DECISION BIASES IN USER AGREEMENT WITH INTELLIGENT DECISION AIDS

By

Jacob Bennion Solomon

A DISSERTATION

Submitted
to Michigan State University
in partial fulfillment of the requirements
for the degree of

Media and Information Studies – Doctor of Philosophy

2015

ABSTRACT

DECISION BIASES IN USER AGREEMENT WITH INTELLIGENT DECISION AIDS

By

Jacob Bennion Solomon

Intelligent Decision Aids (IDAs) are emerging technologies used in areas such as medicine, finance, and e-commerce that leverage artificial intelligence, data mining, or related computational methods to provide recommendations to decision makers. An important goal for designers should be to help users identify and accept good recommendations and ignore poor recommendations. However, considerable research has found that IDA users frequently make poor decisions about which recommendations to follow.

I present findings from three studies that provide evidence of four distinct decision-making biases related to IDA-supported decision making. These biases are characterized by an increase in users' agreement with an IDA's recommendations that is unassociated with the recommendations themselves but associated with some other aspect of the design of the IDA or of the user.

In an experiment that manipulated the perceived customizability of an IDA that assisted users in predicting the outcomes of baseball games, I found that users who believed they had customized the IDA were more likely to follow both good and poor recommendations than other users who received identical recommendations from the IDA but did not customize its logic. This finding is evidence of a *customization bias.* Importantly, this study found that customization bias is not caused by users believing they have improved the algorithm by customizing it.

In a second experiment, subjects were encouraged to believe that the system had either high or low efficacy prior to seeing recommendations. This encouragement created an *expectations bias* in which subjects were more likely to follow both good and poor recom-

mendations when they had higher expectations of the IDA's efficacy than other subjects who had expected the IDA's algorithm to perform poorly.

In the third experiment, I assessed decision making by users of an IDA for recommending exercise activities. Subjects who used a customizable version of this IDA, where the recommendations depended on how users configured the IDA, were more likely to agree with the recommendations than users who received recommendations of similar quality but did not customize the IDA. This finding shows additional evidence of customization bias, demonstrating that it extends to IDAs where the customizability has real influence over the recommendations rather than merely perceived customization as in the first study. In this study I also found that when users believe that an IDA's internal logic is more clear and understandable, they are more likely to follow recommendations regardless of their quality. This finding suggests a *transparency bias*. There was a strong relationship between the quality of recommendations that subjects received and the quality of their decisions, indicating that when decision makers are supported by IDAs, the quality of recommendations is important to system success. However, subjects who performed the decision task unaided by an IDA performed as well as the IDA-supported subjects.

These findings show that when decision makers are aided by an IDA, the system affects the decision making process by requiring users to evaluate recommendations. IDA users may make biased evaluations due to characteristics of the interface and interaction design of the system as well as individual characteristics of the users. In the concluding chapter I discuss the implications of these findings for the design of IDAs and related socio-technical systems, as well as for future work on computer-supported decision making.

ACKNOWLEDGMENTS

There have been many contributions from many different people that have made it possible for me to pursue a Ph.D. and complete this dissertation, and I would like to acknowledge those contributions here.

I have been fortunate to have been supported in countless ways by my family, and I am grateful to them beyond what I can adequately describe. I am grateful to my wife Chalsea for giving me her full support for me to complete a Ph.D. I am also grateful to my parents who have been supportive and encouraging of me unconditionally throughout my life.

My advisor Rick Wash has spent countless hours providing feedback and guidance on my work and to me as a young researcher. He has given me with opportunities to work as a research assistant on his grant-funded research, which provided not only financial support but also invaluable experience at designing and executing research. That experience has been critical to my dissertation work and to my development as a researcher. Rick has encouraged my ideas, given me constructive critiques, and been patient and allowed me flexibility to pursue new ideas (even bad ones) that have helped me learn and grow.

My dissertation committee has made many important contributions to my dissertation and I am grateful for their input and feedback that have improved this work substantially. I first came up with several ideas that were the seeds for this dissertation from the discussion and feedback I received in Wei Peng's course in my first year as a Ph.D. student, and her feedback and suggestions continued to be helpful throughout the process of developing and refining my dissertation. Emilee Rader has provided important feedback on the design and execution of the research, and she has also given me important guidance for improving the way I communicate my ideas. Emilee has also generously provided funding for some of research in my dissertation. Josh Introne made several suggestions on both the design of the studies and the design of the system used in my research that proved to be critical for obtaining interpretable research results. I am also grateful for the guidance I received from former members of my committee Gary Hsieh and Joseph B. Walther, who contributed to the early stages of this research.

I have been a member of the BITLab at MSU, where I have been fortunate to have received helpful feedback on many occasions from the faculty and students there. I am particularly grateful to Chankyung Pak, who helped me figure out how to do some of the more complex statistical procedures that are used in my dissertation.

And I also wish to acknowledge the financial support I received for this work from the College of Communication Arts and Sciences in the form of a Dissertation Completion Fellowship and Summer Research Fellowship.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Advances in machine learning, artificial intelligence, and related computational techniques have been widely applied to help people from many different industries and circumstances make better decisions. A class of systems called Intelligent Decision Aids (IDAs) provide recommendations to decision makers by leveraging large quantities of data and applying artificial intelligence or sophisticated statistical models to generate recommendations.

In this dissertation, I argue that using IDAs to assist in decisions with high uncertainty alters the decision making process by requiring users to evaluate the recommendations that the system provides. I will show that this can be a challenging task for decision makers and that people are not always capable of identifying good and poor recommendations. I will also show that the while quality of recommendations is critical to agreement with recommendations, users can have biases that can influence users' agreement independently of the quality of recommendations. There is a customization bias, where users are inclined to agree with recommendations when they have participated in customizing the IDA's inner logic. There is also an expectations bias, where users are more inclined to agree with recommendations when they expect the system to perform well because they believe its process for producing recommendations is efficacious. A consistency bias occurs when users will be more likely to agree with recommendations that appear to be consistent with the way the IDA was configured. And there is a transparency bias, where users are more inclined to agree with recommendations when they feel they understand the logic that was used to produce them. These biases are evidence that the design of the user experience is influential in the decisions that users make and therefore in the effectiveness of an IDA as a socio-technical system.

These biases have provenance both in the design of systems and in the users of systems. By observing and reporting these biases, I make a contribution to theories of computer-supported decision making and offer knowledge that advances knowledge about how people make decisions when supported by intelligent systems. Additionally, an understanding of these biases makes a practical contribution to the design of IDAs. Because these biases show predictable behavior that is caused by the design of the system, system designers can navigate these biases or even employ these biases in an effort to engineer better decision making by users of their systems.

1.1 Intelligent Systems and Intelligent Decision Aids

Sophisticated computational technologies such as artificial intelligence and "big data" algorithms are increasingly being developed to add automation to knowledge work and to daily life (Carr, 2014). These technologies have been called *intelligent systems* (Guerlain, Brown, & Mastrangelo, 2000). Examples of intelligent systems are the algorithms that filter content on social media or other online content resources (Pariser, 2011), systems that collect, aggregate, and analyze detailed personal information about health and behavior to individuals in what are often called systems for the "quantified self" (Choe, Lee, Lee, Pratt, & Kientz, 2014), search engines that catalog the web and use algorithms to match queries to information within the catalog, recommender systems that suggest products to buy or movies to see, or algorithms that control (under human supervision) unmanned vehicles in military operations (Clare, Cummings, How, Whitten, & Toupet, 2012).

A feature of intelligent systems is that despite sophisticated automation, they still necessitate some human interaction or cooperation (Guerlain et al., 2000). For this reason, intelligent systems are socio-technical systems. Established principles of human-centered design argue for making system functions visible and controllable (Norman, 1990a), yet the computational methods behind intelligent systems are not always amenable to visibility or controllability because of their complexity. This has created a considerable challenge for designers of intelligent systems, who must design the interfaces and interactions between users and the algorithms contained within the system. As a result of this difficulty in designing these interfaces with intelligent systems, many systems suffer from poor usability, utility, or have other practical challenges. "Filter bubbles", for example, result from algorithms that filter online content in such a way that the diversity of content that any one person sees is slowly reduced, and this lack of diversity can go unnoticed by users (Pariser, 2011). Or as will be discussed in chapter 2, intelligent systems used in medicine, aviation, and other domains can lead users to make poor decisions because they are ill-suited to enabling users to properly calibrate their trust in the systems.

Norman (1990b) argued 25 years ago, in response to concerns regarding the unfulfilled potential of automation technologies, that the unfulfilled potential of automation is the result of inadequate interfaces and interaction design between humans and automation. I echo this argument today regarding intelligent systems. Intelligent systems cannot be successful unless the interfaces that afford human interaction and collaboration with them are as powerful and sophisticated as the computational methods that make them "intelligent." I argue for a human-centered design approach to intelligent systems that focuses on developing an understanding of users, designing technical affordances within this system, and evaluating the relationship between the nature of the users and the technical design of the system on the system's outcomes. In this dissertation, I present research that examines individual and aggregated differences in how people use intelligent systems to make decisions, tests different designs of intelligent systems that assist in decision making, and evaluates how the design of the system and the nature of its users combine to determine decision making outcomes.

In this dissertation I have focused on the application of intelligent systems to decision aids. The studies in this work evaluate the users and design of a class of technology I refer to as *intelligent decision aids*. I define intelligent decision aids as computational technologies that:

- Provide a recommendation or set of recommendations about specific actions or items that may be chosen by a decision maker.

- Generate these recommendation by means of artificial intelligence, statistical or mathematical modeling, or similarly complex computational methods that cannot be clearly or efficiently represented to decision makers in entirety.

- Provide a user interface with which a decision maker interacts in order to access these recommendations.

- Refrain from actually selecting or executing an action or decision without the approval of the user.

Some examples of systems that are included in this definition of an intelligent decision aid are:

- A clinical decision support system such as DxPlain (Barnett, Cimino, Hupp, & Hoffer, 1987) where a clinician provides information about a patient's case, then clicks a button to receive a set of potential diagnoses that should be most seriously considered.

- A system that recommends movies that users may enjoy watching by estimating their interests in certain types of movies based on data collected about the user.

- A system that recommends stocks that an investor may wish to purchase.

- A system that alerts a luggage screener of the potential for a hazardous item using image recognition software.

Two types of intelligent decision aids have emerged that have been investigated separately by distinct research communities. *Intelligent Decision Support Systems* have been the topic of research in information systems literature. One of the primary applications of this line of research has been targeted towards developing intelligent clinical decision support systems (Berner, 2007). Other applications within information systems literature have been in finance and operations planning. *Recommender Systems* have been the topic of considerable research in human-computer interaction and computer science. Recommender systems are frequently designed for e-commerce to help customers find items or services that match their preferences. Many recommender systems use a collaborative filtering (Su & Khoshgoftaar, 2009) approach in which recommendations are made to decision makers by finding items popular among people similar to the decision maker.

User-centered research on both intelligent decision support systems and recommender systems reveals considerable overlap in the socio-technical issues that must be considered in the design of these systems. Issues of trust (Wang & Benbasat, 2008; O'Donovan & Smyth, 2005; Massa & Avesani, 2007; Muir, 1987; Sanchez, Fisk, & Rogers, 2004) transparency (Sinha & Swearingen, 2002; Cramer et al., 2008; Herlocker, Konstan, & Riedl, 2000) and usability (Li et al., 2012; Herlocker, Konstan, Terveen, & Riedl, 2004) pervade the research from both research communities.

I have focused this dissertation on IDAs, rather than other types of intelligent systems or on intelligent systems more broadly, for two primary reasons. First, they are an important subclass of intelligent systems that are being increasingly adopted to assist in medicine, law, finance, and other professions (Carr, 2014) where decision makers make difficult decisions that affect lives. By generating new knowledge that can be used to improve IDAs, this research can make an important contribution in the real world. Second, IDAs have several properties that make them ideal for human-centered socio-technical research as described above. They can be used to assist in very explicit decisions for which adherence to the system can be clearly and objectively evaluated. Other types of intelligent systems may have a multiplicity of intended purposes or of different behaviors that are of interest. Filtering algorithms on social media, for example, may have purposes to create an engaging site but also to sell advertisements, and there may be many different outcomes or behaviors of interest like content consumption, content creation, or content sharing. IDAs can, at least in a lab setting, be plausibly presented to have a singular purpose and a singular behavior of interest with clear and objectively defined criteria for evaluation. These properties make IDAs highly useful for quantitative-based research that seeks to understand users and test intelligent system designs.

However, in spite of this focus on IDAs, there is an important reason that the results of these studies might be relevant to intelligent systems more broadly. The issues that I have assessed in this dissertation are largely functions of the design of the interface between users and the sophisticated computational methods embedded in the system. The customizability of an algorithm, the consistency between algorithm input and output, the transparency of the algorithm, or the system's self-evaluation of its own efficacy are aspects of the design that, based on the results of the studies described in the following chapters, can affect users' behavior when interacting with intelligent systems. Although the specific biases and behaviors I observe in this dissertation may manifest themselves differently in non-IDA intelligent systems, they are worthy of consideration in human-centered design and research on those systems as well and this dissertation provides a basis for extensions of this IDA-focused research to other types of intelligent systems.

1.2 Agreement with Recommendations

Consider a doctor who uses a computerized system that suggests treatment options for patients. The doctor may add relevant information about the patient and the diagnosis, and the system will return a list of suggested treatments, and may note which treatment or treatments it believes are most likely to be successful. The doctor must then examine this list and consider these options, as well as consider options that he or she is aware of but that have not been recommended by the system. Imagine if the top recommendation is an option that the doctor would not have considered or with which the doctor has little experience and would be reluctant to prescribe. And also imagine if this top recommendation is the treatment that would truly be most beneficial to the patient. If the doctor decides to follow this recommendation and prescribes this treatment, the system has been effective at improving decision making. However, if the doctor does not follow this recommendation and instead chooses a less beneficial treatment for the patient, the system has failed to improve decision making despite achieving a technical success at finding the best possible treatment for the patient. A reverse situation is also plausible, where the system recommends a less than ideal treatment but the doctor ignores that recommendation and chooses another better treatment that has not been suggested by the IDA. In this case, the system at worst has caused no harm and may even be considered a success if the process of using the system contributed in some way to the doctor making the correct decision, even though its recommendation wasn't followed. But if the doctor chose to follow the poor recommendation even though it is not what he or she would have otherwise chosen, than the system will have actually caused harm.

