more robust analysis of decision accuracy.

## Limitations

The limitations in this study were primarily due to its deployment method and fidelity. First, this experiment was conducted entirely online through a Qualtrics survey and campus-wide email distribution. This means that it was not conducted in a controlled lab setting, so there was variability in the devices used to complete the survey, the extent to which participants paid attention to the task instructions, and whether or not participants completed the survey in one sitting (as opposed to taking a break and returning to the survey later). There also may be variability in the extent to which participants took the survey seriously, since there was no situational pressure from an experiment to focus on the task. Additionally, the participants who completed the survey were predominantly women, meaning that there were unequal group sizes in the gender analyses.

Second, this experiment explored decision-making in a low-stakes, low-fidelity environment. While adding the 45-second time limit with a visible countdown put participants under some time pressure, this was not equivalent to a user making decisions in their professional or personal life with real-world ramifications. Studying this data dashboard in a more realistic environment might show that more or different information is needed on the dashboard and results in different decision-making performance. Additionally, achieving perfectly accurate recommendations in a decision support tool might not be feasible, so decision-making may be affected when the recommendations are incorrect. Similarly, participants only used the dashboard for 12 trials, which may not have been enough time for users to fully calibrate their understanding and trust of the system.

## REFERENCES

Aerts, J. C. J. H., Clarke, K. C., & Keuper, A. D. (2003). Testing Popular Visualization Techniques for Representing Model Uncertainty. *Cartography and Geographic Information Science*, *30*(3), 249–261. https://doi.org/10.1559/152304003100011180

Allen, P. M., Edwards, J. A., Snyder, F. J., Makinson, K. A., & Hamby, D. M. (2014). The Effect of Cognitive Load on Decision Making with Graphically Displayed Uncertainty Information. *Risk Analysis*, *34*(8), 1495–1505. https://doi.org/10.1111/RISA.12161

Apesteguia, J., Amat, G., & Iriberri, N. (2012). The Impact of Gender Composition on Team Performance and Decision Making: Evidence from the Field on JSTOR. *Management Science*, *58*(1), 78–93. https://doi.org/10.1287/mnsc.1110.1348

Arshad, S. Z., Zhou, J., Bridon, C., Chen, F., & Wang, Y. (2015). Investigating user confidence for uncertainty presentation in predictive decision making. *OzCHI 2015: Being Human - Conference Proceedings*, 352–360. https://doi.org/10.1145/2838739.2838753

Atkinson, S. M., Baird, S. B., & Frye, M. B. (2003). Do Female Mutual Fund Managers Manage Differently? *Journal of Financial Research*, *26*(1), 1–18. https://doi.org/10.1111/1475-6803.00041

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191–215. https://doi.org/10.1037/0033-295X.84.2.191

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/J.INFFUS.2019.12.012

Bartik, J., Ruff, H., Calhoun, G., Behymer, K., Goodman, T., & Frost, E. (2019). Visualizations for Communicating Intelligent Agent Generated Courses of Action. In J. Y. C. Chen & G. Fragomeni (Eds.), *Virtual, Augmented and Mixed Reality. Applications and Case Studies: 11th International Conference, VAMR: Vol. 11575 LNCS* (pp. 19–33). Springer. https://doi.org/10.1007/978-3-030-21565-1_2

Beckwith, L., Burnett, M., Wiedenbeck, S., Cook, C., Sorte, S., & Hastings, M. (2005). Effectiveness of end-user debugging software features. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 869–878. https://doi.org/10.1145/1054972.1055094

Behymer, K. J., Mersch, E. M., Ruff, H. A., Calhoun, G. L., & Spriggs, S. E. (2015). Unmanned Vehicle Plan Comparison Visualizations for Effective Human-autonomy Teaming. *Procedia Manufacturing*, *3*, 1022–1029. https://doi.org/10.1016/J.PROMFG.2015.07.162

Blais, A. R., Thompson, M. M., & Baranski, J. V. (2005). Individual differences in decision processing and confidence judgments in comparative judgment tasks: The role of cognitive styles. *Personality and Individual Differences*, *38*(7), 1701–1713. https://doi.org/10.1016/J.PAID.2004.11.004

Bleidorn, W., Arslan, R. C., Denissen, J. J. A., Rentfrow, P. J., Gebauer, J. E., Potter, J., & Gosling, S. D. (2016). Age and gender differences in self-esteem—A cross-cultural window. *Journal of Personality and Social Psychology*, *111*(3), 396–410. https://doi.org/10.1037/PSPP0000078

Brodlie, K., Allendes Osorio, R., & Lopes, A. (2012). A Review of Uncertainty in Data Visualization. *Expanding the Frontiers of Visual Analytics and Visualization*, 81–109. https://doi.org/10.1007/978-1-4471-2804-5_6

Brooke, J. (1996). SUS -- A quick and dirty usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Eds.), *Usability Evaluation in Industry* (pp. 189–194). Taylor & Francis.

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To Trust or to Think. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 21. https://doi.org/10.1145/3449287

Burnett, M. (2020). Doing Inclusive Design: From GenderMag in the Trenches to Inclusive Mag in the Research Lab. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3399715.3400871

Burnett, M., Fleming, S. D., Iqbal, S., Venolia, G., Rajaram, V., Farooq, U., Grigoreanu, V., & Czerwinski, M. (2010). Gender differences and programming environments: Across programming populations. *ESEM 2010 - Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*. https://doi.org/10.1145/1852786.1852824

Burnett, M., Stumpf, S., Macbeth, J., Makri, S., Beckwith, L., Kwan, I., Peters, A., & Jernigan, W. (2016). GenderMag: A Method for Evaluating Software's Gender Inclusiveness. *Interacting with Computers*, *28*(6), 760–787. https://doi.org/10.1093/IWC/IWV046

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*(1), 116–131. https://doi.org/10.1037/0022-3514.42.1.116

Cai, C. J., Reif, E., Hegde, N., Hipp, J., Kim, B., Smilkov, D., Wattenberg, M., Viegas, F., Corrado, G. S., Stumpe, M. C., & Terry, M. (2019). Human-centered tools for coping with imperfect algorithms during medical decision-making. *Conference on Human Factors in Computing Systems - Proceedings*, 14. https://doi.org/10.1145/3290605.3300234

Cassell, J. (2002). Genderizing Human-Computer Interaction | The Human-Computer Interaction Handbook. In J. A. Jacko & A. Sears (Eds.), *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications* (pp. 401–412). L. Erlbaum Associates Inc. https://dl.acm.org/doi/10.5555/772072.772100

Charness, G., & Gneezy, U. (2012). Strong Evidence for Gender Differences in Risk Taking. *Journal of Economic Behavior & Organization*, *83*(1), 50–58. https://doi.org/10.1016/J.JEBO.2011.06.007

Chen, H., Wood, M. D., Linstead, C., & Maltby, E. (2011). Uncertainty analysis in a GIS-based multi-criteria analysis tool for river catchment management. *Environmental Modelling & Software*, *26*(4), 395–405. https://doi.org/10.1016/J.ENVSOFT.2010.09.005

Chen, J. Y. C., Barnes, M. J., Selkowitz, A. R., & Stowers, K. (2017). Effects of Agent Transparency on human-autonomy teaming effectiveness. *2016 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2016 - Conference Proceedings*, 1838–1843. https://doi.org/10.1109/SMC.2016.7844505

Chen, J. Y. C., Lakhmani, S. G., Stowers, K., Selkowitz, A. R., Wright, J. L., & Barnes, M. (2018). Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomics Science*, *19*(3), 259–282. https://doi.org/10.1080/1463922X.2017.1315750

Chen, J. Y. C., Procci, K., Boyce, M. W., Wright, J. L., Garcia, A., & Barnes, M. (2014). *Situation Awareness-Based Agent Transparency* [US Army Research Laboratory]. https://doi.org/10.21236/ADA600351

Chiesi, F., & Primi, C. (2015). Gender differences in attitudes toward statistics: Is there a case for a confidence gap? *CERME 9 - Ninth Congress of the European Society for Research in Mathematics Educatio*, 622–628. https://hal.science/hal-01287050

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. In *Statistical Power Analysis for the Behavioral Sciences*. Routledge. https://doi.org/10.4324/9780203771587

Cook, W. D., & Seiford, L. M. (2009). Data envelopment analysis (DEA) – Thirty years on. *European Journal of Operational Research*, *192*(1), 1–17. https://doi.org/10.1016/J.EJOR.2008.01.032

Corbett, C., & Hill, C. (2015). *Solving the Equation: The Variables for Women's Success in Engineering and Computing*. American Association of University Women. 1111 Sixteenth Street NW, Washington, DC 20036.