This scenario illustrates why it is important to understand what causes users of IDAs to follow or ignore recommendations. IDAs can only be effective at improving decisions if they a) provide recommendations that are better than what decision makers would otherwise choose, and b) persuade decision makers to follow those good recommendations. If both of those conditions are not met, an IDA may be ineffective or even harmful. Designers must find ways to improve both the quality and the acceptability of the recommendations that the system produces simultaneously, and this can be a serious challenge. It is further complicated by the fact that if IDAs are not successful or are inconsistent at creating high quality recommendations, than any efforts to promote agreement with recommendations may actually be counterproductive because they will lead users to make bad decisions.

IDAs, in spite of their efforts to provide critical information to decision makers, may create new uncertainty for decision makers if they do not fully understand how the system has produced its recommendations. The doctor in the example above may choose to ignore the good recommendation because he or she does not understand why it was suggested, instead opting for a lesser option because it is better understood. IDAs have been shown to frequently suffer from a lack of *transparency* in that users do not fully understand how they work or why specific recommendations were provided (Sinha & Swearingen, 2002). However, there has been considerable effort in IDA research and design communities to design more transparent systems, primarily by providing users with clear explanations for recommendations (Lim, Dey, & Avrahami, 2009; Ehrlich et al., 2011; Tintarev & Masthoff, 2011). Nevertheless, transparency is not easily achieved in IDA designs (Herlocker et al., 2000).

Research on transparency in IDAs has often not made a distinction between *understanding* how a system works and *preferring* how it works. I argue that users can loosely be categorized as those who do not understand how a system works, those who understand how it works and believe the process is effective, or those who understand and do not think the process is effective. In this dissertation I will show that these different types of users will often make different decisions in regard to agreement with IDA recommendations, even when the recommendations themselves are no different. I will show that when users understand how a system works and feel that the system's process has good efficacy, they will be more likely to follow both good and poor recommendations than others who have the same level of understanding but feel the system has low efficacy. This is important because it provides a target for system designers. As long as the system can produce good recommendations, designers can encourage strong agreement by taking measures to increase users' expectations about the efficacy of the system's recommendation logic so that they are likely to follow the good recommendations. At the same time, systems need to develop methods to help users understand when recommendations may be unreliable.

One way that designers may be able to simultaneously improve recommendations and users' beliefs about its efficacy is by designing customizable systems that require the decision maker to provide specific input to the system's logic or algorithm. This input may be used to provide local expertise or information to help the IDA provide a recommendation that best suits the specific circumstances of the decision. And by allowing users control over the algorithm, customization can allow users to give it a configuration that they believe has high efficacy.

I have previously found, however, that customization leads users to more agreement with recommendations in their decisions (Solomon, 2014). This is a decision making bias that I call *customization bias*. In chapter 3, I will show that users are biased by perceived customization even when they do not believe it has led to any greater efficacy of the IDA. This has an important implication for IDA research and design, as it is an example of how the process and experience of using an IDA can impact the decision making in ways that are unrelated to the quality of recommendations the system produces. In chapter 5 I will show that this bias is observable under a state of true customization where users' influence over the recommendations more visible than in the study in chapter 3. This additional evidence provides some external validity to the finding of customization bias.

When users have an understanding of the configuration of a system's logic, whether because they have customized that configuration themselves or have simply been made aware of it in the design of the IDA, there is an opportunity for them to evaluate whether the recommendations are consistent with the configuration. For example, in an IDA for recommending stocks to purchase, if the user customizes the system to focus on energy-related stocks, and the recommendations do not suggest any energy-related stocks, the user might think that the system has malfunctioned or is simply a poor system, and then might choose to ignore the recommendations. But the system may have a good reason for not suggesting energy-related stocks in spite of the configuration (e.g. its algorithm thinks all energy stocks are poor investments at that moment), and therefore ignoring this recommendation would be a bad choice by the user. In this scenario, the lack of consistency between the configuration and the recommendations is causing a bias, because the user is ignoring the recommendation on the basis of a lack of consistency even though the system is making a good recommendation that should be followed. I observed this bias in the study presented in chapter 3, and I argue that it is another example of the design of the interaction between user and IDA influencing decisions independently of the quality of recommendations.

CHAPTER 2

BACKGROUND

Designing intelligent decision aids presents a number of challenges due to the socio-technical nature of computer-supported decision making. Computing technologies can be powerful and capable of providing valuable insight to decision makers. However, decision makers are human and therefore diverse in their capabilities and characteristics, prone to using heuristics and having biases in decision making, and may have different decision making goals than system designers. For this reason, any design that keeps humans in the loop of decision making must account for these human factors.

In this section, I will discuss existing research on the design of intelligent decision aids. I will describe some common technical approaches to providing computerized decision support. I will also give an overview of research on human factors in IDAs, with a focus on issues of trust in automation, transparency of systems' inner logic, user control and the division of labor between user and system.

From this overview, I will make a theoretical argument that a completely top-down approach to IDA design where system engineers make all determinations about how the system produces recommendations may be inadequate for maximizing the potential of IDAs. I will provide evidence from the literature that allowing end-users to have some control over the design of an IDA's inner logic, a process called *customization*, can serve to make IDA designs less vulnerable to known human factors problems with IDAs. I will also discuss the theoretical basis for new human factors problems that may arise from an IDA design that affords end-user customization of its recommendation-producing process.

2.1 Computational Techniques for Decision Support

There is a wide variety of computational techniques that have been developed and deployed within IDAs to provide support to decision makers. For the most part, these computational techniques are used to process and analyze data that is available to the system in order to produce recommendations for the user about their decision. In this section I will briefly describe a few of the most common computational techniques that are used in IDAs, including a discussion of the strengths and weaknesses of these approaches.

2.1.1 Collaborative Filtering

Many IDAs make use of a collaborative filtering (Su & Khoshgoftaar, 2009) approach to generating recommendations. In collaborative filtering, users generally provide ratings or other information about their preference for items in a system's catalog. Collaborative filtering techniques typically create a matrix or set of matrices that represent users or items within the system. A User-Item matrix represents a rating or other form of valuation for every combination of user and item within the system, although many of the cells of this matrix may be empty because users typically only rate a small portion of all possible items. A User-User (as well as an Item-Item matrix) stores a 'similarity' value calculated using a distance metric such as cosine distance or Euclidean distance between each pair of users or items. For example, consider two users who have each rated a set of movies. The similarity between these two users can be calculated by the cosine distance between each users' vector of rated movies. If the two users mostly agree on their ratings of movies, the cosine distance will be small and the similarity large. This similarity value is then stored in the User-User matrix.

From these matrices or similar data representations, recommendations can be generated using one of a variety of computational techniques. A common approach to producing recommendations from these matrices is k-nearest neighbors (Ekstrand, Riedl, & Konstan, 2011).

In this approach, the system finds a small group (of size k) of users from the system that are most similar to the current user in terms of their explicit valuations of items, then recommends to the current user items that are most popular within their ‘neighborhood’. Another approach is matrix factorization. Matrix factorization techniques (Koren, Bell, & Volinsky, 2009) seek to find latent factors within the matrices that indicate some underlying concept or variable among a set of users’ preferences for items. For example, matrix factorization techniques may reveal that movies may fall on some scale between ‘serious’ and ‘escapist’ and that users’ preferences for the different extremes of this scale is manifested in their ratings of movies (Koren et al., 2009). Thus, by identifying such latent factors, a system can determine both where on the scale a user’s preference lies and where on the scale each item lies, and recommend items that are a close match to the inferred preference.

A related approach to collaborative filtering is a network-based approach to generating recommendations. In this approach, a social network is constructed from the set of users and items in the system, and various attributes of the network are used to make recommendations such as the centrality of nodes. For example, Qin et al. (2010) built a recommender system for YouTube videos by building a network of videos as nodes and connecting any two videos if they share at least one common user who has commented on the video. They then make recommendations by using network properties to produce expected utilities for videos and users and recommending to users the videos with the highest expected utilities.

A social networking approach to recommendations can be particularly useful when social connections themselves are the items being recommended. Social networking sites can recommend other users to each other by evaluating the number of shared nodes within the network of users, as well as other features of the network structure, and use this information to suggest new connections (Roth et al., 2010).

One of the greatest strengths of collaborative filtering is that it is “content-independent” (Park & Chu, 2009). This means that the system does not need to have very much explicit information about the items it recommends. For example, movie recommendations can be made using collaborative filtering with just the title of each movie and a set of ratings from users, since the recommendations are entirely determined by the ratings and not anything in the content of the items. This content-independence can also be beneficial in that it can lead to more "serendipitous" recommendations. Collaborative filtering can find items that are similar to items a user likes but in ways the user has not previously considered (Park & Chu, 2009).

The content-independence, however, comes at a cost for collaborative filtering of being highly user dependent. This user-dependence is known as the "cold-start problem" (Lam, Vu, Le, & Duong, 2008). Collaborative filtering requires data about users, particularly ratings of items, in order to be useful. It requires a relatively large existing user base in order to find meaningful clusters of users that share preferences. And for any given user, it needs ratings or other information about that user in order to find similar users from which to make recommendations. New systems do not have the data they need to be useful to users, which makes it difficult to build the critical mass of users required to ever become useful.

2.1.2 Content-Based Recommendations

Content-based recommender systems produce recommendations by using known and explicit attributes of items. These systems maintain an Item-Attribute matrix where all items have been evaluated on the attributes, and the systems then elicit preferences from users for attributes and recommend items that are similar to users' stated preferences for attributes (Leino, 2014).

One of the primary advantages of content-based recommenders is that they suffer less from the cold-start problem that plagues collaborative filtering (Schein, Popescul, Ungar, & Pennock, 2002). Content-based recommenders do not require a critical mass of existing users to be able to make recommendations, meaning that new systems can be more useful from inception. However, content-based recommenders do still need data about a given user in order to make personalized recommendations, meaning the cold-start problem is not entirely eliminated.

A problem with content-based recommenders is that curating the content features can be difficult and costly. Pandora.com, for example, uses a content-based recommender (Glaser, Westergren, Stearns, & Kraft, 2006) that requires every song in their database to be rated by a musical expert on over 400 musical features[1]. Developing and curating the data in such a way may not be practical for all types of recommendations.

2.1.3 Artificial Neural Networks

Other innovations in artificial intelligence have been adopted in some contexts, most notably in medicine. Artificial Neural Networks (ANN's) are a form of artificial intelligence that tries to find patterns among large and complex data (Hill, Marquez, O'Connor, & Remus, 1994). ANN's are a replication of biological neurological processes and have been found to be powerful in many analytic tasks, particularly in contexts where traditional statistical approaches based on regression are problematic (Hill et al., 1994). ANN's have been found to be particularly useful for IDA that support clinical decision making (Berner, 2007). A known shortcoming of ANN's is that unlike many other techniques, ANN's do not provide clear reasoning about how they have come to conclusions (Berner, 2007).

2.1.4 Genetic Algorithms

Genetic algorithms are another technique from artificial intelligence that have been deployed with some success in IDAs (Berner, 2007). Genetic algorithms simulate an evolutionary process based on the notion of survival of the fittest. Attributes of recommendable items that are distributed among a dataset are randomly combined and the resulting combination is evaluated according to an established criteria or "fitness function." Combinations that yield the best evaluations are kept, while weaker combinations are dropped, and new combinations

[1] https://www.pandora.com/about/mgp

formed from the remaining set. This process is continued until performance stops improving, at which point a solution is apparent that can be used as a basis for recommendations. Genetic algorithms have the same shortcoming as ANN's in that they struggle to provide clear reasoning for the recommendations that are produced (Berner, 2007).

2.2 Applications of IDA

One of the challenges in researching and developing effective IDAs is the tremendous variety in the decision contexts for which they are implemented. The applications of IDA technologies range from e-commerce to finance to medicine to journalism. E-commerce websites such as Amazon.com have adopted recommender systems to suggest products to customers and help them choose items from an enormous catalog (Linden, Smith, & York, 2003). Recommender systems on large sites such as Amazon may make recommendations about specific items within a class of alternatives (e.g. which book to read) as well as suggesting different classes of items for its customers to consider (e.g. whether to look for a book or a camera). Other e-commerce recommender systems focus on a more specific decision context, such as putting together an outfit or wardrobe (Tu & Dong, 2010).

Some highly successful IDA implementations make recommendations about media consumption. Netflix has sought to improve the user experience of its service by researching and deploying a recommender system [2]. MovieLens (Miller, Albert, Lam, Konstan, & Riedl, 2003) is a similar movie recommendation service which has been developed not only to make recommendations to decision makers but as a testing ground for experimental approaches to recommender system design. Online music services such as Pandora [3] and LastFM [4] have used recommender systems to engage users by providing personalized but serendipitous music recommendations. Social networking sites such as as Facebook, Twitter, or LinkedIn can

[2] http://www.netflixprize.com/index

[3] http://www.pandora.com

[4] http://www.last.fm

help people make decisions about who people to interact or communicate with. LinkedIn, for example, makes recommendations to users about who they might connect with for professional development and networking, as well jobs or other career opportunities (Skeels & Grudin, 2009). An important application of IDAs is in recommending news articles or other types of web-content for users to consume. Large web-portals such as Google and Yahoo use techniques taken from recommender systems to produce personalized news portals that recommend content that is expected to be of interest to users (Liu, Dolan, & Pedersen, 2010).

2.3 IDA Effectiveness

IDAs have been found to be effective at reducing the time and effort required to make decisions, (Hostler, Yoon, & Guimaraes, 2005; Amento, Terveen, Hill, Hix, & Schulman, 2003; Chen & Pu, 2009; Xiao & Benbasat, 2007). IDAs can automate the acquisition and analysis of information, reducing the number of alternatives that a decision maker must consider (Häubl & Trifts, 2000) and thus reducing effort. This is an important benefit that IDAs provide to decision makers that justifies their development and implementation.

However, some researchers argue that IDAs should benefit not only decision processes such as the efficiency of making decisions, but also the decision outcomes themselves (Knijnenburg, Willemsen, & Kobsa, 2011; Xiao & Benbasat, 2007). Researchers have had difficulty in establishing the efficacy of IDAs when used in practice. For example, in clinical settings, Bright et al. (2012) conducted a systematic review of clinical trials of IDAs. They found that only 20% of randomized trials of clinical IDAs even evaluated decision outcomes, whereas most were focused on assessing decision process measures and the economic justification of IDAs. Among the trials that did assess decision outcomes, they found only limited evidence that IDAs were beneficial. They noted however that the lack of clear evidence may be due to the tremendous difficulty that exists in executing clinical trials for IDAs that evaluate decision outcomes. There may be a serious selection bias in that systems and circumstances for which clinical trials can be easily executed are also those for which the particular systems are ineffective. Bright et al. concluded not that IDAs were ineffective but that there is insufficient evidence to draw a conclusion.

In e-commerce, there are no comparable reviews or even any studies to my knowledge that assess the broad impact of IDAs on decision quality or decision outcomes that are based on data outside of a laboratory setting. This may be due to the horizontal differentiation of most decisions in e-commerce in which users may have widely varying criteria for what they prefer, making it difficult to develop measures for decision quality that are objective. However, there has been considerable work to evaluate e-commerce IDAs in lab settings where decision quality can be objectively defined and measured (Xiao & Benbasat, 2007). Some examples of such measures include a match between the attributes of a chosen item and subjects' preferences for those attributes (Pereira, 2001) or the frequency with which subjects change their decision later when given a cost-free opportunity to change (Häubl & Trifts, 2000). Xiao and Benbesat (2007) conducted a systematic review of IDA research in e-commerce that used such measures, and noted that there were mixed findings, with some studies finding the IDAs helped decision making while others found no effect or even that IDAs harmed decisions. However, they found that variations in the design of the IDAs used in these studies, such as the way preferences were elicited and the way recommendations presented, had an impact on decision quality. This suggests that the design of the IDA is an important factor for decision making. It also suggests that there is a strong need for socio-technical theory of IDA design to help designers engineer good decision making.

2.4 Agreement with Recommendations

2.4.1 Trust

Technologies that automate decisions, actions, information processing, or other functions within a human-machine interaction place a burden on users to develop trust in the system. Trust in technologies has been a key area of research on IDAs and in automation technologies more generally. Researchers in these areas have sought to understand how users develop trust in systems, how users lose trust in systems, and how users' trust in an automated technology influences their behaviors or decisions.

Lee and Moray (1992) argue that as automation is increased in a system, the control that users exert over a system and its output shifts from an active control to "supervisory control." This role of a supervisor rather than a direct controller demands that users place an increased amount of trust in the system.

In seeking to understand how users develop and maintain trust in automated technologies, researchers have disagreed over the appropriate way to define trust within the context of a human-machine interaction. One approach that has frequently been taken is to borrow definitions of trust from research on interpersonal relationships (Hoffman, Johnson, Bradshaw, & Underbrink, 2013; Muir, 1987; Madhavan & Wiegmann, 2007; Wang & Benbasat, 2008). Interpersonal trust has been defined as "a willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that party?" (Mayer, Davis, & Schoorman, 1995, p. 712). In this sense, trust in automation can be viewed as a user's willingness to be vulnerable to a subordinate machine's actions in the same way that supervisors in a hierarchical interpersonal relationship are vulnerable to the action of those they supervise. The "Computers-Are-Social-Actors" paradigm (Nass & Moon, 2000), which demonstrates a human propensity to form social relationships with interactive machines, has often been used as a justification for treating and measuring trust in technology using an interpersonal conceptualization of trust (Madhavan & Wiegmann, 2007).

However, much work has found that people trust technology in different ways than they trust other people. Lee and See (2004) claim that while interpersonal trust is symmetrical in that both parties have a need to develop trust in one another through repeated interaction, trust in automation is generally asymmetrical. Users may need to develop beliefs and attitudes that allow them to be willingly vulnerable, but automated systems generally do not need to develop trust in their users.

Other work has found that people's level of trust in systems often differs from their level of trust in a human performing the same role with the same reliability. For example, Dzindolet and colleagues (Dzindolet, Pierce, Beck, & Dawe, 2002; Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003), in experiments that control the reliability and utility of both an automated decision aid and a human decision aid, found that people have an initial tendency to trust and rely on the automated aid more than they trust the human. However, they also found that in cases where the aid made an obvious error, users' reliance on the automated aid plummeted considerably to a level far below both the reliability of the automated aid and below the reliance observed in a condition using another human as the decision aid in place of the automated system. This sensitivity presents an important challenge for designing IDAs. Designers must find ways to appropriately set expectations of system quality. Expectations can be a powerful influence on behavior in socio-technical systems (Wash, 2013), and IDA users may have difficulty forming accurate expectations of the system.

Trust, particularly in the context of automation use, is typically defined as an attitude or a belief (J. D. Lee & See, 2004). This distinguishes trust from a related concept known as *reliance*, which is a behavior in which users rely on or defer action to an automated system. Lee and See (2004) argue that while trust influences reliance, the two concepts are not identical. They argue that trust is explicitly an attitude and should be treated and measured as such in research on trust and automation, whereas reliance is the behavior of deferring actions or decisions to an automated system.

An important question that has been explored in the research on trust in automation relates to how trust is developed. One hypothesis is that trust is the result of reliability over time. As a system demonstrates that it is reliable, users grow to trust it more and become more reliant on the automation. Lee and Moray (1992) suggest that trust in automation is enhanced when users understand the process that the automation uses to produce its output.

A number of individual differences between users can also impact trust in automation. Sanchez et al. (2004) found that older adults were more sensitive than younger adults to declines in the reliability of an automated decision support system for a driving task, losing trust in the system more quickly as the system became less reliable. Merritt and colleagues (Merritt & Ilgen, 2008; Merritt, Heimbaugh, LaChapell, & Lee, 2012) showed that some people have more of a general propensity to trust automation than others.

Some researchers (O'Donovan & Smyth, 2005; Massa & Avesani, 2007) have used the notion of trust as a basis for making recommendations. In a collaborative filtering system, similarity between users in terms of preferences can be replaced by a trust-metric, where users' degree of trust of other users is elicited or estimated in order to make recommendations.

Lee and See (2004) argue that an important goal of research on automation is to find ways to design for calibration between users' trust in a system and its reliability. If a system is highly reliable, than users should trust and follow its recommendations and by doing so will make optimal decisions. If a system is only moderately reliable, users' should be less trusting and carefully scrutinize its recommendations. As a design objective, IDA designers should find ways to help users properly calibrate their trust in an IDA in order to optimize decision making.

2.4.2 Automation Bias

Human factors engineering researchers have examined whether automated decision aids such as alert systems and screening tools help users make better decisions. Skitka and Mosier (Skitka, Mosier, & Burdick, 1999) have observed considerable evidence of *automation bias.*

Automation bias occurs when a decision maker fails to seek out evidence that contradicts recommendations provided by a decision aid, leading them to follow poor recommendations and make poor decisions (Manzey, Reichenbach, & Onnasch, 2012). Skitka and Mosier (Skitka et al., 1999) conducted an experiment on a flight simulator task where an automated alert system was 94% accurate. They found that when the system gave an incorrect alert, subjects followed the recommendation 65% of the time. They also found that subjects failed to take appropriate action when they were using an automated alert system and it incorrectly failed to alert them, as compared to a control group which did not have an automated alert system. Similar findings have been reported using other aviation tasks (Mosier, Skitka, Heers, & Burdick, 1998), luggage screening (Madhavan & Phillips, 2010), process control (Manzey, Reichenbach, & Onnasch, 2008), and mammography (Alberdi, Povyakalo, Strigini, & Ayton, 2004).

Automation bias has become a serious problem in clinical decision making. Goddard et al. (2012) conducted a meta analysis of studies on automation bias in clinical decision making. Their review included automation bias induced both by intelligent decision aids as well as other types of automated systems such as alerts. They found that when a decision aid gave incorrect recommendations, it increased the risk of a poor clinical decision by 26%. Coiera et al. (2006) have argued that automation bias has been one of the primary culprits for why intelligent systems have failed to meet expectations for improving care in clinical settings. They argue that while intelligent systems provide many benefits, automation bias and other related human factors problems are simultaneously created by the introduction of intelligent systems that may offset the gains.

Carr (2014) provides a number of examples of automation bias as a cause of devastating decision making in contexts such as transportation, finance, and law. He argues that an over-reliance on intelligent systems that automate knowledge and intellectual work may dull the analytical skills required for good decision making and creative problem solving. Part of this, he argues, is that such system may deprive people of the more enjoyable aspects of their work and make them indifferent to the decision making process by removing an emotional attachment to it.

Fortunately, there is evidence that automation bias is not an inevitable consequence of using intelligent systems, and that automation bias can be minimized through system design. Dzindolet et al. (2003) found that automation bias was reduced when the system explained its reasoning process. The transparency of a system may make it easier for users to scrutinize recommendations and appropriately calibrate their reliance and agreement with them. Minimizing the prominence of an information display (Berner, Maisiak, Heudebert, & Young Jr, 2003) can also help users avoid becoming too reliant on an automated decision aid. By avoiding information overload, users have more cognitive resources available to process information on their own and scrutinize recommendations made by an aid.

2.5 Transparency

One of the most important topics in IDA research and design is *transparency*. Transparency refers to the degree to which users understand why the system gave its recommendations, or understand the system logic for how they were generated. Transparency is often accomplished in IDA by giving users explanations for recommendations. For example, e-commerce sites like Amazon often explain recommendations by stating "People who purchased this item also purchased these items."

Transparency is important for IDAs for several reasons. Herlocker, Konstan, and Reidl (2000) have argued that most IDA are "black boxes" and that transparency is important so that users can handle errors in recommendations. Since recommender systems are rarely perfect, it is helpful for users to understand why a recommendation was given so they can determine for themselves whether a recommendation contains error. This approach allocates the function of recognizing errors to the user but places the demand of articulating or communicating how recommendations are generated to the system. Tintarev and Masthoff

(2012) have suggested that providing explanations in recommender systems serves to establish user trust in the system and to make the recommendations more persuasive. A lack of transparency in intelligent systems for home automation has been found to be one of the primary frustrations of users and an impediment to more widespread adoption (R. Yang & Newman, 2013).

In section 2.1 above, I outlined a set of technical approaches for the inner logic of IDAs. There is considerable variability across these technical approaches such that two different systems designed to assist the same type of decision may reasonably use one of many different technical designs.This suggests that from the perspective of users, it is unreasonable for designers to expect even people experienced with using IDAs to know or have an expectation that a system works in any particular way. Users should not be expected to think that the black box that generates recommendations in one system is the same as the black box of another system. Therefore, users' beliefs about the inner logic must primarily come from what they are told within the system interface, documentation, or training about the system.

Much of the work on transparency in IDAs or related systems has focused on evaluating different ways to design explanations into the interface of an IDA (Tintarev & Masthoff, 2011). Lim et al. (2009) have compared explanations that explain recommendations in terms of "why," "why not," "how to," or "what if." They found that for enhancing user understanding of the system, "why" and "why not" explanations worked best and that other forms of explanations did not help users better understand a system than providing no explanations. Other work in this area has explored using visualizations to explain IDA recommendations (Verbert, Parra, Brusilovsky, & Duval, 2013; Knijnenburg, Bostandjiev, O'Donovan, & Kobsa, 2012), providing detailed tutorials (Kulesza, Stumpf, Burnett, & Kwan, 2012), using and explaining simple formulas as the recommendation logic (Aksoy, Bloom, Lurie, & Cooil, 2006), describing the tradeoffs between items as a way to justify recommendations (Wang & Benbasat, 2007), and hierarchical explanations that allow users to "drill down" further and further down a decision tree until they are satisfied with the explanation (Kay & Kummerfeld, 2012).

Many studies of IDAs find that providing transparency leads to user satisfaction with the system and makes them show greater trust in recommendations (Herlocker et al., 2000; Tintarev & Masthoff, 2012, 2008; Sinha & Swearingen, 2002; Cramer et al., 2008). These studies show that users prefer transparent IDA interfaces and report that they are useful in the decision making process. However, a more pertinent question is whether transparency actually helps users make better decisions. Cummings argues that the "inability of the human to understand complex algorithms only exacerbates the tendency towards automation bias" (Cummings, 2004, p. 3). When users don't understand how a system produces its recommendations, they may find it harder to identify when the system has made an error and therefore may make poor decisions when the system provides poor recommendations.

Some research has shown though that transparency alone may not help decision making, and in some case may actually lead to decision errors. Ehrlich et al. (2011) found that in a DSS for helping IT admins detect security attacks, providing explanations led some users to have unwarranted confidence in the recommendations and subsequently make poor decisions when the system made poor recommendations. Tintarev and Masthoff (2012) found that in recommender systems for movies and for cameras, users who were given explanations for recommendations were actually more likely to change their mind about their decision later on, suggesting that the explanations prompted them to make a decision that they later regretted. These results have some important implications. One reason that users may make poor decisions when a system is transparent is that they may be convinced by the explanation for the recommendation more than the recommendation itself. In other words, users may develop a preference for how recommendations are generated that is the basis for their decisions more than an independent evaluation of recommendation quality. This would suggest that transparency, in contrast to Cumming's argument mentioned above, can increase automation bias if user preference for system logic supersedes a critical analysis of the recommendations by the user.

2.6 Customization

An emerging design approach to IDAs is to afford end-user customization of the system. This means that users are given some control over some aspects of the system. Users may be able to customize the visual layout or other aesthetic details of the interface, the data that are stored or used in the system, or even the algorithm or reasoning process can be controlled or influenced by the user.

Customization is a widespread approach to designing all types of interactive systems, including IDAs. Customization in IDA has often been conceived as allowing users control over the data that they input, and findings ways to elicit the best data (e.g. item ratings or user profile information) from users (McNee, Lam, Konstan, & Riedl, 2003). However, recently many experimental IDA have explored customization of an IDA's algorithm (Bostandjiev, O'Donovan, & Höllerer, 2012; Han, He, Jiang, & Yue, 2013; Schafer, Konstan, & Riedl, 2004; Bostandjiev, O'Donovan, & Höllerer, 2013; Parra, 2013). A shortcoming of the published research on these kinds of systems is that they have not been evaluated in terms of their effect on decision making. Rather, user-centered evaluations of usability or user satisfaction have been favored in most cases. Thus, it is not clear from the existing work on customizable IDA how they affect user decision making.