Correll, M., & Gleicher, M. (2014). Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 2142–2151. https://doi.org/10.1109/TVCG.2014.2346298

Cramer, R. J., Neal, T. M. S., & Brodsky, S. L. (2009). Self-Efficacy and Confidence: Theoretical Distinctions and Implications for Trial Consultation. *Consulting Psychology Journal*, *61*(4), 319–334. https://doi.org/10.1037/A0017310

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/XGE0000033

Dilla, W. N., & Steinbart, P. J. (2005). Using Information Display Characteristics to Provide Decision Guidance in a Choice Task under Conditions of Strict Uncertainty. *Journal of Information Systems*, *19*(2), 29–55. https://doi.org/10.2308/JIS.2005.19.2.29

Dong, X., & Hayes, C. C. (2012). Uncertainty Visualizations. *Journal of Cognitive Engineering and Decision Making*, *6*(1), 30–56. https://doi.org/10.1177/1555343411432338

Durndell, A., & Haag, Z. (2002). Computer self efficacy, computer anxiety, attitudes towards the Internet and reported experience with the Internet, by gender, in an East European sample. *Computers in Human Behavior*, *18*(5), 521–535. https://doi.org/10.1016/S0747-5632(02)00006-7

Eberhard, K. (2021). The effects of visualization on judgment and decision-making: a systematic literature review. *Management Review Quarterly 2021 73:1*, *73*(1), 167–214. https://doi.org/10.1007/S11301-021-00235-8

Edwards, J. A., Snyder, F. J., Allen, P. M., Makinson, K. A., & Hamby, D. M. (2012). Decision making for risk management: a comparison of graphical methods for presenting quantitative uncertainty. *Risk Analysis : An Official Publication of the Society for Risk Analysis*, *32*(12), 2055–2070. https://doi.org/10.1111/J.1539-6924.2012.01839.X

Estes, R., & Hosseini, J. (1988). The Gender Gap on Wall Street: An Empirical Analysis of Confidence in Investment Decision Making. *The Journal of Psychology*, *122*(6), 577–590. https://doi.org/10.1080/00223980.1988.9915532

Faccio, M., Marchica, M. T., & Mura, R. (2016). CEO gender, corporate risk-taking, and the efficiency of capital allocation. *Journal of Corporate Finance*, *39*, 193–209. https://doi.org/10.1016/J.JCORPFIN.2016.02.008

Fehr-Duda, H., De Gennaro, M., & Schubert, R. (2006). Gender, financial risk, and probability weights. *Theory and Decision*, *60*(2–3), 283–313. https://doi.org/10.1007/s11238-005-4590-0

Fern, A., Burnett, M., Davidson, J., Doppa, J. R., Pesantez-Cabrera, P., & Kalyanaraman, A. (2024). AgAID Institute—AI for agricultural labor and decision support. *AI Magazine*. https://doi.org/10.1002/AAAI.12156

Fisher, A., & Margolis, J. (2002). Unlocking the clubhouse. *ACM SIGCSE Bulletin*, *34*(2), 79–83. https://doi.org/10.1145/543812.543836

Grigoreanu, V., Cao, J., Kulesza, T., Bogart, C., Rector, K., Burnett, M., & Wiedenbeck, S. (2008). Can feature design reduce the gender gap in end-user software development environments? *Proceedings - 2008 IEEE Symposium on Visual Languages and Human-Centric Computing, VL/HCC 2008*, 149–156. https://doi.org/10.1109/VLHCC.2008.4639077

Gupta, S., Modgil, S., Bhattacharyya, S., & Bose, I. (2022). Artificial intelligence for decision support systems in the field of operations research: review and future scope of research. *Annals of Operations Research*, *308*(1–2), 215–274. https://doi.org/10.1007/S10479-020-03856-6

Hallström, J., Elvstrand, H., & Hellberg, K. (2015). Gender and technology in free play in Swedish early childhood education. *International Journal of Technology and Design Education*, *25*(2), 137–149. https://doi.org/10.1007/S10798-014-9274-Z/METRICS

Hartzel, K. (2003). How self-efficacy and gender issues affect software adoption and use. *Communications of the ACM*, *46*(9), 167–171. https://doi.org/10.1145/903893.903933

Harvey, N. (1997). Confidence in judgment. *Trends in Cognitive Sciences*, *1*(2), 78–82. https://doi.org/10.1016/S1364-6613(97)01014-0

He, X., Inman, J. J., & Mittal, V. (2008). Gender Jeopardy in financial risk taking. *Journal of Marketing Research*, *45*(4), 414–424. https://doi.org/10.1509/jmkr.45.4.414

Hyde, K., Maier, H. R., & Colby, C. (2003). Incorporating uncertainty in the PROMETHEE MCDA method. *Journal of Multi-Criteria Decision Analysis*, *12*(4–5), 245–259. https://doi.org/10.1002/MCDA.361

Jianakoplos, N. A., & Bernasek, A. (1998). Are Women More Risk Averse? *Economic Inquiry*, *36*(4), 620–630. https://doi.org/10.1111/J.1465-7295.1998.TB01740.X

Kahneman, D. (2013). *Thinking, Fast and Slow*. Farrar, Straus, and Giroux.

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020). Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3313831.3376219

Kelleher, C. (2009). Barriers to Programming Engagement. *Advances in Gender and Education*, *1*, 5–10.

Klayman, J., Soll, J. B., González-Vallejo, C., & Barlas, S. (1999). Overconfidence: It Depends on How, What, and Whom You Ask. *Organizational Behavior and Human Decision Processes*, *79*(3), 216–247. https://doi.org/10.1006/OBHD.1999.2847

Krieger, S., Allen, M., & Rawn, C. (2015). Are females disinclined to tinker in computer science? *SIGCSE 2015 - Proceedings of the 46th ACM Technical Symposium on Computer Science Education*, 102–107. https://doi.org/10.1145/2676723.2677296

Laerd Statistics. (2017). *Statistical tutorials and software guides*. Statistical Tutorials and Software Guides. https://statistics.laerd.com/

Lauriola, M., & Levin, I. P. (2001). Personality traits and risky decision-making in a controlled experimental task: an exploratory study. *Personality and Individual Differences*, *31*(2), 215–226. https://doi.org/10.1016/S0191-8869(00)00130-6

Lem, S., Onghena, P., Verschaffel, L., & Van Dooren, W. (2013). On the misinterpretation of histograms and box plots. *Educational Psychology*, *33*(2), 155–174. https://doi.org/10.1080/01443410.2012.674006

Loft, S., Bhaskara, A., Lock, B. A., Skinner, M., Brooks, J., Li, R., & Bell, J. (2023). The Impact of Transparency and Decision Risk on Human–Automation Teaming Outcomes. *Human Factors*, *65*(5), 846–861. https://doi.org/10.1177/00187208211033445

Lundeberg, M. A., Fox, P. W., & Punćochaŕ, J. (1994). Highly Confident but Wrong: Gender Differences and Similarities in Confidence Judgments. *Journal of Educational Psychology*, *86*(1), 114–121. https://doi.org/10.1037/0022-0663.86.1.114

MacEachren, A. M. (1992). Visualizing Uncertain Information. *Cartographic Perspectives*, *13*, 10–19. https://doi.org/10.14714/CP13.1000

MacEachren, A. M., Robinson, A., Hopper, S., Gardner, S., Murray, R., Gahegan, M., & Hetzler, E. (2005). Visualizing Geospatial Information Uncertainty: What We Know and What We Need to Know. *Cartography and Geographic Information Science*, *32*(3), 139–160. https://doi.org/10.1559/1523040054738936

MacEachren, A. M., Roth, R. E., O'Brien, J., Li, B., Swingley, D., & Gahegan, M. (2012). Visual Semiotics &amp; Uncertainty Visualization: An Empirical Study. *IEEE Transactions on Visualization and Computer Graphics*, *18*(12), 2496–2505. https://doi.org/10.1109/TVCG.2012.279

Mercado, J. E., Rupp, M. A., Chen, J. Y. C., Barnes, M. J., Barber, D., & Procci, K. (2016). Intelligent Agent Transparency in Human-Agent Teaming for Multi-UxV Management. *Human Factors*, *58*(3), 401–415. https://doi.org/10.1177/001872081562120

Meyers-Levy, J. (1988). Gender Differences in Information Processing: A Selectivity Interpretation. In P. Cafferata & A. Tybout (Eds.), *Cognitive and Affective Responses to Advertising* (pp. 219–260). Lexington Books.