In this section, I will discuss the literature on customization in interactive systems and what is known about making human-machine systems customizable from a design and user-experience perspective. I will then discuss the literature specifically about customizable IDA and make an argument based on both theory and empirical findings that customization can be beneficial to IDA-supported decision making. However, I will also point to evidence that suggests new human factors design problems that may arise as a result of customizable IDA.

2.6.1 Customization of Interactive Systems

Much of the existing work on customization has looked at customization within interactive media such as web portals, newsfeed readers, and online communities with user-generated content. Sundar (Sundar, 2008; Sundar, Oh, Bellur, Jia, & Kim, 2012) has argued that interactive media such as these that afford users the ability to customize some aspect of the interface nurture a sense of agency in users, and that this sense of agency is a powerful predictor of user satisfaction when using interactive media. Sundar describes the sense of agency experienced by users of customizable media as a feeling of "self-as-source," meaning that users feel rewarded by simultaneously being both a consumer and a producer of the medium.

One purpose for customization in interactive media is *personalization.* Personalization means that content provided through the medium is personally relevant or engaging to the user. Personalization may be achieved by a system trying to estimate what content a user will prefer or find most relevant and then inserting that content into the medium. Blom (2000) describes a continuum between *personalization* and *customization*, where *personalization* is system-initiated tailoring of content whereas customization is user-initiated tailoring of content. Personalized systems gather and process information autonomously about users and perform tailoring of content often without the direction or even awareness of users. Customization provides an interface for users to explicitly influence this tailoring. For example, targeted advertising that tries to show users the most relevant ads based on information it collects through cookies about users browsing would be an example of personalization, as the tailoring of the content is entirely system-initiated. However, a system that gives users prompts and asks questions to explicitly allow users to tailor their ads would be a customizable system. This distinction highlights a difference in the process of how content becomes tailored to a user, but no distinction in the outcome of the process in which the content is highly relevant and engaging to a particular user.

An important question that follows from this distinction is whether or not differences in the process of providing tailored content, user-controlled vs. system controlled, has any impact on users' experiences with interactive media. Sundar et al. (2010) have explored this question, seeking to determine whether customization is purely a form of user-controlled personalization, or whether it has other effects on users. They compared a customizable version of an RSS news reader with one that was personalized to provide relevant content using a collaborative filtering technique. They found that "power users" were able to generate content that they most preferred when using the customizable interface rather than the personalized interface. However, non-power users preferred the content in the personalized interface. The authors performed a follow-up study to determine whether a perception of control over the interface can explain the preference that power users had for the content, but this study found that perceived control did not explain the results. This result suggests that the benefit of customization is that it allows skilled users to maximize a system's output to match their preferences better than a system-maintained algorithm might. When users are given control of a medium and they have the skill to use it properly, they can tailor its output to maximize its quality.

Other work, however, has found that the sense of agency afforded by customization provides additional benefits to user satisfaction beyond simply enabling a system to provide high-quality content to be consumed. Marathe and Sundar (2011) argue that customization enables an interface to be predictable to the user. They also argue that customization fosters not only a sense of control or agency, but also a sense of identity for users within the medium. Customization of a medium allows users to express who they are through the medium, and this affordance of self-expression is beneficial to user satisfaction and user experience with an interactive system. In an empirical test of this hypothesis, they found that using a customizable web portal led users to both have a greater sense of control and to feel that the system was a more precise expression of their identity.

The theme resulting from this work on customization in interactive media is that customization enables skilled users to tailor content or output to match their preferences, and that this affordance also leads to a generally more satisfactory user experience by fostering a sense of control, identity, and predictability when using interactive systems.

2.6.2 Customization in IDAs

Is customization a good design for intelligent decision aids? How does it impact the decision outcomes, decision processes, and user experience of IDA? In this section, I will detail a theoretical argument for the benefits of customization in IDAs. I will then examine empirical evaluations of customizable IDA and discuss how they support this argument. I will also review some research that demonstrates potentially negative consequences of customization in IDAs.

2.6.2.1 Theoretical Basis for Customizable IDAs

In any human-machine system, there is some division of labor between the user and the system, where users fulfill some functions or sub-tasks necessary to the primary task and the system performs other functions. Designers of IDAs and other human-machine systems must determine which functions are to be performed by the machine and which functions should be left to users, and making this determination can be a difficult design challenge. In the field of human factors engineering, this division of labor is the *function allocation*, and it has been an important area of research for over 60 years (de Winter & Dodou, 2014).

As technologies have advanced in capabilities, this question has generally become even more complicated because there is increasingly overlap between the functions that can feasibly be performed by either humans or machines. Many theories and frameworks have been developed to assist designers in allocating the functions of a system. The most prominent of these frameworks is Fitts' list (Fitts, 1951). This list, sometimes known as the MABA-MABA list ("Men Are Better At-Machines Are Better At") makes declarations about types of functions that each entity is likely to excel at in comparison to the other entity. As examples, this list suggests that humans are better at perceiving patterns, improvising or using flexible procedures, and reasoning inductively. It suggests that machines are better at following procedures with precision and performing multiple functions simultaneously.

Since its original publication, Fitts list has frequently been adapted and modified to apply to new types of machines and technologies. However, its basic premise of creating heuristics that distinguish the relative strengths of humans and machines has been well preserved in engineering and design (de Winter & Dodou, 2014).

IDA designers face many function allocation decisions. Functions such as inputing and validating data, conducting analysis or filtering output may conceivably be completed by either the user or the system. For example, designers of an IDA may allocate to the user function of acquiring or curating relevant information, and allocate to the machine the function of calculating probabilities or retrieving a document from a database. Or the designers may allocate the function of inputing data to the system through an automated web crawler, and the function of analyzing the relevance of the data to the user or users. Designers may need to balance factors of utility, reliability, and usability in making function allocation choices.

Parasuraman, Sheridan, and Wickens (2000) have proposed a classification framework to determine an IDA's *level of automation*. The level of automation can thought of as the proportion of functions that have been allocated to the system rather then the user. This framework proposes a scale between completely autonomous systems that decide and execute all decisions and completely unassisted decision making. This framework is useful for describing systems where functions are uniquely allocated to either the user or the system in a "divide-and-conquer" design.

Figure **??** illustrates Parasuraman et al.'s framework, including some examples of hypothetical system designs and where they fit into the framework. Parasuraman et al. explain that the first two stages, information acquisition and information analysis, can be thought of

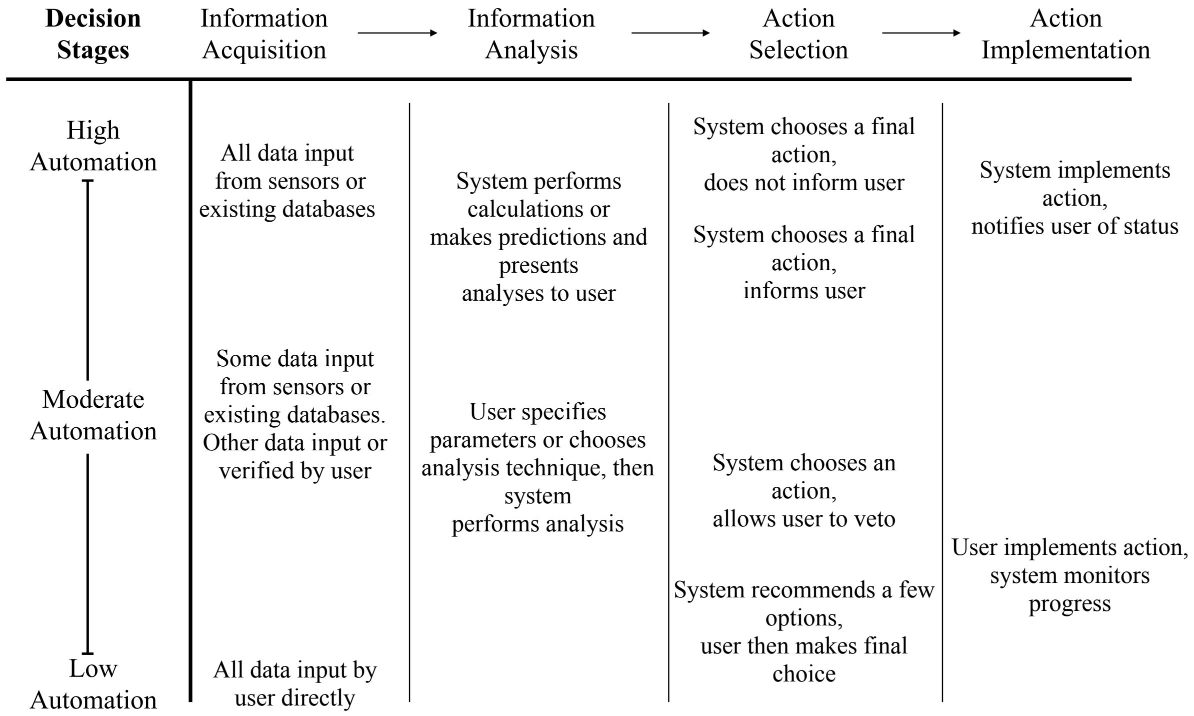

Decision Stages
Information Acquisition
Information Analysis
Action Selection
Action Implementation
High Automation
Moderate Automation
Low Automation
All data input from sensors or existing databases
Some data input from sensors or existing databases. Other data input or verified by user
All data input by user directly
System performs calculations or makes predictions and presents analyses to user
User specifies parameters or chooses analysis technique, then system performs analysis
System chooses a final action, does not inform user
System chooses a final action, informs user
System chooses an action, allows user to veto
System recommends a few options, user then makes final choice
System implements action, notifies user of status
User implements action, system monitors progress

Figure 2.1: Some hypothetical system designs and their placement within Parasuraman et al.'s of decision making stages and levels of automation.

as the "input" to the system whereas the latter stages are its output. Customizable IDAs, as discussed in this dissertation, refers to user customization of system inputs, and therefore fits into the first two stages of decision making. By allowing users to influence inputs, customizable IDAs would best be described as having a moderate to low level of automation in these input stages. Complete manual control, the lowest level of automation in the framework, would fall outside the definition of an intelligent decision aid however.

Research on automation at the input stages has generally found that a higher level of automation leads to improved task performance as long as the automation performs reliably. But when the automation fails, this leads to worse performance than if the input stages had been allocated primarily to the user (Onnasch, Wickens, Li, & Manzey, 2013; Schuster, Jentsch, Fincannon, & Ososky, 2013). Onnasch et al. (2013) attributed this to a loss of *situational awareness*. This means that users become less aware of all the conditions and information that should impact their decision and this makes it harder for them to recognize a failure of the system.

A customizable decision aid may be able to take advantage of the performance benefits of automation while avoiding the loss of situational awareness that leads to poor decision making. By making a system personalized to the specific decision context, customization may make the system more reliable, i.e. provide better recommendations. But the involvement required of the user to customize the system may demand that users maintain situational awareness and may be better able to recognize when the system has failed or provided poor recommendations. In other words, customization may be able to create a middle ground both the benefits of manual control and automation are realized.

In support of this effort to find an ideal compromise between allocating input functions to the user or to the system, there has been an increasing call among automation researchers to explore a more collaborative style of human-machine interaction (Cummings & Bruni, 2009) than the one prescribed by tradition function allocation theories, including Parasuraman et al.'s framework. In human-machine collaboration, functions are not necessarily divided between user and system and the level of automation is not necessarily a measure of "how much" is done by either entity. Instead, the interaction is designed to allow functions and roles to be shared to varying degrees and communication encouraged and simplified to allow the user and system to collaboratively arrive at an optimal solution.

Cummings and Bruni (Cummings & Bruni, 2009) have built on the Parasuraman et al.'s levels of automation framework to make it more inclusive of the concept of human-machine collaboration. They identify three important roles in the decision process, and each of these roles can be allocated either in full or in part to either the human user or the machine. The *generator* produces a set of decision alternatives or recommendations. The *decider* makes a final choice from among the generated alternatives. And the *moderator* keeps things moving forward towards making and executing a final decision.

Customizable IDAs create a mixed function allocation for the *generator* role. The system's recommendations are influenced both by the system's internal (automated) logic as well as by the users input. Cummings and Bruini's empirically evaluated their framework and explicitly tested varying degrees of automation within the *generator* role. They found that decision making outcomes suffered in a design where the *generator* role was mostly performed by the machine, in comparison to two other designs where this role was either mostly performed by the user or evenly split between user and machine. This experiment tested a system that was primarily a search engine, and we cannot say whether its result would extend to IDAs. However, it does support the hypothesis that customization could have a positive effect on decision making if implemented as a system design.

Function allocation and levels of automation are theories that have been developed to explain human behavior and performance as assisted by a broad spectrum of assistive technologies. IDAs are included in the technologies but the research used to build and validate these theories often involves systems that fall outside of the definition of an intelligent decision aid. For this reason, there is not yet convincing evidence in the function allocation literature that customizable IDAs will improve decision making, even though it offers theoretical support for this hypothesis if the findings from the broader class of decision aids are maintained within the subset of intelligent aids.

2.6.2.2 Customizable IDA Research

Additional evidence supporting the use of customization in IDAs can be found within the IDA literature.

Automation technologies such as IDAs have been criticized for being too rigid (Norman, 1990b). Rigid systems do not account well enough for the variability in the contexts in which they will be used, variability in the people who will use them, and this rigidity leads to various errors either by the system or by users. McDonald and Ackerman (2000) have applied this criticism specifically to recommender systems that use a collaborative filtering approach where items are recommended by aggregating groups with similar overall preferences. They argue that varying contexts often demand different approaches to making recommendations, but most systems are designed to maximize the accuracy of a single approach.

One way to create flexibility in an IDA is to let the user customize the system to meet the needs of the specific decision context. In other words, allowing customization enables a system to reliably be tailored to a specific decision context. Much research in HCI has sought to design and evaluate new ways to allow users to control or customize the output of an IDA. One approach to affording users some control is for the system to accept feedback from users about the recommendations. Chen and Pu (2012) describe various ways to for a system to accept feedback from users about recommendations, including structured interfaces for users to critique specific aspects of the recommendation set as well as natural language designs that let users express qualitative feedback that can be interpreted by the system and used to build a user profile. This critique-based approach to feedback is designed to afford some control to users by letting them dialogue with the system to help it build a more accurate user-profile.