Meyers-Levy, J., & Loken, B. (2015). Revisiting gender differences: What we know and what lies ahead. *Journal of Consumer Psychology*, *25*(1), 129–149. https://doi.org/10.1016/j.jcps.2014.06.003

Meyers-Levy, J., & Maheswaran, D. (1991). Exploring Differences in Males' and Females' Processing Strategies. *Journal of Consumer Research*, *18*(1), 63. https://doi.org/10.1086/209241

Meyers-Levy, J., & Sternthal, B. (1991). Gender Differences in the Use of Message Cues and Judgments. *Journal of Marketing Research*, *28*(1), 84. https://doi.org/10.2307/3172728

Meyers-Levy, J., & Zhu, R. J. (2010). Gender differences in the meanings consumers infer from music and other aesthetic stimuli. *Journal of Consumer Psychology*, *20*(4), 495–507. https://doi.org/10.1016/J.JCPS.2010.06.006

Miskioglu, E., & Martin, K. M. (2019). Is it Rocket Science or Brain Science? Developing an Instrument to Measure "Engineering Intuition." *ASEE Annual Conference and Exposition, Conference Proceedings*. https://doi.org/10.18260/1-2--33027

Noseworthy, T. J., Cotte, J., & Lee, S. H. (2011). The Effects of Ad Context and Gender on the Identification of Visually Incongruent Products. *Journal of Consumer Research*, *38*(2), 358–375. https://doi.org/10.1086/658472

O'Neill, T. A., Flathmann, C., McNeese, N. J., & Salas, E. (2023). Human-autonomy Teaming: Need for a guiding team-based framework? *Computers in Human Behavior*, *146*, 107762. https://doi.org/10.1016/J.CHB.2023.107762

O'Neill, T., McNeese, N. J., Barron, A., & Schelble, B. (2022). Human-Autonomy Teaming: A Review and Analysis of the Empirical Literature. *Human Factors*, *64*(5), 904–938. https://doi.org/10.1177/0018720820960865

Padilla, L. M., Hansen, G., Ruginski, I. T., Kramer, H. S., Thompson, W. B., & Creem-Regehr, S. H. (2015). The influence of different graphical displays on nonexpert decision making under uncertainty. *Journal of Experimental Psychology: Applied*, *21*(1), 37–46. https://doi.org/10.1037/XAP0000037

Pallier, G., Wilkinson, R., Danthiir, V., Kleitman, S., Knezevic, G., Stankov, L., & Roberts, R. D. (2002). The Role of Individual Differences in the Accuracy of Confidence Judgments. *The Journal of General Psychology*, *129*(3), 257–299. https://doi.org/10.1080/00221300209602099

Parasuraman, R., & Manzey, D. H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. *Https://Doi.Org/10.1177/0018720810376055*, *52*(3), 381–410. https://doi.org/10.1177/0018720810376055

Pelissari, R., Oliveira, M. C., Abackerli, A. J., Ben-Amor, S., & Assumpção, M. R. P. (2018). Techniques to model uncertain input data of multi-criteria decision-making problems: a literature review. *International Transactions in Operational Research*, *28*(2), 523–559. https://doi.org/10.1111/ITOR.12598

Pfaff, M. S., Klein, G. L., Drury, J. L., Moon, S. P., Liu, Y., & Entezari, S. O. (2013). Supporting complex decision making through option awareness. *Journal of Cognitive Engineering and Decision Making*, *7*(2), 155–178. https://doi.org/https://doi.org/10.1177/1555343412455799

Powell, M., & Ansic, D. (1997). Gender differences in risk behaviour in financial decision-making: An experimental analysis. *Journal of Economic Psychology*, *18*(6), 605–628. https://doi.org/10.1016/S0167-4870(97)00026-3

Pugh, S. (1981). Concept Selection: A Method that Works. In S. Pugh (Ed.), *Proceedings of International Conference on Engineering Design* (pp. 497–506).

Reeves, R. V. (2022). *Of Boys And Men: Why The Modern Male Is Struggling, Why It Matters, And What To Do About It*. Brookings Institution Press.

Reichheld, F. F. (2003). The one number you need to grow. *Harvard Business Review*, *81*(12), 46-54+124.

Shim, J. P., Warkentin, M., Courtney, J. F., Power, D. J., Sharda, R., & Carlsson, C. (2002). Past, present, and future of decision support technology. *Decision Support Systems*, *33*(2), 111–126. https://doi.org/10.1016/S0167-9236(01)00139-7

Stowers, K., Kasdaglis, N., Rupp, M. A., Newton, O. B., Chen, J. Y. C., & Barnes, M. J. (2020). The IMPACT of Agent Transparency on Human Performance. *IEEE Transactions on Human-Machine Systems*, *50*(3), 245–253. https://doi.org/10.1109/THMS.2020.2978041

Tatasciore, M., Bowden, V., & Loft, S. (2023). Do concurrent task demands impact the benefit of automation transparency? *Applied Ergonomics*, *110*. https://doi.org/10.1016/J.APERGO.2023.104022

Thompson, M. M., Naccarato, M. E., Parker, K. C. H., & Moskowitz, G. B. (2013). The Personal Need for Structure and Personal Fear of Invalidity Measures: Historical Perspectives, Current Applications, and Future Directions. *Cognitive Social Psychology*, 19–39. https://doi.org/10.4324/9781410605887-3

Tomsett, R., Preece, A., Braines, D., Cerutti, F., Chakraborty, S., Srivastava, M., Pearson, G., & Kaplan, L. (2020). Rapid Trust Calibration through Interpretable and Uncertainty-Aware AI. *Patterns*, *1*(4). https://doi.org/10.1016/j.patter.2020.100049

Tsakalerou, M., Efthymiadis, D., & Abilez, A. (2022). An intelligent methodology for the use of multi-criteria decision analysis in impact assessment: the case of real-world offshore construction. *Scientific Reports 2022 12:1*, *12*(1), 1–14. https://doi.org/10.1038/s41598-022-19554-1

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/SCIENCE.185.4157.1124

Vaidya, O. S., & Kumar, S. (2006). Analytic hierarchy process: An overview of applications. *European Journal of Operational Research*, *169*(1), 1–29. https://doi.org/10.1016/J.EJOR.2004.04.028

van de Merwe, K., Mallam, S., & Nazir, S. (2024). Agent Transparency, Situation Awareness, Mental Workload, and Operator Performance: A Systematic Literature Review. *Human Factors*, *66*(1), 180–208. https://doi.org/https://doi.org/10.1177/00187208221077804

Vorvoreanu, M., Zhang, L., Huang, Y. H., Hilderbrand, C., Steine-Hanson, Z., & Burnett, M. (2019). From gender biases to gender-inclusive design: An empirical investigation. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3290605.3300283

Wagh, A., Cook-Whitt, K., & Wilensky, U. (2017). Bridging inquiry-based science and constructionism: Exploring the alignment between students tinkering with code of computational models and goals of inquiry. *Journal of Research in Science Teaching*, *54*(5), 615–641. https://doi.org/10.1002/TEA.21379

Wang, L., & Yu, Z. (2023). Gender-moderated effects of academic self-concept on achievement, motivation, performance, and self-efficacy: A systematic review. *Frontiers in Psychology*, *14*, 1136141. https://doi.org/https://doi.org/10.3389/fpsyg.2023.1136141

Weber, E. U., Blais, A. R., & Betz, N. E. (2002). A Domain-specific Risk-attitude Scale: Measuring Risk Perceptions and Risk Behaviors. *Journal of Behavioral Decision Making*, *15*(4), 263–290. https://doi.org/10.1002/BDM.414

Wilke, C. O. (2019). *Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures*. O’Reilly Media.

Wohleber, R. W., Stowers, K., Barnes, M., & Chen, J. Y. C. (2023). Agent transparency in mixed-initiative multi-UxV control: How should intelligent agent collaborators speak their minds? *Computers in Human Behavior*, *148*, 107866. https://doi.org/10.1016/J.CHB.2023.107866

Wynne, K. T., & Lyons, J. B. (2018). An integrative model of autonomous agent teammate-likeness. *Theoretical Issues in Ergonomics Science*, *19*(3), 353–374. https://doi.org/10.1080/1463922X.2016.1260181

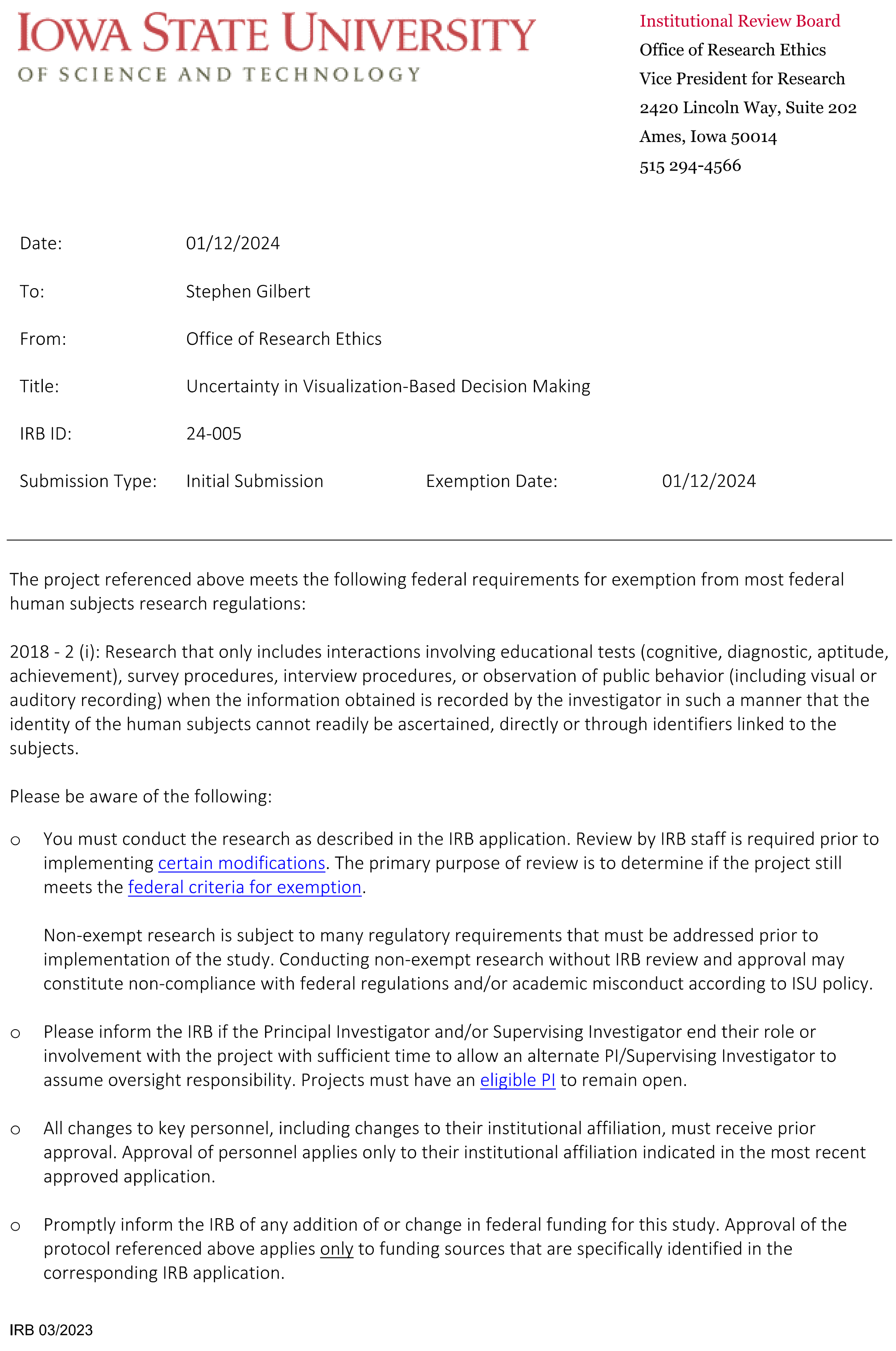
Yang, Q., Steinfeld, A., & Zimmerman, J. (2019). Unremarkable AI: Fitting intelligent decision support into critical, clinical decision-making processes. *Conference on Human Factors in Computing Systems - Proceedings*, *11*. https://doi.org/10.1145/3290605.3300468

Yuan, L., Gao, X., Zheng, Z., Edmonds, M., Wu, Y. N., Rossano, F., Lu, H., Zhu, Y., & Zhu, S. C. (2022). In situ bidirectional human-robot value alignment. *Science Robotics*, *7*(68), 4183. https://doi.org/https://doi.org/10.1126/scirobotics.abm4183

Zakay, D., & Wooler, S. (1984). Time pressure, training and decision effectiveness. *Ergonomics, 27*(3), 273–284. https://doi.org/10.1080/00140138408963489

Zuk, T., & Carpendale, S. (2006). Theoretical analysis of uncertainty visualizations. *Proceedings Volume 6060, Visualization and Data Analysis, 6060*, 66–79. https://doi.org/10.1117/12.643631

# APPENDIX A. IRB APPROVAL LETTER



IOWA STATE UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Institutional Review Board
Office of Research Ethics
Vice President for Research
2420 Lincoln Way, Suite 202
Ames, Iowa 50014
515 294-4566

|  |  |  |  |
|---|---|---|---|
| Date: | 01/12/2024 |  |  |
| To: | Stephen Gilbert |  |  |
| From: | Office of Research Ethics |  |  |
| Title: | Uncertainty in Visualization-Based Decision Making |  |  |
| IRB ID: | 24-005 |  |  |
| Submission Type: | Initial Submission | Exemption Date: | 01/12/2024 |

The project referenced above meets the following federal requirements for exemption from most federal human subjects research regulations:

2018 - 2 (i): Research that only includes interactions involving educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures, or observation of public behavior (including visual or auditory recording) when the information obtained is recorded by the investigator in such a manner that the identity of the human subjects cannot readily be ascertained, directly or through identifiers linked to the subjects.

Please be aware of the following:

- You must conduct the research as described in the IRB application. Review by IRB staff is required prior to implementing certain modifications. The primary purpose of review is to determine if the project still meets the federal criteria for exemption.

  Non-exempt research is subject to many regulatory requirements that must be addressed prior to implementation of the study. Conducting non-exempt research without IRB review and approval may constitute non-compliance with federal regulations and/or academic misconduct according to ISU policy.

- Please inform the IRB if the Principal Investigator and/or Supervising Investigator end their role or involvement with the project with sufficient time to allow an alternate PI/Supervising Investigator to assume oversight responsibility. Projects must have an eligible PI to remain open.

- All changes to key personnel, including changes to their institutional affiliation, must receive prior approval. Approval of personnel applies only to their institutional affiliation indicated in the most recent approved application.

- Promptly inform the IRB of any addition of or change in federal funding for this study. Approval of the protocol referenced above applies only to funding sources that are specifically identified in the corresponding IRB application.

IRB 03/2023

- A brief status update is required every three years to keep IRB oversight active. You will be notified as this date approaches.

- Immediately inform the IRB of (1) all serious and/or unexpected adverse experiences involving risks to subjects or others; and (2) any other unanticipated problems involving risks to subjects or others.

- Approval from other entities may also be needed. For example, access to data from private records (e.g., student, medical, or employment records, etc.) that are protected by FERPA, HIPAA or other confidentiality policies requires permission from the holders of those records. Similarly, for research conducted in institutions other than ISU (e.g., schools, other colleges or universities, medical facilities, companies, etc.), investigators must obtain permission from the institution(s) as required by their policies. An IRB determination of exemption in no way implies or guarantees that permission from these other entities will be granted.

- Your research study may be subject to post-approval monitoring by Iowa State University's Office of Research Ethics. It may also be subject to formal audit or inspection by federal agencies and study sponsors.

- Upon completion of the project, transfer of IRB oversight to another IRB, or departure of the PI and/or Supervising Investigator, please initiate a Project Closure in IRBManager to officially close the project. For information on instances when a study may be closed, please refer to the IRB Study Closure Policy.

- All research involving human participants must undergo IRB review. Only the IRB or its designees may make the determination of exemption, even if you conduct a study in the future that is exactly like this study.

If you have questions or concerns, please do not hesitate to contact us at 515-294-4566 or IRB@iastate.edu.

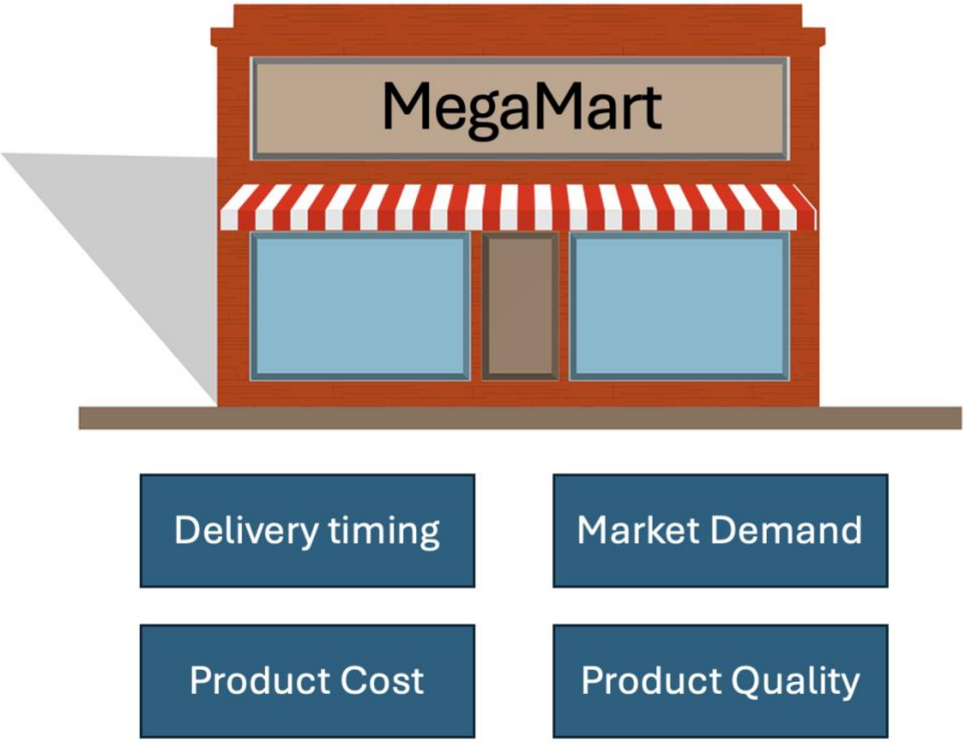
# APPENDIX B. TASK TRAINING MATERIAL

The training video consisted of four modules, which were combined to create a unique training video for each of the six conditions according to the table below. Each module included a static image with voiceover text (transcripts below).