Customization may also help user "buy in" to recommendations more than personalized systems that give users little control. Lee and Lee (2009) showed that IDAs in e-commerce that recommend items to buy using personalization techniques can create a sense of psychological reactance in which users feel their freedom to choose is being restricted. In their study, the sense that the IDA was restricting free choice led users to reject using the system altogether. Customization may provide a solution to this problem. By giving users choice in the recommendation-producing process, customizable IDAs may avoid this psychological reactance will still getting personalized recommendations.

Customizing IDAs has generally been shown to increase user engagement and satisfaction with the system and its recommendations (Hijikata, Kai, & Nishida, 2012; Knijnenburg et al., 2012; Parra, 2013; Burkolter, Weyers, Kluge, & Luther, 2014). However, despite these clear benefits of customization as a design choice, much less research has looked at customization and its impact on decision making. This is an important gap in the IDA research because for many IDAs, decision-centered criteria are more important than user-centered criteria. For example, if doctors enjoy using a system and come to rely on it because the customization is engaging, but it leads them to make worse decisions, than the system as a whole is harmful to its true purpose of helping to provide better care for patients. Likewise, an e-commerce recommender system that users love to use but leads them to buy products that don't actually match their preferences may be counter-productive, as users will likely return products or stop patronizing the site.

One reason that customization may help users make better decisions is by affording a "what-if" style of analysis where users can can repeatedly try out different configurations or inputs to the system and evaluate the output. What-if analysis is a common technique in business intelligence as a way to obtain decision support (Golfarelli, Rizzi, & Proli, 2006).

What-if analysis allows users test how variations in important parameters in a decision context will affect the outcome of the decision (Golfarelli et al., 2006). Customizable IDAs can enable this by allowing these changes in parameters to be specified as part of the input to a system, and users can then use the recommendations that are produced as an estimation of the effect these changes will have. What-if analysis can be effective at improving decision making when certain conditions favor it (Kottemann, Boyer-Wright, Kincaid, & Davis, 2009). Therefore, as a means of improving decision making, customization may be helpful to decision makers by affording what-if analysis.

Bostandjiev at al. (2012) have argued that a customizable IDA that allows users to adjust weights and other algorithm features can help users learn about how the algorithm work. They argue that what-if analysis is afforded by this type of customization and that it can make the system more transparent.

Users of computing technologies have a tendency to be focused and engaged with the immediate interface that they are using, and generally do not treat interaction with a computer as proxy for interaction with the people who designed and built the computer (Sundar & Nass, 2000; Solomon & Wash, 2014). It is unnatural for users to try to "get inside the heads" of system designers to reason about how the system might work. Rather, users generally are oriented towards the immediate interface with which they are interacting as the source of the interaction, and are not oriented towards other sources like the system designers.

Customization of system logic provides a means in the interface for users to naturally think about how a system *should* work based on their own understanding of the decision problem and their expertise in its domain. By requiring users to think like *a* system designer without requiring them to think like *the* system designers, customization may help users not only improve the system's recommendation but also improve their situational awareness.

In line with the notion of "self-as-source", there is evidence that when an IDA works or thinks in a fashion that is similar to users, they will like using the system and find it useful. Aksoy et al. (2006) found in controlled experiments that similarity between a recommender system's logic and users' decision making strategy helped users make better decisions and prefer to use the recommender system. This research suggests that it helps users when the system "thinks" like they do. One way to ensure that an IDA thinks like its user is to allow users to configure to match their preferred method of generating recommendations. However, this study has an important limitation that necessitates further research. The recommender system used was quite simple as it simply ranked items as the weighted sum of four attribute scores. The formula was described to users clearly, meaning that the system had an exceptionally high degree of transparency. Such a high degree of transparency at this time not realistic for many more complex methods of generating recommendations. Therefore, it is not clear whether this similarity effect will hold true when users can only infer similarity or have only partial information about the similarity between their way of thinking and the system's. And more importantly, if users can only change a part of the algorithm or system logic through customization, how is similarity between user logic and system logic perceived?

In a similar study Al-Natour et al. (2008) found that users expressed more positive opinions about a recommender system that used logic that matched the users' decision making style. In the study, subjects made a decision about a laptop purchase and then completed a survey that measured their decision making style. They then used a decision aid to get a recommendation for the decision they had just made, and this system explained the process it used to produce the recommendation. In some cases, the system used a similar decision process to the one the subject had used, and other times it used a different process. Subjects then reported their attitudes and intentions to use the decision aid, with subjects who had seen a system that shared their personal decision making process reporting more positive attitudes and intentions to use the system for future decisions.

These studies demonstrate the potential for customization in IDAs. When systems provide explanations of their logic or take other measures to establish transparency, user reactions to the system may largely depend on whether the logic used is preferred by the user. Customization provides a means to not only enhance transparency, but to also ensure that the logic is preferred by users. Establishing this assertion within a complex IDA and evaluating the actual decision making that ensues using the IDA has not been addressed in the literature.

2.6.3 Potential Problems with Customization in IDAs

In my previous work, I showed that when users customize an IDA, they become biased towards accepting its recommendations (Solomon, 2014). In this study, subjects played a fantasy baseball game, and used an IDA that recommended likely outcomes for baseball games to help them make predictions. Some users were given the opportunity to customize the system by choosing some attributes of the game that they wanted the system to emphasize when producing a recommendation. In the study, customizing the algorithm did not actually affect the recommendations, although users believed that it did. These users who customized the IDA were more likely to accept both good and poor recommendations than a set of users who were not allowed to customize the system. This demonstrates a new form of decision making bias that I have called *customization bias*. When recommendations are high quality, having customized a system may lead to improved decision making, regardless of their actual influence on the recommendations. However, if the system gives a poor recommendation, users may not scrutinize the recommendations well enough and make poor decisions.

Customization bias is a form of automation bias in that customization leads users to rely too much on the IDA's recommendations and become poor judges of recommendation quality. One reason for this may be that customization creates an *illusion of control* in IDA users. Langer (1975) has demonstrated an illusion of control in human decision making whereby people become overconfident that random events will have positive outcomes because they have made some type of choice associated with the outcome. For example, Langer showed that people were willing to bet more in a card game when they were able to blindly choose a card from the deck rather than be dealt the top card. Choosing a card from a shuffled deck does not affect the probability of a getting a good card, but people believed they could control the quality of card they drew.

However, some other recent work on illusory control has provided a different interpretation of this finding. Gino et al. (2011) have shown that findings from illusory control studies are not likely the result of a universal tendency for people to overestimate their control, but rather of a tendency to have poor perception of their actual control over outcomes. In their studies, they varied the degree of actual control over an outcome between no control (a completely random outcome) to complete control where the outcome was determined entirely by a subject's choice. Some subjects were given more moderate amounts of actual control, where their choices affected outcomes with some probability. They found that when actual control was low, people tended to overestimate their control just as in illusory control studies. However, when actual control was high, people tended to underestimate their control. Perception of control was only mildly correlated with actual control. This work suggests that people are perhaps poor judges of their control, rather than inherently biased towards overestimation.

A few studies have explored whether the illusion of control is manifest in using decision support systems. Kotteman and Davis (1994) found that users of a spreadsheet-based financial forecasting system were more confident in their decisions when the system allowed them to make adjustments to its inputs and values. However, their confidence was not warranted as they performed at the same level in their decision making as others who had used a locked-down version of the system that could not make changes. Kahai et al. (1998) replicated both of these findings in a scenario where users customized a decision aid by helping to build the statistical model it used to generate forecasts.

Another potential problem with using customizable input in IDAs is that it may enable confirmation bias. Confirmation bias is an information seeking behavior in which people seek or interpret evidence that is partial to existing beliefs or hypotheses (Nickerson, 1998). Confirmation bias has been observed in IDAs where users control the system's input (Woolley, 2007; Berner et al., 2003; Solomon, 2014; Messier Jr, Kachelmeier, & Jensen, 2001).

Related to confirmation bias is another problem with customization. Effectively customizing an IDA may require a fair amount of expertise both in using the system and in the decision domain. Berner et al. (2003) studied a clinical IDA that allowed users to customize queries for recommendations that those users who had the expertise to create good custom queries generally already knew the best decision, and merely confirmed it with the system. Those without expertise also tended to confirm their initial hypotheses using the system, but these initial hypotheses were often incorrect and therefore poor decisions were made. This finding suggests a possible paradox for customization. The people who have the expertise to use it to generate good recommendations may not actually need the IDA, while those who could most benefit from it may not be able to use it effectively to produce good recommendations.

2.6.4 Summary of Customization

Designing IDAs to be customizable by end-users has a number of important theoretical advantages. These include:

- Allowing users to personalize the system's inner logic to best meet their specific circumstances, potentially leading the system to make better recommendations.
- Adding transparency to the system so that users have an adequate understanding of how it works and how its recommendations have been produced.
- Giving users an opportunity to think critically about their decision and build situational awareness.
- Helps users build trust in the IDA so that they will rely on it when it gives good recommendations.
- Allows users to conduct "what-if" analysis that can help them learn both about the system as well as the data that inform it, potentially leading to insight about the decision.
- Situates the system at a moderate degree of control for activities of information acquisition and information analysis, reducing the potential for automation bias and complacency.

However, customization may also present several new challenges that need to be investigated and considered in theories that inform IDA design. Some of these new challenges are:

- Creating a bias whereby users are prone to agreement with system recommendations even when the recommendations are poor.

- Creating a misplaced sense of control, where users incorrectly judge the effect of their input on the system's output, leading to poor decision making.
- Enabling confirmation bias by letting users tailor information acquisition and analysis so that it confirms prior beliefs
- Users may be able to customize a system so that it matches their subjective preference for how a system should work or how decisions should be made, which may not objectively be the best approach for a given situation.
- Customizing an IDA effectively may require expertise that users who can most benefit from an IDA may not possess.

2.7 Summary

Intelligent Decision Aids can be powerful in helping their users acquire and analyze information and select actions as part of a decision making process. A variety of powerful technologies have been developed that can automate the processing and analysis of large-scale data and use that processing to make recommendations to users about a particular decision. The application of these technologies to decision making in medicine and health care, business and e-commerce, and many other important domains suggests that computer-assisted decision making is seen as having tremendous potential for improving decision processes and outcomes.

However, to date there is some concern that the theorized benefits of these technologies are not being realized. One of the concerns is that while the technologies that drive these systems are powerful, the design of the interaction between users and systems has not been perfected. Human factors problems such as miscalibrated trust, automation bias, confirmation bias, and other cognitive biases have been shown to limit the effectiveness of these systems in helping users and decision makers actually improve their decision making. For this reason, user-centered approach to IDA design that considers users as an integral part of the system is necessary. Such a user-centered approach to IDA design demands a strong theoretical understanding of how characteristics and features of an IDA design impact users' decision making.

One of the most widely studied design characteristics in IDA research is the notion of transparency. Since the underlying system logic, including both algorithms or statistical models as well as databases, is often too large, complex, or varied for users to easily see, IDAs are often thought of as "black-boxes." However, much work has sought to design transparent IDAs that offer some form of insight to users about how a system works to produce recommendations. Many benefits have been demonstrated to making IDAs transparent. However, it is not yet clear whether improved decision making is one of those benefits. The little research that examines that question is inconclusive, with some work even suggesting that transparent IDA may inhibit some aspects of decision making. For this reason, IDA designs that provide transparency must be carefully evaluated in order to determine how the design may be impacting decision making.

One such design feature that requires a careful evaluation is customization. Customization may make an IDAs more transparent by allowing users to play a dual role as both system user and system designer. A customizable system that allows users to configure it in a way that they expect will help it personalize its recommendations provides some automatic transparency to users. Since users know what they have done to a system in customizing, they gain at least some transparency into how it works.

However, customization as a design feature may have other consequences for system usability and for computer-supported decision making. Customizing a system may create or enhance some biases that arise when users rely on automated aids to assist in decision making. Users may become biased towards agreeing with recommendations, they may try to customize a system to confirm an existing hypothesis, they may have miscalibrated trust or unrealistic expectations for how well the system will work as a result. One of the difficulties with customization of an IDA's internal logic is that in most cases, customization only provides partial control to users. Therefore, their actions to configure the system must interact with a number of other potentially invisible factors within the system's logic, and users may not be able to easily interpret how they have affected the system's output by way of their customization of the system.

CHAPTER 3

CUSTOMIZATION BIAS

3.1 Introduction

Customization is a design approach to creating personalized recommendations. Rather than completely using artificial intelligence or computational techniques to personalize and tailor recommendations to the user and the specific decision context, customizable IDAs allocate some of the responsibility of personalizing recommendations to the users themselves.

From the perspective of a designer, there can be several goals behind using this design. One important goal can be to leverage users' knowledge of their local circumstances, preferences, and situational awareness to help the system produce better recommendations for users. Another goal can be to give users a sense of control or agency that produces a positive user experience and attitude towards the system.

There is an assumption behind this goal, which is that when a system produces good recommendations, then users will accept these recommendations and consequently make good decisions. Herlocker et al. (2004) suggest that recommendation quality is often the fundamental focus of IDA designers and that other evaluation criteria, particularly human-centered criteria, are often ignored. I argue that in most cases, IDA should be most critically evaluated based on the decisions that their users make rather than offline technical criteria or user-centered criteria such as usability or user satisfaction. However, this is a tremendous challenge for IDA research because of the complexity of evaluating decision making, particular the types of decisions that IDA are frequently designed to assist that have high uncertainty, high stakes, time pressure, and variability in decision makers' expertise (Klein, 2008). Also, for many decisions, not everyone agrees on the best criteria for which decisions should be evaluated. IDAs may for example be used to recommend products that are *hori-*

zontally differentiated (Cremer & Thisse, 1991), which means that a good decision for some will be a poor decision for others.

Furthermore, since decision quality may be difficult to evaluate, particularly in real time, the quality of recommendations that a system provides cannot be easily communicated to users or even known to the system prior to providing recommendations. This presents a challenge for IDA-supported decision making. When an IDA gives a recommendation, should users follow the recommendation? How can users know how how good a recommendation is and whether or not recommendations are trustworthy? Using IDA to support decision making adds some complexity to decisions in that IDAs provide new but aggregated information in the form of recommendations that users must evaluate as part of the decision process.