| | Cond. 1 | Cond. 2 | Cond. 3 | Cond. 4 | Cond. 5 | Cond. 6 |
|---|---|---|---|---|---|---|
| Module 1 Introduction | 2 | 1 | 1 | 2 | 1 | 1 |
| Module 2. Explanation of Charts | 1 | 1 | 1 | 2 | 2 | 2 |
| Module 3. Example | 1 | 1 | 1 | 2 | 2 | 2 |
| Module 4. Explanation of Recommendations | N/A | 1 | 1 | N/A | 2 | 2 |

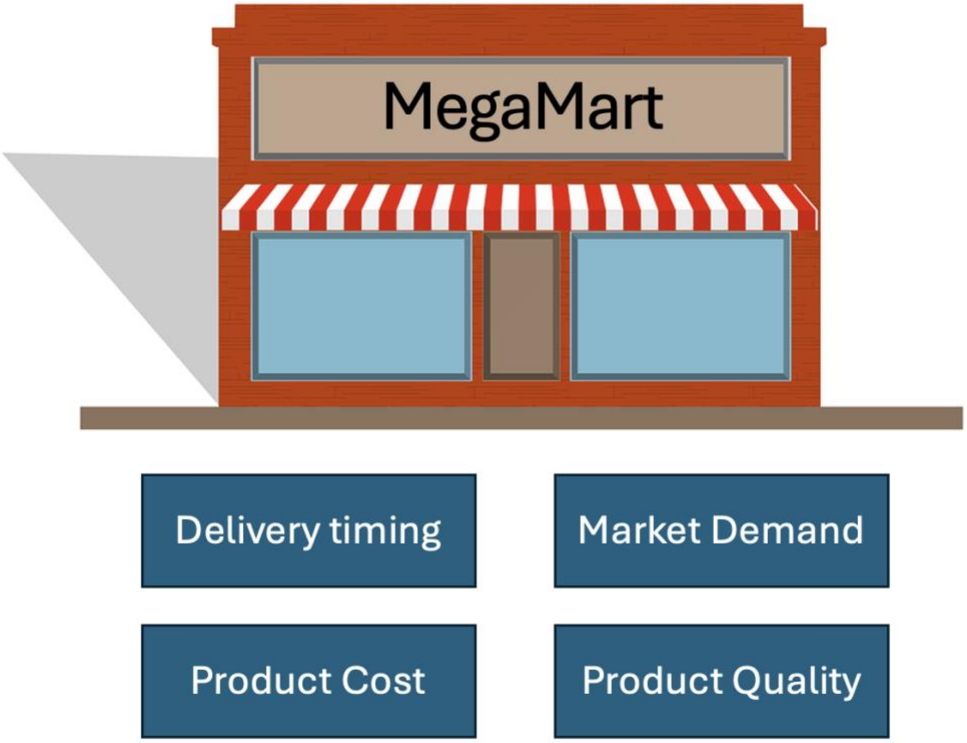
## Module 1: Introduction

*Version 1 (Conditions 2, 3, 5, 6, with ProdStar recommendations)*



MegaMart is always trying to keep up with their customers, and now they’re looking for help deciding which new products to start selling in their stores. They are trying to balance the importance of timing, cost, market demand, and quality to choose the most successful products. They put together information to compare the options, and they’re using a new tool called ProdStar to recommend the best option.

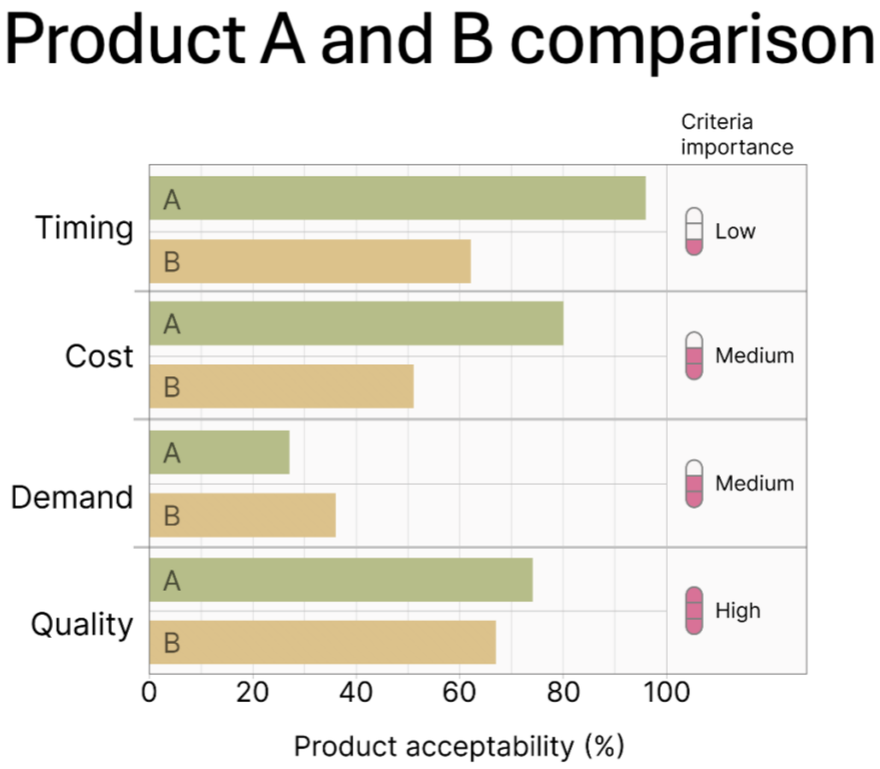
*Version 2 (Conditions 1 and 4, without ProdStar recommendations)*



MegaMart is always trying to keep up with their customers, and now they're looking for help deciding which new products to start selling in their stores. They are trying to balance the importance of timing, cost, market demand, and quality to choose the most successful products.

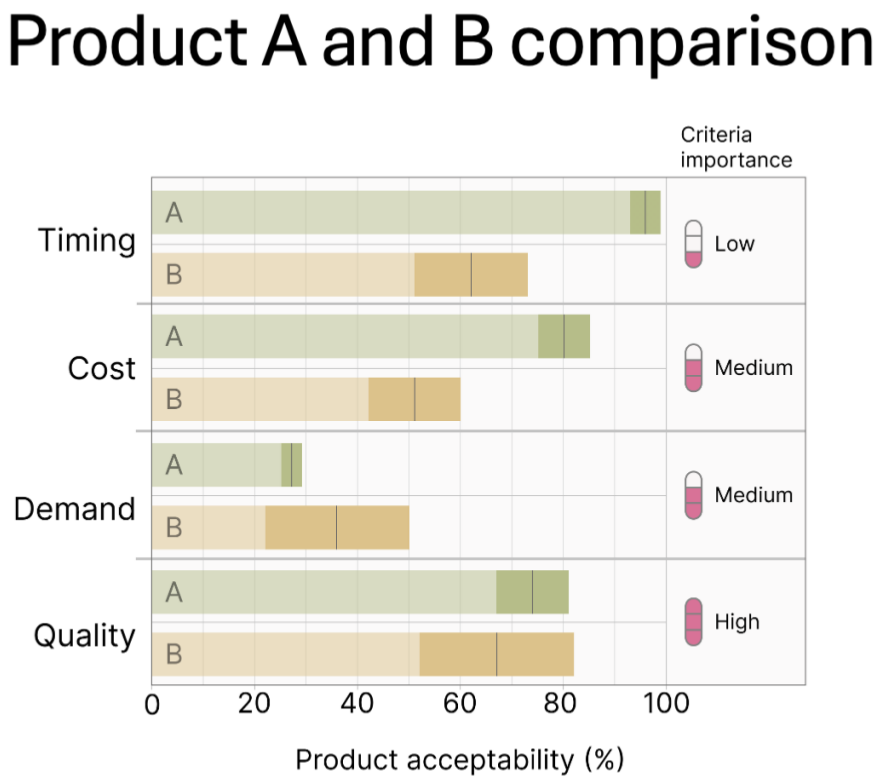
**Module 2: Explanation of Charts**

*Version 1 (Conditions 1, 2, and 3, charts without uncertainty)*



These charts show how good of a fit each product is for their four decision criteria. The icons and text on the right side of the chart indicate how important each criterion is to MegaMart. Keep an eye on these, since their relative importance will vary from comparison to comparison.