As discussed in Chapter 2, there is an abundance of evidence that IDA users often make decision errors. Users often follow poor recommendation from a system (Skitka et al., 1999), although users may also fail to follow good recommendations which hurts their decision quality. From a human factors engineering perspective, understanding what leads users to follow or deter from IDA recommendations is a critical aspect of understanding how to make systems that improve decision making. As argued by Dzindolet et al. (2003), a goal of human-centered IDA design should be to calibrate users' reliance on IDAs' recommendations with the system's "reliability," which is to say the quality of the recommendations they produce. If users can effectively detect recommendation quality, they can use an IDA to make good decisions.

In this chapter, I will present a study that examines how IDA designs that afford end-user customization can impact user decision making, leading to a decision making bias called *customization bias*. Customization bias occurs when users become partial to accepting recommendations from IDAs as a result of their involvement in customizing its algorithm. I first observed this bias in my previous work (Solomon, 2014) on customizable IDAs. This study will build on that work by testing a theoretical mechanism by which customization bias is enabled.

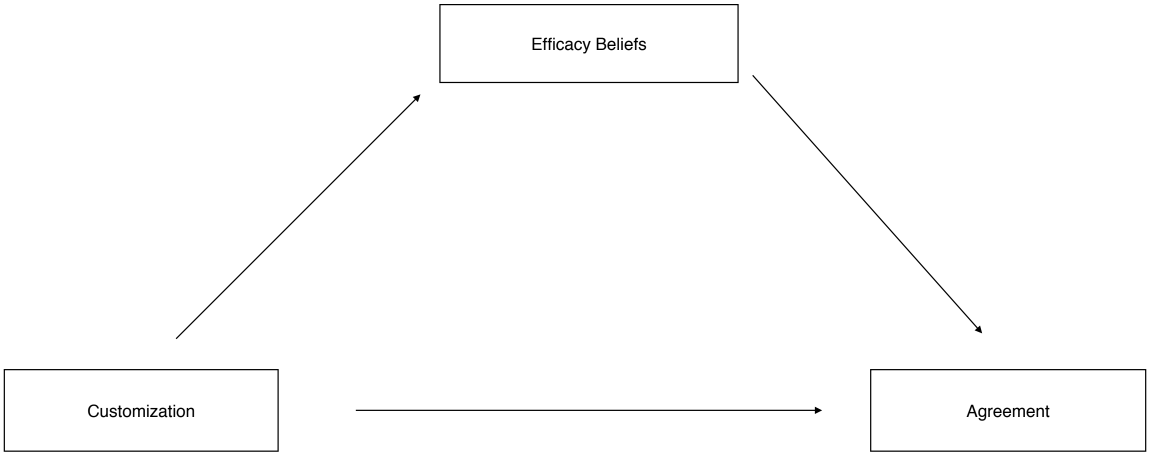


Figure 3.1: How customization can create agreement with IDA recommendations.

IDAs that are customizable afford users the opportunity to tailor the system's inner logic to match their preferences for how recommendations are produced. The intent of this tailoring is to allow the user to create an algorithm that works in a way that the user believes is efficacious. In other words, customization allows users to make the system work in a way that they believe will be successful at producing good recommendations. I will refer to users' beliefs about the quality of recommendations that users expect a system to produce as *efficacy beliefs.*

One possible reason for customization bias is that by allowing users to tailor the algorithm, they develop inflated beliefs about the system's efficacy. Users may believe that their actions in customizing the IDA will uniformly improve the system, ignoring the possibility that they have harmed the algorithm's performance or had little effect.

The concept of users' efficacy beliefs is an important construct both for understanding customization bias as well as more generally for IDA design. How do users form expectations or beliefs about the efficacy of a system? Particularly when an IDA lacks transparency or when the user has little experience with a system, users may have little information that allows them to adequately assess how well a system is likely to work at producing good recommendations. But this belief may nonetheless impact their decision. If users have little knowledge about a decision or about a system, they may have little to fall back on when making decisions other than a belief.

Figure 3.1 illustrates the theoretical relationship between customization, efficacy beliefs, and agreement. This figure represents three relationships. First, it suggests that customization causes increased efficacy beliefs. Second, it suggests that efficacy beliefs cause agreement. Together, these two relationships are an argument for *mediation*, in that the effect of customization on agreement is mediated by users' efficacy beliefs. This figure also suggests a third relationship, which is a direct relationship between customization and agreement that is not related to efficacy beliefs.

This diagram suggests a causal chain from customization to agreement. Customization causes users to increase their beliefs in system efficacy, and because they have increased their efficacy beliefs they will then be inclined to agree with the IDA's recommendation. It also suggests that the direct influence of customization on agreement is causal in nature. In this chapter, I will present an experiment that tests the causal relationships between customization and efficacy beliefs, as well as the direct relationship between customization and agreement. This experiment will also look at the relationship between efficacy beliefs and agreement, although it will only offer weak evidence that the relationship is causal. However, the causality of efficacy beliefs on agreement will be tested in a subsequent study reported in chapter 4.

3.2 Methods

To evaluate the role of customization in decision making with IDA, I created an experiment where IDA users are given recommendations purportedly generated by a complex algorithm. Some users had the chance to customize the IDA to influence its recommendations, but in reality the customization had no effect on the recommendations. This design tests whether the act of customizing an IDA influences decisions while holding the quality of those recommendations constant. This allows for a comparison of decision making between users who customize an IDA and those who do not but who receive identical recommendations. By holding recommendation quality constant between conditions, this design evaluates the effect customization has directly on the decisions that are made by users, rather than any effect that is due to the change in recommendations that comes from customizing the IDA.

The decision in the experiment was a fantasy baseball prediction game in which subjects tried to predict the scores of Major League Baseball games after being shown statistics about the teams involved. This task has several advantages. First, it is a task with a low threshold for expertise, since many people in the general population follow baseball and play similar games. Second, it is a task for which IDA-like tools are frequently used to help people make choices. Third, it is a task that requires a decision to be made without the availability of all possibly relevant information. Even the best statistical simulators are not perfectly reliable in predicting game outcomes, and therefore some judgement or extra knowledge from the decision maker is required. Fourth, it is a task that can involve both a binary, ”yes/no” type decision (”Which team will win?”), as well as a continuous outcome (”How many runs will each team score?”), and each of these outcomes can be objectively compared to actual outcomes.

249 subjects were recruited from Amazon Mechanical Turk to play this game as part of a study to “help improve an algorithmic tool for aiding decisions in fantasy baseball.” In order to complete the experiment, subjects had to first take a timed test on the basic rules and

statistics of baseball. This quiz is described in greater detail in Appendix B. 73 of subjects did not successfully complete this quiz. These subjects were paid $0.40 and screened out of further participation. This basic knowledge was equivalent to the minimum knowledge required to play fantasy baseball. In order to enroll in the study, subjects were required through Mechanical Turk's system to be in the United States and to have completed at least 95% of their previous assignments on Mechanical Turk. Mechanical Turk workers who had participated in pilots of this study or in my previous study using this task (Solomon, 2014) were also not eligible to enroll. Subjects were paid $2 for participation. Subjects were also promised an additional payment that would depend on their performance in the game, and were told that the average expected payment would be $2.25. Subjects took an average of 14.4 minutes to complete the experiment. Six subjects were removed from the final data set because they completed the study in less than five minutes. In pilot testing I determined that five minutes was not sufficient to be able to complete the study while giving any thought to the decisions. Two additional subjects were removed because they had no matching subject from the non-customization condition (this is explained in detail in section 3.2.3).

The final dataset contained 168 subjects. The subject pool was 73% male with an average age of 33 years old.

3.2.1 Game Play

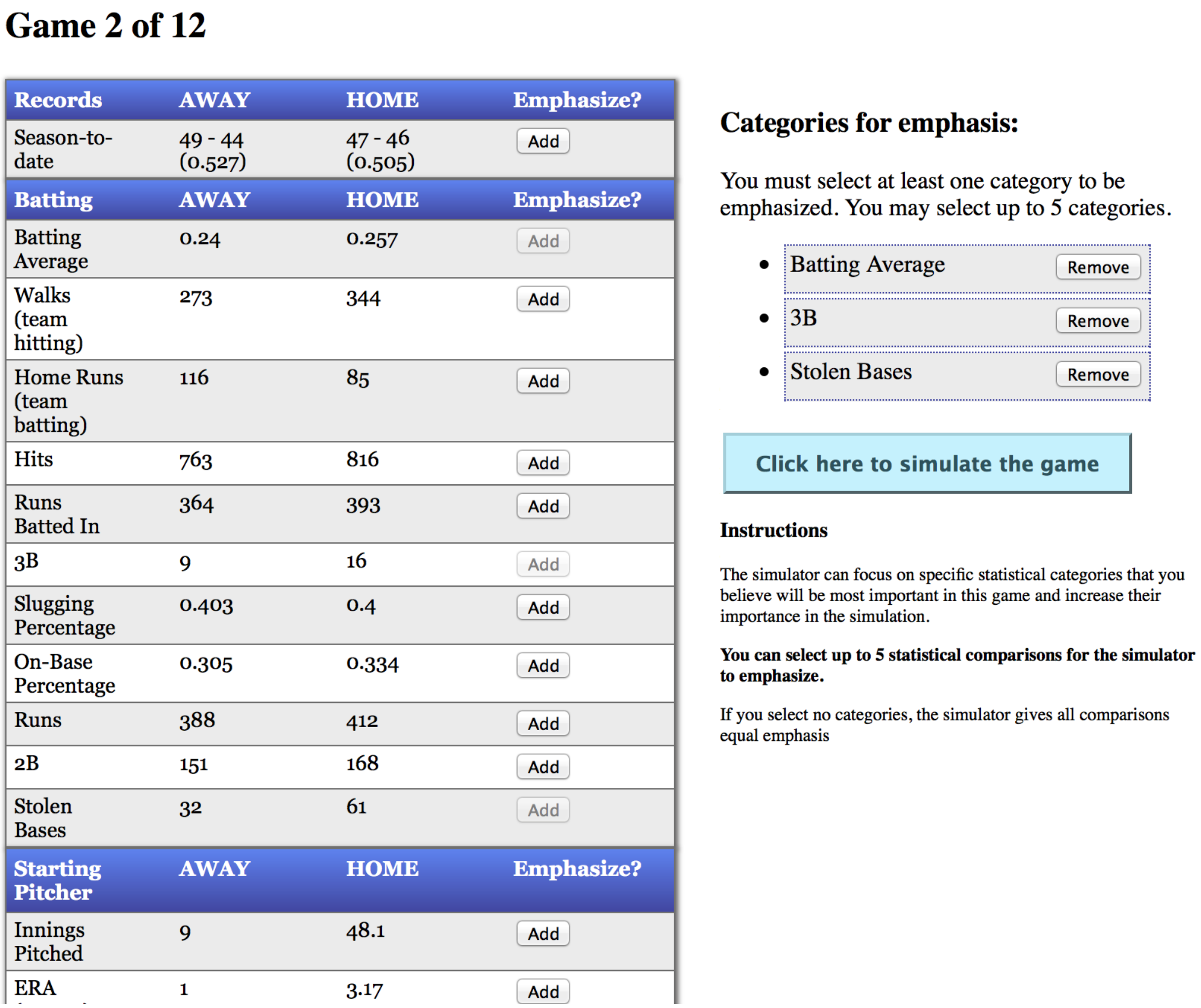
Subjects first read instructions and were required to pass a difficult quiz on the instructions. On average, subjects required 2.1 attempts to pass this quiz. The difficulty and time required to pass the quiz ensured that the distributed online sample had sincere motive to participate and that they adequately understood the game. Subjects played 12 rounds of the fantasy baseball prediction game. In this game, subjects were shown extensive statistics about two teams and asked to make a prediction about the score of the game between the two teams. To ensure that only the available statistical information was used to inform decisions, the names of the teams were not revealed to subjects. Additionally, the baseball games that subjects were predicting were games that had already been played. Subjects were told that even though the games were past games, all statistics and algorithms in the study treated the games as if they were in the future.

I selected games for the experiment from the 2011 and 2012 Major League Baseball seasons using several criteria. I fit an existing statistical model (T. Y. Yang & Swartz, 2004) for assessing the probability of a home victory to games from these seasons. This model estimates the probability that a home team will win using the relative strength of each team in three categories: winning percentage, the Earned Run Average of the starting pitcher, and Batting Average. The model also includes an adjustment for home field advantage. All data about Major League Baseball games and players was taken from Baseball-Reference.com[1].

I selected games to match the approximate distribution of probabilities estimated by the model. I chose two games where the predicted winner had less than a 60% chance of winning, four games where the predicted winner had a 60-70% chance of winning, four games were selected with probability between 70 and 80%, and two games with probability greater than 80%. Only games where the team with the higher expected probability actually won the game were included in the final set of 12 games. These 12 games were presented in a random order to each subject.

Subjects earned points in the game by making accurate predictions about the outcome of the game. Subjects earn 20 points if they predict the exactly correct score. If they choose the wrong winner, they lose 10 points from their score. They also lose one point for the absolute difference between the predicted number of runs for each team and the actual number of runs. For example, if the final score of a game was Away 5 — Home 3 and the subject predicted Away 4 — Home 6, the subject would lose 10 points for choosing the wrong winner, lose 1 point for missing the Away run total by 1, and 3 points for missing the Home run total by 3, leaving a total of 6 points for the game. This baseball task, like many decisions, has a clear "best possible outcome" yet no clear "worst-possible outcome" since

[1] http://www.Baseball-Reference.com



Game 2 of 12

Records	AWAY	HOME	Emphasize?
Season-to-date	49 - 44 (0.527)	47 - 46 (0.505)	Add
Batting	**AWAY**	**HOME**	**Emphasize?**
Batting Average	0.24	0.257	Add
Walks (team hitting)	273	344	Add
Home Runs (team batting)	116	85	Add
Hits	763	816	Add
Runs Batted In	364	393	Add
3B	9	16	Add
Slugging Percentage	0.403	0.4	Add
On-Base Percentage	0.305	0.334	Add
Runs	388	412	Add
2B	151	168	Add
Stolen Bases	32	61	Add
Starting Pitcher	**AWAY**	**HOME**	**Emphasize?**
Innings Pitched	9	48.1	Add
ERA	1	3.17	Add

Categories for emphasis:

You must select at least one category to be emphasized. You may select up to 5 categories.

- Batting Average Remove
- 3B Remove
- Stolen Bases Remove

Click here to simulate the game

Instructions

The simulator can focus on specific statistical categories that you believe will be most important in this game and increase their importance in the simulation.