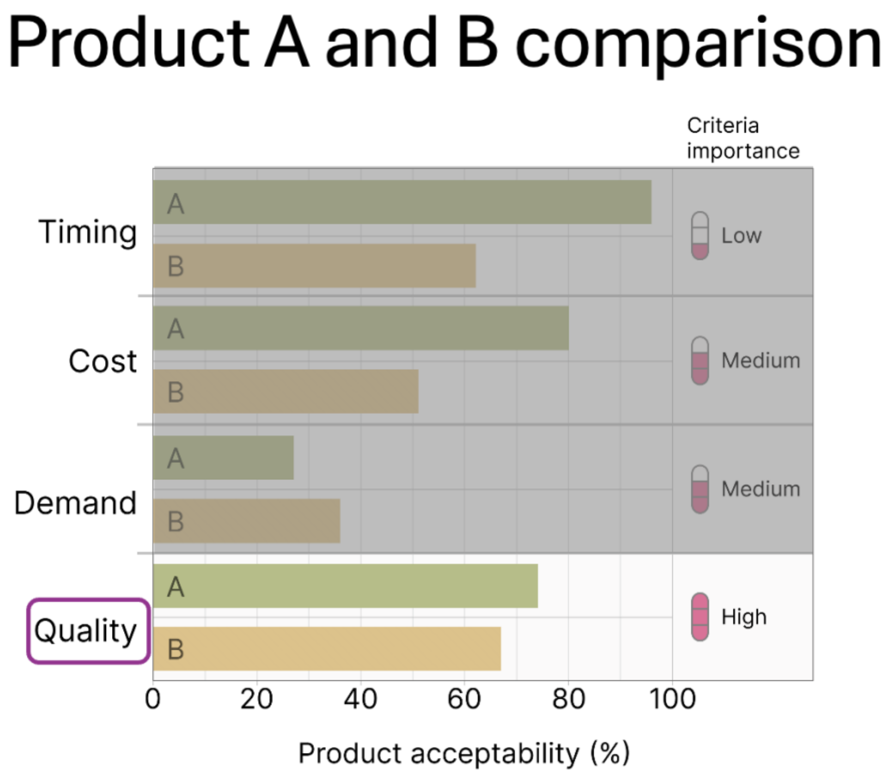
*Version 2 (Conditions 4, 5, and 6, charts with uncertainty)*



These charts show how good of a fit each product is for their four decision criteria. Sometimes MegaMart must make decisions with uncertainty in their data, so this chart shows a range of possible scores. These are 95% confidence intervals, which means that 95% of the time, the actual score would be in this range. The icons and text on the right side of the chart indicate how important each criterion is to MegaMart. Keep an eye on these, since their relative importance will vary from comparison to comparison.

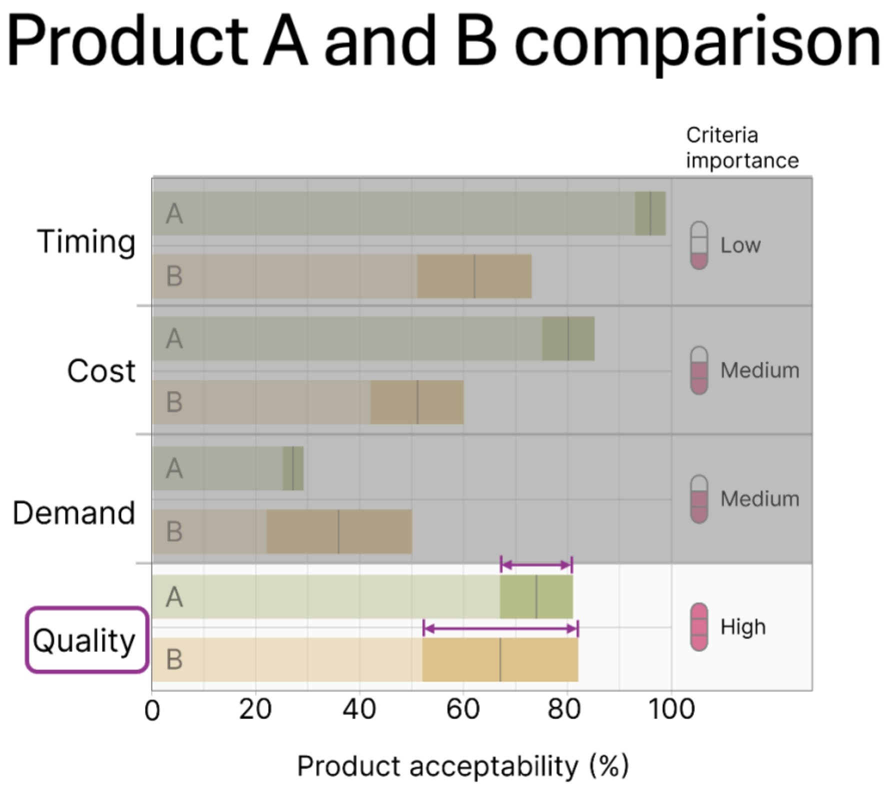
**Module 3: Example**

*Version 1 (Conditions 1, 2, and 3, charts without uncertainty)*



In this example, MegaMart's highest priority is product quality. Product A has a slightly higher score for product quality. This might suggest that Product A is the best choice, but they need to consider the other factors too.

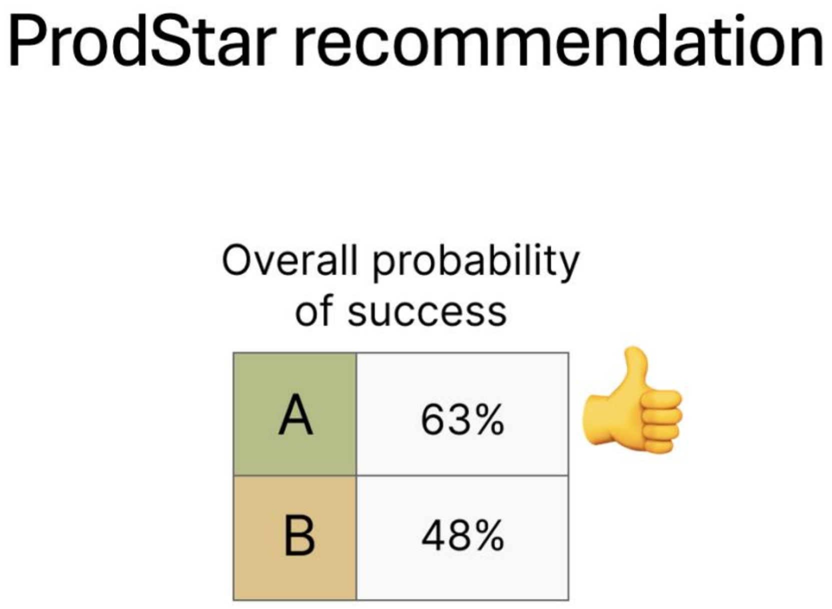
*Version 2 (Conditions 4, 5, and 6, charts with uncertainty)*



In this example, MegaMart's highest priority is product quality. The widths of the green and yellow bars show that there's about twice as much uncertainty for Product B than Product A, and Product A looks like it might be the better option. But MegaMart needs to also consider the other criteria.

**Module 4: Explanation of Recommendations (omitted from Conditions 1 and 4)**

*Version 1 (Conditions 2 and 5, basic recommendations)*



ProdStar recommendation

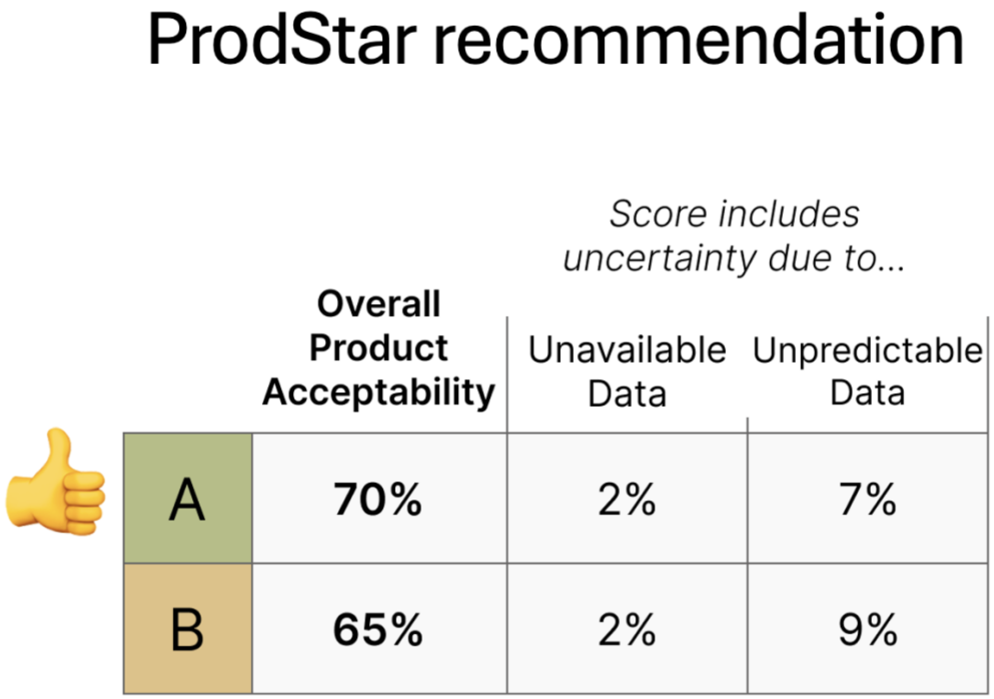
Overall probability of success

| | | |
|---|---|---|
| A | 63% | 👍 |
| B | 48% | |

MegaMart is using an automated system called ProdStar, which offers product recommendations.

But just like a weather forecast, it’s not always correct.

*Version 2 (Conditions 3 and 6, detailed recommendations)*



## ProdStar recommendation

|  |  | *Score includes uncertainty due to...* |  |
|---|---|---|---|
|  | **Overall Product Acceptability** | Unavailable Data | Unpredictable Data |
| 👍 A | **70%** | 2% | 7% |
| B | **65%** | 2% | 9% |

MegaMart is using an automated system called ProdStar, which offers product recommendations. This helps them analyze all the decision factors and uncertainty in the data. It also tells them if uncertainty in the scores is due to unavailable data, like when their marketing team hasn’t gathered enough information, or due to unpredictable data, like volatility in market demand for their products. But just like a weather forecast, ProdStar is not always correct.

# APPENDIX C. SURVEY

## Informed Consent

Q11 Title of Study: **Uncertainty in Visualization-Based Decision Making**

Investigators: **Amanda Newendorp and Stephen Gilbert, Ph.D.**

This form describes a research project. It has information to help you decide whether or not you wish to participate. Research studies only include people who choose to take part – your participation is completely voluntary. Please discuss any questions you have about the study or about this form with the project staff before deciding to participate.

**CONCISE SUMMARY**
We are interested in understanding whether different people have different styles of making decisions with different information. This anonymous survey asks you to help a fictional company select a new product to sell in their store.

Participation will take approximately 15 minutes. Afterward, you will have the option to enter a drawing for one of three $25 eGift cards.

**INTRODUCTION**
We are interested in how different people make decisions with multiple decision factors. To participate in this study, you must be 18 years of age or older.

**DESCRIPTION OF PROCEDURES**
If you agree to participate, you will watch a brief (approximately 1.5 min.) video and answer a set of sample questions for training. Then, you will be asked to complete twelve scenarios, in which you will select the optimal choice and rate your confidence in that selection. These questions will be timed; you will have 45 seconds to make each selection. After this, you will be asked answer questions about your experience in this study, as well as questions about your demographics, decision-making preferences, risk tolerance, and attitudes toward technology.

**RISKS**
If you are uncomfortable with time pressure, this survey may lead to negative feelings that you'd prefer to avoid by not participating.

**BENEFITS**
If you decide to participate in this study, there will be no direct benefit to you. However, it is hoped that the information gained in this study will benefit society by improving our understanding of decision-making.

**COSTS AND COMPENSATION**

You will not have any costs based on your participation in this study. After the survey, if you choose, you may type your name and email to enter a drawing to win one of three $25 USD eGift cards. If you win, it will be emailed to you. Please know that payments may be subject to tax withholding requirements, which vary depending on if you are a legal resident of the U.S. or another country.

**PARTICIPANT RIGHTS**

Your participation in this study is completely voluntary and you may refuse to participate or leave the study at any time. If you decide to not participate in the study or leave the study early, it will not result in any penalty or loss of benefits to which you are otherwise entitled. If you are a student or employee at Iowa State University, your status will not be affected by your decision to participate or not. This research study is not related to any ISU course or any course assignments, meaning your participation is not a course requirement. Your course grade and/or standing in ISU courses will not be influenced by your decision to participate or not. You can skip any questions that you do not wish to answer.

If you have any questions about the rights of research subjects or research-related injury, please contact the IRB Administrator, (515) 294-4566, IRB@iastate.edu, or Director, (515) 294-3115, Office of Research Ethics, Iowa State University, Ames, Iowa 50011.

**CONFIDENTIALITY**

Records identifying participants' personally identifiable information (PII) will be kept confidential to the extent permitted by applicable laws and regulations and will not be made publicly available. However, U.S. federal government regulatory agencies, auditing departments of Iowa State University, and the Institutional Review Board (a committee that reviews and approves human subject research studies) may inspect and/or copy your records for quality assurance and data analysis. These records may contain private information.

The particular results of your participation in this study will not be directly linked to your personal identity. Your answers to the free-response questions will be scanned for any information that might possibly identify you, (e.g., "I made decisions like this when I worked at Company XYZ"), and that information will be removed so that all results data are de-identified. Your de-identified data may be shared with other researchers or used in future research without your additional consent. De-identified data may also be shared on a data repository such as Open Science Framework, for consideration for publication in journals with an Open Data requirement or recommendation.

**CONTACT INFORMATION**

If you have any questions or concerns, you can contact Amanda Newendorp (aknowen@iastate.edu) or Stephen Gilbert (gilbert@iastate.edu).

You may print this document for your records if you desire.

**By clicking Agree below, you agree that you are 18 years or older and that you're willing to participate.**

○ Agree. I am 18 years old or older, and I will participate in the study. (1)

○ Disagree. No, thank you. I'd prefer not to participate. (2)

**Training**

Q85 Please watch this video for instructions.

Training video, see details in APPENDIX B.

After the video finishes, select "next page."

**Practice Trial and 12 Experiment Trials**

**Page 1, auto-advances after 45 seconds**

Dashboard display (REF to figure)

Which product should MegaMart choose to sell in their store?

○ A

○ B

---

Should MegaMart seek additional information before making a final decision?

○ No

○ Yes

**Page 2, untimed**

You selected Product ${Trn_Cond[#]_select/ChoiceGroup/SelectedChoices}. How confident are you that you selected the optimal product?

○ 1 - Not confident at all

○ 2 - Slightly confident

○ 3 - Somewhat confident

○ 4 - Fairly confident

○ 5 - Very confident

**Post-Task Survey**

Next, please answer the following questions about how you made your decisions.

*Display This Question:*

*If Visualization_condition = 1*

*Or Visualization_condition = 2*

*Or Visualization_condition = 3*

How much did you rely on the bar charts to make your decisions?

| | Not at all | | | Very much | |
|---|---|---|---|---|---|
| 1 | 2 | 3 | 3 | 4 | 5 |

*Display This Question:*

*If Visualization_condition = 4*

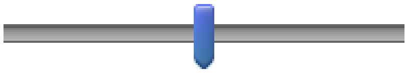
*Or Visualization_condition = 5*

*Or Visualization_condition = 6*

How much did you rely on the bar charts to make your decisions?

Not at all Very much

1 2 3 3 4 5



*Display This Question:*

*If Visualization_condition = 4*

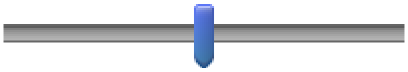
*Or Visualization_condition = 5*

*Or Visualization_condition = 6*

How much did you rely on the ranges of uncertainty to make your decisions?

Not at all Very much

1 2 3 3 4 5



How much did you rely on the criteria weights to make your decisions?

| | | | | | | |
|---|---|---|---|---|---|---|
| | | Not at all | | | Very much | |
| | 1 | 2 | 3 | 3 | 4 | 5 |

*Display This Question:*

*If Visualization_condition = 2*

*Or Visualization_condition = 3*

*Or Visualization_condition = 5*

*Or Visualization_condition = 6*

How much did you rely on the ProdStar recommendations to make your decisions?

| | | | | | | |
|---|---|---|---|---|---|---|
| | | Not at all | | | Very much | |
| | 1 | 2 | 3 | 3 | 4 | 5 |

*Display This Question:*

*If Visualization_condition = 3*

*Or Visualization_condition = 6*

How much did you rely on the ProdStar's explanation of uncertainty?

| | | Not at all | | | Very much | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 3 | 4 | 5 |

*Display This Question:*

*If Visualization_condition = 2*

*Or Visualization_condition = 3*

*Or Visualization_condition = 5*

*Or Visualization_condition = 6*

How much did you trust the recommendations to be accurate?

| | | Not at all | | | Very much | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 3 | 4 | 5 |

*Display This Question:*

*If Visualization_condition = 2*

*Or Visualization_condition = 3*

*Or Visualization_condition = 5*

*Or Visualization_condition = 6*

How accurate do you think the recommendations were?

| | Not at all (0% accurate) | Very much (100% accurate) |
|---|---|---|

| | 0 | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|---|

How likely is it that you would recommend this type of information display to a friend or colleague that is interested in decision-making support?

| | Not likely at all | Extremely likely |
|---|---|---|

| | 1 | 2 | 3 | 4 | 5 | 6 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|

Please briefly describe your process for deciding between A and B in a few sentences or bullet points.

________________________________________________________________

What did you like about the way information was displayed in this study?

________________________________________________________________

What did you NOT like about the way information was displayed in this study?

________________________________________________________________

End of Block: Post-task survey

Start of Block: System Usability Scale

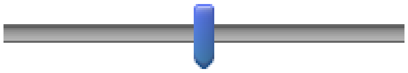
Next, please answer the following questions about this decision support tool.

Assume that "tool" refers to all of the information that was provided to help you make a product decision.

I think that I would like to use this tool frequently.

Strongly disagree Strongly agree

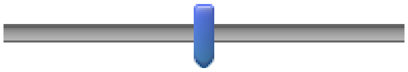
1 2 3 3 4 5



I found this tool unnecessarily complex.

Strongly disagree Strongly agree

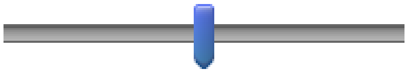
1 2 3 3 4 5



I thought this tool was easy to use.