You can select up to 5 statistical comparisons for the simulator to emphasize.

If you select no categories, the simulator gives all comparisons equal emphasis

Figure 3.2: Customizable IDA.

one can theoretically predict scores that deviate infinitely from the actual score. Similarly, doctors could prescribe the wrong medicine and also prescribe a dosage and duration that deviate infinitely from the option that in truth would be most beneficial to a patient. The presence of clear “best options” means that any deviation is a loss to the decision maker. For this reason, the scoring system identified a clear best option and the IDA was capable of recommending this best option, and deviation from this optimal decision was represented as a loss.

3.2.2 IDA and Conditions

All subjects used an IDA that provided extensive statistical information about the teams involved in each of the games. In addition to providing statistical information, the IDA also recommended its own prediction about the score of the game. Subjects were told this prediction was based on a statistical algorithm. However, the recommendations were actually pre-determined for each game. There were two types of recommendations. Good recommendations suggested the actual score of the game, yielding 20 points if followed exactly. Poor recommendations suggested the wrong winner, as well as a score that would yield 5 points. Subjects were given poor recommendations for four games (one random game for each difficulty level), and good recommendations for the remaining games. Over the 12 games, the average score of the IDA's recommendations was 15. Subjects were told of this average, and that there was considerable variation in the quality of the recommendation.

There were two conditions of the experiment. In the customizable condition, subjects had the opportunity to make adjustments to the IDA's recommendation algorithm after seeing a table of statistical comparisons between the teams (see Figure 3.2). Subjects were asked to choose between one and five statistical categories to receive extra emphasis in the simulation algorithm. For example, a subject could select winning percentage, home runs, and starting pitcher ERA and the algorithm would then emphasize the contribution of these statistics when estimating the game's outcome. The instructions stated that good customization improves the performance of the algorithm, but poor customization could harm performance.

In the non-customization condition, subjects saw the same table of statistical comparisons as the customization condition. The interface had pre-loaded a set of categories that would be emphasized, and the buttons used to configure the system were disabled so that changes could not be made to the configuration. In the instructions, subjects were informed that the pre-loaded categories were configurations that had been used by previous users of the system who were completing the same task. Details of how the configuration was selected for users in the non-customizable condition are described below in the section 3.2.3.

After viewing the statistics (and if in the customization condition customizing the IDA), subjects clicked a button to generate a recommendation about the outcome of the game that they could use to help them make a decision. Prior to seeing the recommendations, subjects answered a three question survey to assess their belief about how well they expected the system to perform. This survey is described in the measures section below. Subjects were then shown the IDA's recommendation and given the opportunity to make their own prediction about the game outcome. Once subjects had submitted their decision, they were directed to the next round until all 12 rounds were completed. Subjects then took a post-test questionnaire, and were given a code to return to Mechanical Turk to be submitted for payment. After the entire study was completed, subjects were sent a message with a breakdown of their score for each game and a debriefing statement about the true nature of the IDA. Subjects were shown all their scores after the study, rather than immediately following each round, to reduce the nuisance factor of learning about the IDA or the decision scenario over the course of the study.

3.2.3 Subject Matching

A potential confound can arise in any study that compares a customizable system to a non-customizable system. When subjects customize a system, they have set its configuration. Subjects who do not customize a system must nevertheless use a system that has been configured in some way. If the configuration for the system used by the non-customizers is pre-determined, as was done in my previous work (Solomon, 2014), there is a confound in that the customization condition differs both in the *act* of customizing as well as the *product* of customizing (i.e. the configuration that is used). Differences between the conditions may be because of the act of customizing. But they also could be because of specific configurations that are used. For example, if non-customizers simply do not think that the default configuration is a good configuration, they may be less prone to agreement not because they haven't chosen the configuration but because they believe it is truly a poor configuration.

The purpose of this study is to understand how the *act* of customizing an IDA influences decision making, and is not concerned with *product* of customizing. This is because the act of customizing is generalizable to other systems and other decision scenarios besides the one used in this study. The product of customizing is specific to the task and IDA used in the study but will not have applicability to other systems (e.g. knowing which categories of baseball statistics people select won't tell us anything about what parameters an investor should use to customize a stock recommender).

For this reason, this study was designed to remove this confound and ensure that only the act of customizing was varied between conditions. To do this, the study was run one condition at a time. The customization condition was run first, followed by the non-customization condition. The configurations that subjects in the customization condition used were assigned to subjects in the non-customization condition. This ensured that subjects in the non-customization control condition were using the same configurations as those in the customization condition.

Rather than randomly assigning configurations to non-customizers, a collaborative filtering technique was used to match non-customizers to the most similar subject from the customization condition in terms of their attitudes about the specific baseball statistical categories. This was done to reduce differences in subjective opinions about which categories are most effective for a computerized simulator in estimating game outcomes. For example, some people may feel that Home Runs and Earned Run Average are the most informative statistical categories, whereas others may think On-Base Percentage and Stolen Bases are better categories. If we assume that customizers will choose their preferred categories most of the time (and this assumption is supported by the data shown in Figure **??**), then subjects who customize may have different outcomes in the study as a result of this preference for the configuration rather than the act of customizing. The collaborative filtering technique

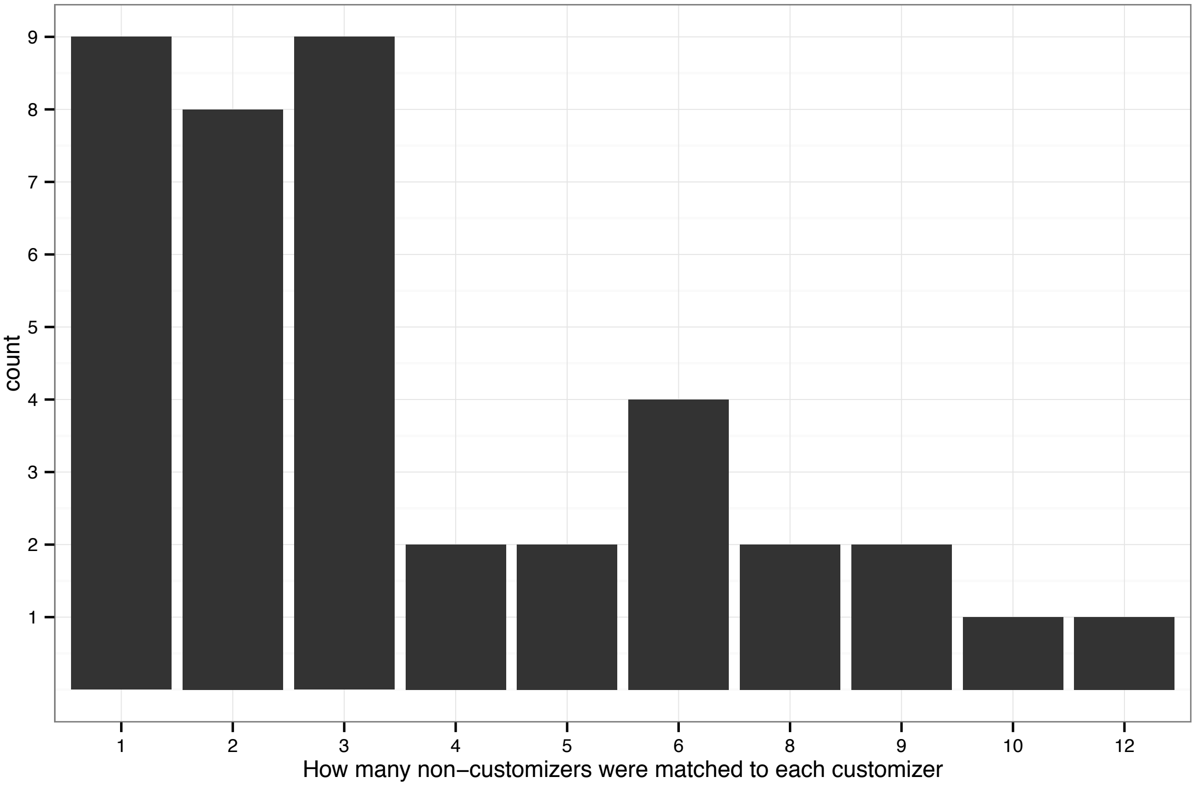


Figure 3.3: Distribution of matches per customizer.

is intended to, as best as possible, assign non-customizers configurations that they prefer equivalently to the customizers.

To execute this collaborative filtering matching, subjects took a pre-test survey that asked them to assess how informative each of the 27 statistical category options might be to a computerized baseball outcome simulator. They answered a question about each category on a 5-point scale. When each non-customizer finished this survey, the cosine similarity between their responses and the responses of every subject from the customization condition was calculated. Then, the non-customizer was matched with the customizer with the highest cosine similarity. The non-customizer then did the baseball prediction task using the same order of games, and for each game the IDA's configuration was shown as the same configuration that the customizer had used.

A consequence of this matching approach is that two subjects who customized the system never got matched to any non-customizers. These subjects were removed from the final data set, leaving 49 subjects in the customization group. In an effort to find matches for the majority of customizers, and to obtain multiple non-customizers, 119 additional subjects were recruited for the non-customization condition. The matching distribution was not uniform. 9 customizers only received one matching non-customizer. One customizer received 12 matches, which was the maximum. This distribution is shown in Figure 3.3. The average cosine similarity between matches was 0.983, with a standard deviation of 0.014. Overall, subjects in the non-customization condition had very similar responses to their match from the customization condition and there was only a small amount of variance across subjects in their similarity. For this reason, the uneven distribution of matches is not likely problematic because non-customizers all had very nearly the same similarity to their matching subject from the customization group.

A limitation of this study design is that the matching technique creates a mild violation of random assignment because subjects who signed up earlier for the study were more likely to be in the customization group. An analysis of all demographic and pre-treatment variables did not find any statistically significant or even potentially meaningful differences between subjects in each condition. Therefore, I have no reason to suspect that this bias impacted the results of the experiment.

3.2.4 Measures

Agreement. There are two measures of agreement between subjects' predictions for the outcomes of a game and the IDA's recommendation. *Winner Agreement* is a binary measure of agreement. Winner agreement is coded as 1 if the subject picks the same team to win as the IDA's recommendation, and coded as 0 if they choose the opposite team.

Score Agreement is a continuous measure that uses the game's scoring mechanism. Score agreement is measured using the same method as subjects' point totals as described above, however rather than using the true outcome of the game as the comparison, it uses the IDA recommendation. For example, if the IDA predicts Away 2 — Home 6, and the subject predicts Away 4 — Home 5, score agreement would be 15 since the subject starts with 20 points, then loses 4 for the difference in Away scores and 1 for the difference in Home scores. This measure is included to provide a granular indication of agreement and reliance on an IDA. Most studies that examine reliance on decision aids use only a binary task so that reliance or agreement is measured as a frequency. However, many IDA-supported decisions have more granularity in the options users have. A doctor, for example, can use an IDA to determine a dosage or duration of a treatment. The IDA may give a specific recommendation, and the doctor may be influenced by the recommendation but may make a small adjustment to it. Using only a binary measure of agreement would not capture that influence, and therefore both a continuous and binary measure are valuable to understanding agreement.

Beliefs of System Efficacy. Subjects' beliefs about system efficacy were measured using three indicators. Subjects answered the following question on a 7-point Likert scale: *Based on the categories that are being emphasized, how well do you expect the simulator to perform at making its prediction?* A second measured asked subjects to estimate the number of points that the IDA's prediction would earn if it were scored. The third measure asked subjects to assess the probability that the simulator's prediction would choose the correct team to win the game.

Table 3.1: Factor analysis of efficacy beliefs measures.

Statistic	Factor Loading (Std. Error)	Mean	St. Dev.	Min	Max
7-point Likert	1.000 (–)	4.773	1.291	1	7
Expected Points	2.272 (0.095)	12.931	3.714	0	20
Correct Winner	9.240 (0.370)	72.859	12.934	50	100
		Fit Indices			
Comparative Fit Index		1.000			
Root Mean Square Error		0.000			

A confirmatory factor analysis was used to evaluate the reliability between these measures from the data. Table 3.1 describes this factor analysis. Overall, the three measures were consistent with each other and loaded onto a latent variable with a strong fit. In the analyses presented in the results section, the factor score for each observation that resulted from this analysis was used as a single variable called *efficacy beliefs*. The analyses were later repeated using each of the individual indicators separately, and this made no difference to any of the conclusions that were drawn from the analyses.

Propensity to trust automation. Since subjects may have individual differences in their attitudes about automation or decision aids, I measured subjects' propensity to trust automated decision aids using the scale developed and validated by Merritt et al. (2012). This scale has 6 items that are listed in Appendix A. Cronbach's alpha for these items was 0.72. To obtain a factor score for each subject as a measurement of propensity to trust automated decision aids, I conducted a confirmatory factor analysis that included all six items onto a single factor. This analysis is described in Table 3.2. The factor scores for each subject were calculated and used as a single variable called *automation trust propensity* in the subsequent analyses. These items were administered in the pre-test, before subjects saw the IDA or had any other information about it.

Table 3.2: Factor analysis of propensity to trust automated decision aids items.

Statistic	Factor Loading (Std. Error)	Mean	St. Dev.	Min	Max
Trust 1	1.00 (–)	3.494	0.922	1	5
Trust 2 (Reverse)	-0.733 (0.088)	2.589	0.993	1	5
Trust 3	0.941 (0.071)	3.190	0.928	1	5
Trust 4	1.125 (0.066)	3.137	0.966	1	5
Trust 5	1.126 (0.073)	3.327	1.024	1	5
Trust 6	0.947 (0.084)	3.107	1.033	1	5
		Fit Indices			
Comparative Fit Index		0.977			
Root Mean Square Error		0.104			

Category ratings. For a single configuration of the IDA, which could emphasize between one and five categories, subjects' overall rating of the configuration was calculated as the average rating given to all the categories within a configuration. This is a measure of subjects' beliefs about whether the emphasized categories are good indicators of game outcomes.

Control, helpfulness, accuracy, and importance. In the post-test questionnaire, subjects answered questions and responded on 7-point Likert scales about how they used the IDA for their decisions. These questions and their associate constructs were:

- *Control*- I was able to control the accuracy of the simulator.
- *Helpful*- How helpful was the simulator?
- *Accurate*- How accurate do you think the simulator was at predicting the outcome of games?
- *Important*- How important were the simulator's predictions in informing your predictions?