Strongly disagree Strongly agree

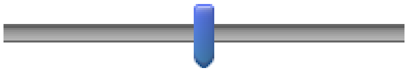
1 2 3 3 4 5



I think that I would need the support of a technical person to be able to use this tool.

Strongly disagree Strongly agree

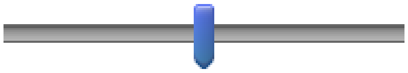
1 2 3 3 4 5



I found the various functions in this tool were well integrated.

Strongly disagree Strongly agree

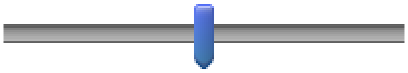
1 2 3 3 4 5



I thought there was too much inconsistency in this tool.

Strongly disagree Strongly agree

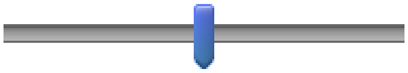
1 2 3 3 4 5



I would imagine that most people would learn to use this tool very quickly.

Strongly disagree Strongly agree

1 2 3 3 4 5



I found the tool very cumbersome to use.

Strongly disagree Strongly agree

1 2 3 3 4 5

I felt very confident using this tool.

Strongly disagree Strongly agree

1 2 3 3 4 5

I needed to learn a lot of things before I could get going with this tool.

Strongly disagree Strongly agree

1 2 3 3 4 5

End of Block: System Usability Scale

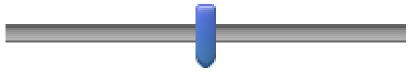
Start of Block: GenderMag facets

Please rate how strongly you agree or disagree with the following statements.

I am able to use new software and technology when I have just the built-in help for assistance.

| Disagree Completely | Neither Agree Nor Disagree | Agree Completely |
|---|---|---|

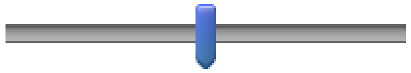
1 2 3 3 4 5 6 7 7 8 9



I am able to use new software and technology when I have seen someone else using it before trying it myself.

| Disagree Completely | Neither Agree Nor Disagree | Agree Completely |
|---|---|---|

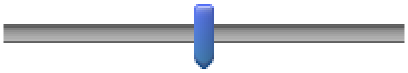
1 2 3 3 4 5 6 7 7 8 9



I am able to use new software and technology when no one is around to help if I need it.

| Disagree Completely | Neither Agree Nor Disagree | Agree Completely |
|---|---|---|

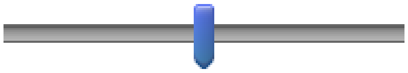
| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 3 | 4 | 5 | 6 | 7 | 7 | 8 | 9 |



I am able to use new software and technology when someone else has helped me get started.

| | | |
|---|---|---|
| Disagree Completely | Neither Agree Nor Disagree | Agree Completely |

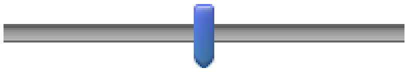
| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 3 | 4 | 5 | 6 | 7 | 7 | 8 | 9 |



I am able to use new software and technology when someone shows me how to do it first.

| | | |
|---|---|---|
| Disagree Completely | Neither Agree Nor Disagree | Agree Completely |

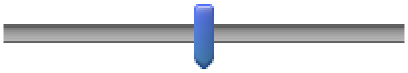
| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 3 | 4 | 5 | 6 | 7 | 7 | 8 | 9 |



I am able to use new software and technology when I have used similar technology before, to do the same task.

Disagree Completely | Neither Agree Nor Disagree | Agree Completely

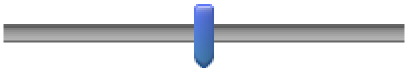
1 2 3 3 4 5 6 7 7 8 9



I am able to use new software and technology when I have never used anything like it before.

Disagree Completely | Neither Agree Nor Disagree | Agree Completely

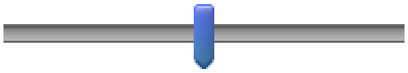
1 2 3 3 4 5 6 7 7 8 9



I am **not** confident about my ability to use and learn new software and technology. I have other strengths.

Disagree Completely | Neither Agree Nor Disagree | Agree Completely

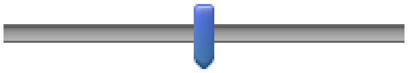
1 2 3 3 4 5 6 7 7 8 9



I make time to explore new software and technology that is not critical to my job.

| Disagree Completely | Neither Agree Nor Disagree | Agree Completely |
|---|---|---|

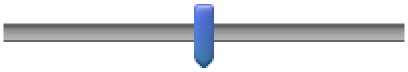
1 2 3 3 4 5 6 7 7 8 9



One reason I spend time and money on new software and technology is because it's a way for me to look good with peers.

| Disagree Completely | Neither Agree Nor Disagree | Agree Completely |
|---|---|---|

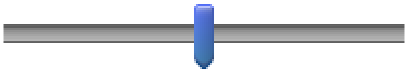
1 2 3 3 4 5 6 7 7 8 9



It's fun to try new software and technology that is not yet available to everyone, such as being a participant in beta programs to test unfinished new software and technology.

| Disagree Completely | Neither Agree Nor Disagree | Agree Completely |
|---|---|---|

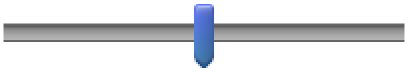
1 2 3 3 4 5 6 7 7 8 9

| | |
|---|---|
| 1 () |  |

I enjoy finding the lesser-known features and capabilities of the new software and technology I use.

| Disagree Completely | Neither Agree Nor Disagree | Agree Completely |
|---|---|---|

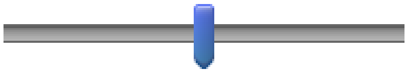
1 2 3 3 4 5 6 7 7 8 9

| | |
|---|---|
| |  |

I explore areas of new software and technology before it is time for me to use it.

| Disagree Completely | Neither Agree Nor Disagree | Agree Completely |
|---|---|---|

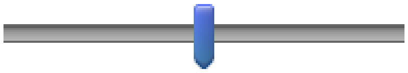
1 2 3 3 4 5 6 7 7 8 9

| | |
|---|---|
| |  |

I'm never satisfied with the default settings for my new software and technology; I customize them in some way.

| Disagree Completely | Neither Agree Nor Disagree | Agree Completely |
|---|---|---|

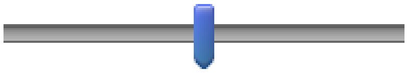
1 2 3 3 4 5 6 7 7 8 9



I want to get things right the first time, so before I decide how to take action, I gather as much information as I can.

Disagree Completely | Neither Agree Nor Disagree | Agree Completely

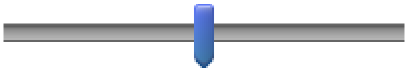
1 2 3 3 4 5 6 7 7 8 9



I always do extensive research and comparison shopping before making important purchases.

Disagree Completely | Neither Agree Nor Disagree | Completely Agree

1 2 3 3 4 5 6 7 7 8 9



When a decision needs to be made, it is important to me to gather relevant details before deciding, in order to be sure of the direction we are heading.

| Disagree Completely | Neither Agree Nor Disagree | Agree Completely |
|---|---|---|

1 2 3 3 4 5 6 7 7 8 9

I avoid "advanced" buttons or sections in new software and technology.

| Disagree Completely | Neither Agree Nor Disagree | Agree Completely |
|---|---|---|

1 2 3 3 4 5 6 7 7 8 9

I avoid activities that are dangerous or risky.

| Disagree Completely | Neither Agree Nor Disagree | Agree Completely |
|---|---|---|

1 2 3 3 4 5 6 7 7 8 9

Despite the risks, I use features in new software and technology that haven't been proven to work.

| Disagree Completely | Neither Agree Nor Disagree | Agree Completely |
|---|---|---|

1 2 3 3 4 5 6 7 7 8 9

End of Block: GenderMag facets

Start of Block: Demographics survey

Finally, please answer a few questions about yourself.

How do you currently describe yourself?

○ Female

○ Non-binary

○ Male

○ Other

○ Prefer not to say

What is your age?

- ○ Under 18
- ○ 18 - 24
- ○ 25 - 34
- ○ 35 - 44
- ○ 45 - 54
- ○ 55 - 64
- ○ 65 - 74
- ○ 75 - 84
- ○ 85 or older

In what college is your major or home department?

- ○ Agriculture and Life Sciences
- ○ Business
- ○ Design
- ○ Engineering
- ○ Human Sciences
- ○ Liberal Arts and Sciences

What is your current student status?

○ Freshman

○ Sophomore

○ Junior

○ Senior

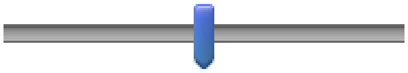
○ Graduate student

○ Non-student

How knowledgeable are you about managing probabilistic data (e.g., experience with statistics or machine learning)?

Not knowledgeable at all    Somewhat knowledgeable    Extremely knowledgeable

1    2    3    3    4    5



End of Block: Demographics survey