Decision quality. The number of points earned from a decision was used to measure decision quality.

3.3 Hypotheses

H1. *Subjects who customize the IDA will believe that the system has higher efficacy than those who do not customize the system.* The illusion of control (Langer, 1975) would predict that users who make choices about how the IDAs algorithm works will believe it works better than other users who use an identical algorithm but do not make the choice to use it in the IDA. Importantly, this illusion is strictly the result of making the choice about how the IDA is configured, and not the result of users configuring it using categories that they feel are better predictors of game outcomes.

H2. *Subjects' beliefs about the IDA's efficacy will be predictive of their agreement with its recommendations. Subjects will have more agreement when they believe the system has greater efficacy.* I expect users who believe the IDA's process for producing recommendations is effective to evaluate recommendations to be more accurate and trustworthy, and therefore these users will agree with recommendations more than those who feel the process is ineffective.

H3. *Subjects who customize the IDA will have greater agreement with its recommendations than those who do not customize.* I expect the results of this study to be consistent with my previous work (Solomon, 2014) which found that users are more likely to agree with recommendations when they believe they have customized the IDA. The design of this study allows this hypothesis to be broken into two pieces:

- **H3a.** *The effect of customization on agreement will be partially mediated by efficacy beliefs. Subjects who customize will agree more with the IDA in part because by customizing the system they increase their beliefs in its efficacy (H1 and H2).* If both H1 and H2 are supported, it would follow that customization can cause agreement by causing an increase in efficacy beliefs, which then causes more agreement.

- **H3b.** *There will be a direct effect of customization on agreement. Subjects who customize will agree more with the IDA for reasons other than an increased belief in its efficacy.* There are at least two mechanisms other than efficacy beliefs by which customization might plausibly cause greater agreement. One is a confirmation bias. If, for example, a user looks over the statistics of the teams and forms an opinion that the away team will win because they have a better pitcher, she might configure the IDA to emphasize pitching statistics, which are consistent with her initial opinion. Then if the IDA recommended the away team to win, she may treat the recommendation as confirmation by the system that her initial opinion has merit and be inclined to agree with the recommendation. Another potential mechanism is the effort required to configure the IDA. The IKEA effect (Norton, Mochon, & Ariely, 2011) illustrates that people are more inclined to buy products when they have participated in created them. If this effect extended to following IDA recommendations it would predict greater agreement with recommendations by customizers. Effort might also cause agreement by creating fatigue in users, such that after exerting effort to customize the IDA users are fatigued and not interested in exerting more cognitive effort to scrutinize and evaluate recommendations closely, and instead just choose to agree because that is easier than forming their own prediction. And there may also be other unconsidered reasons that customization may increase agreement other than through efficacy beliefs. The purpose of this study is not to identify any mechanisms other than efficacy beliefs. Rather it is intended to identify whether efficacy beliefs are a mechanism for customization's effect on agreement and whether there is a need for future research that explore and identify other mechanisms.

H4. *Subjects who customize the IDA will make better decisions, earning them more points in the game, than those who do not customize.* I expect this to happen because customizers are inclined to agree with the IDA's recommendations (H3), and since the IDA gives mostly reliable recommendations, it will be more useful to them in making good decisions in the baseball prediction game.

H5. *Subjects who customize the IDA will report feeling more control over the system than subjects who do not customize.* Previous work on customization in web portals has found that customization increases users' sense of control (Marathe & Sundar, 2011), and I expect that this can be replicated by users of an IDA.

H6. *Subjects who customize the IDA will report that the system is a) more helpful to them as they make decisions, b) generates more accurate recommendations, and c) is a more* *important part of their decision-making process than subjects who did not customize.* Customization has been found to be beneficial to users' perceptions of the system and user experience (Hijikata et al., 2012; Knijnenburg et al., 2012). I expect this to hold true in this study and cause users feel the system is more accurate and beneficial to their decision process.

H7. *Consistency between the recommendations that the IDA gives and the configurations used will lead to greater agreement by subjects. When the team that is predicted to win the game by the IDA is stronger in the categories that were used in the configuration, subjects will be more likely to agree with the IDA, even if the recommendation is poor.* In my previous work (Solomon, 2014) I found that users who customized were more likely to agree with recommendations when they appeared to be consistent with how they had configured the system. That study could not assess whether this would be true for users who do not customize. This study however enables this assessment, and I expect subjects in both conditions to agree more when the system's recommendations are consistent with configurations.

- **H7a.** *This effect will be stronger for subjects who have customized the IDA than those who did not customize.* The confirmation bias described under H3b will make the effect of consistency stronger for users who customize because they have the opportunity to configure the IDA to match their initial opinion more precisely than those who do not customize.

3.4 Results

3.4.1 Descriptive Statistics

The means and standard deviations for the three measures of efficacy beliefs are reported in Table 3.1. In general, subjects had fairly positive beliefs that the IDA would produce

Table 3.3: Descriptive statistics.

Statistic	N	Mean	St. Dev.	Min	Max
Winner Agreement	2,016	0.767	0.423	0	1
Agreement	2,016	15.446	5.877	−2	20
Points Earned	2,016	14.861	5.905	0	20
Control	168	2.988	1.672	1	7
Helpful	168	5.024	1.624	1	7
Accurate	168	4.497	1.202	1	7
Important	168	4.976	1.529	1	7

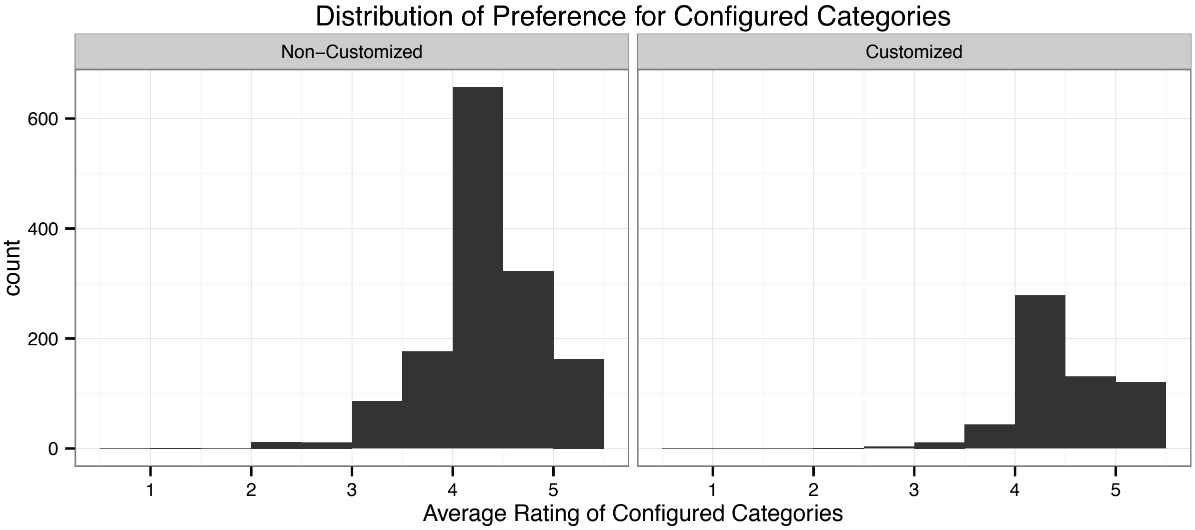


Figure 3.4: Preference for configured categories.

good recommendations, although responses across the spectrum of the three variables were observed.

On average, subjects agreed with the IDA's predicted winner 77% of the time, with an average score agreement of 15.47 (a perfect agreement score is 20). These numbers closely match the reliability of the IDA, which predicted the correct winner 67% of the time and scored an average of 15 points. This indicates that subjects had reasonably well-calibrated reliance on the IDA in terms of the frequency with which they followed its recommendations. However, as will be discussed below, calibration of the frequency of reliance is not the same as discernment of good and poor recommendations. Subjects did not always make good decisions by following good recommendations or rejecting poor recommendations. On average subjects earned 14.86 points per round, although the range of points earned spread between the minimum and maximum possible (0 to 20 points).

Distributions of the other measures are listed in Table 3.3. Figure **??** shows the distribution of the subjects' category ratings, broken into the two conditions of the study. As can be seen from this figure, in both conditions subjects were very favorable towards the categories that the IDA used in a round. On only a handful of occasions did anyone give an average rating of the categories lower than the neutral point of 3 on the scale. So it appears that in both conditions, subjects felt the categories being used were good categories that are informative to predicting the outcome of a baseball game. This supports the assumption stated above that users will tend to choose categories that they believe work work well in producing recommendations.

It is important to evaluate how well the matching mechanism of the experiment worked at assigning configurations that matched the preferences of subjects in the non-customizable IDA condition. A linear model was fit to test whether subjects in the customizable IDA condition rated their chosen categories higher than those in the non-customizable IDA condition, who had been assigned configurations expected to match their preferences. This model included a random effect for each subject to account for having multiple observations for each subject. The model found a statistically significant difference between conditions ($p < .05$), with customizers preferring their categories more than non-customizers. However, although statistically significant, the magnitude of this effect was very small. The difference in means between conditions was 0.17 on the five point scale. Overall it appears that the matching mechanism worked reasonably well at pairing subjects with similar beliefs about how statistical categories might help a computerized simulator.

Because subjects played the game repeatedly, there was the possibility for them to learn or adapt their decision making over the course of the experiment. Figure 3.5 shows how subjects

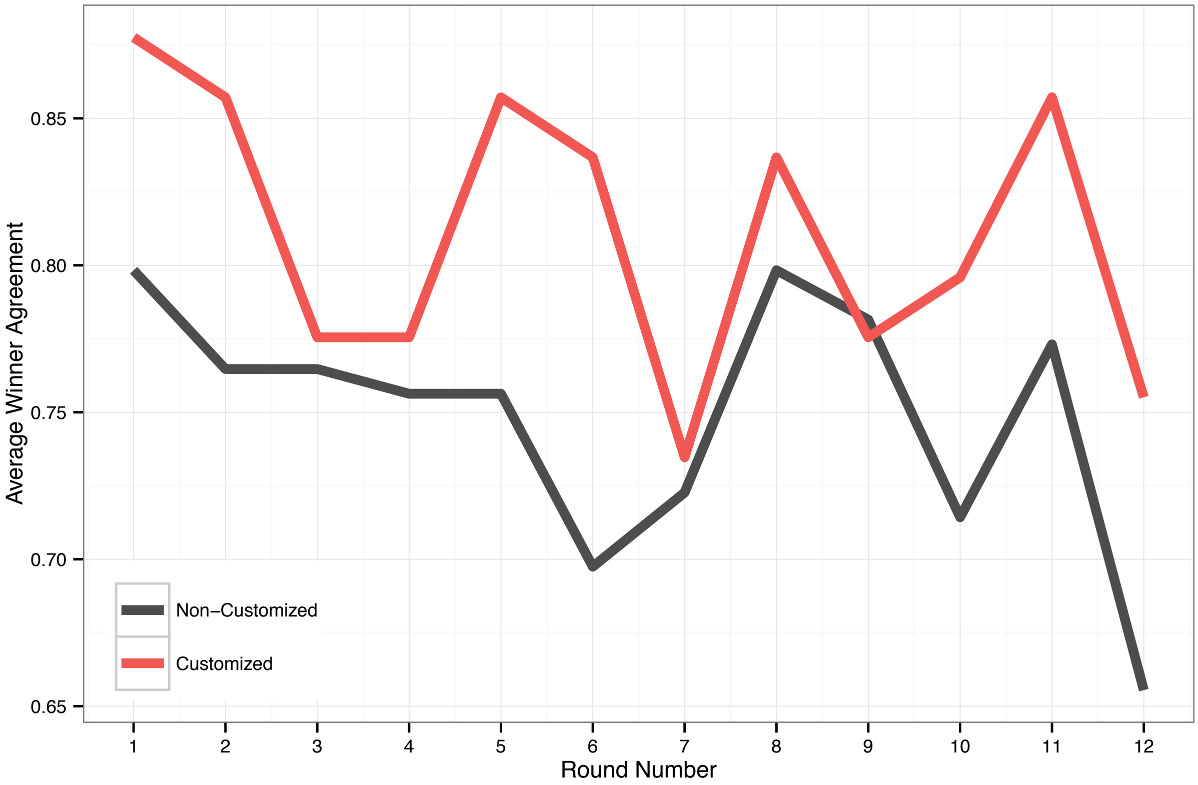


Figure 3.5: Average winner agreement by round number.

adjusted their agreement with the IDA over the rounds on average. There was overall a slight trend against agreeing with the IDA as the experiment went on. However, this was not a smooth decline but rather some waves of increased and decreased agreement. This pattern was similar for both conditions of the experiment, suggesting that customization did not lead to users learning about the IDA in any way that differed from the non-customization users. Because of this negative trend, the round number for each observation was included in all models of decision making presented below.

3.4.2 Efficacy Beliefs and Agreement

Efficacy Beliefs. To test the effect of customization on efficacy beliefs (H1), I fit a multilevel linear regression model to the data. This model estimates the efficacy beliefs score for each round of the game for each subject. It includes the condition the subject was in

Table 3.4: Efficacy beliefs model.

	Dependent variable:
	Efficacy Beliefs
Intercept	-1.444^{***} (0.225)
Customized	0.040 (0.093)
Category Ratings	0.367^{***} (0.051)
Trust Propensity	0.129^{**} (0.055)
Round Number	-0.020^{***} (0.005)
Random Effects Std. Deviation	0.649
Log Likelihood	−2,434.228
Log Likelihood χ^2	23.423^{***}
Note:	*p<0.1; **p<0.05; ***p<0.01

(customization or non-customization) as well as subjects' average category rating for the emphasized categories, their propensity for trust in automated decision aids, and the round number as covariates. Because there are multiple observations per subject, the multilevel model included a random effect that varied the intercept of the model for each individual subject. This random effect accounts for the lack of independence between observations from the same subject.

H1 was not confirmed, as there was not a statistically significant difference between customizers and non-customizers. The model did find that when the IDA used categories that the subject had rated highly, they had a higher belief of its efficacy. The model also found that people with a higher propensity to trust automated decision aids believed the IDA had higher efficacy than those with a lower propensity to trust automated decision aids.

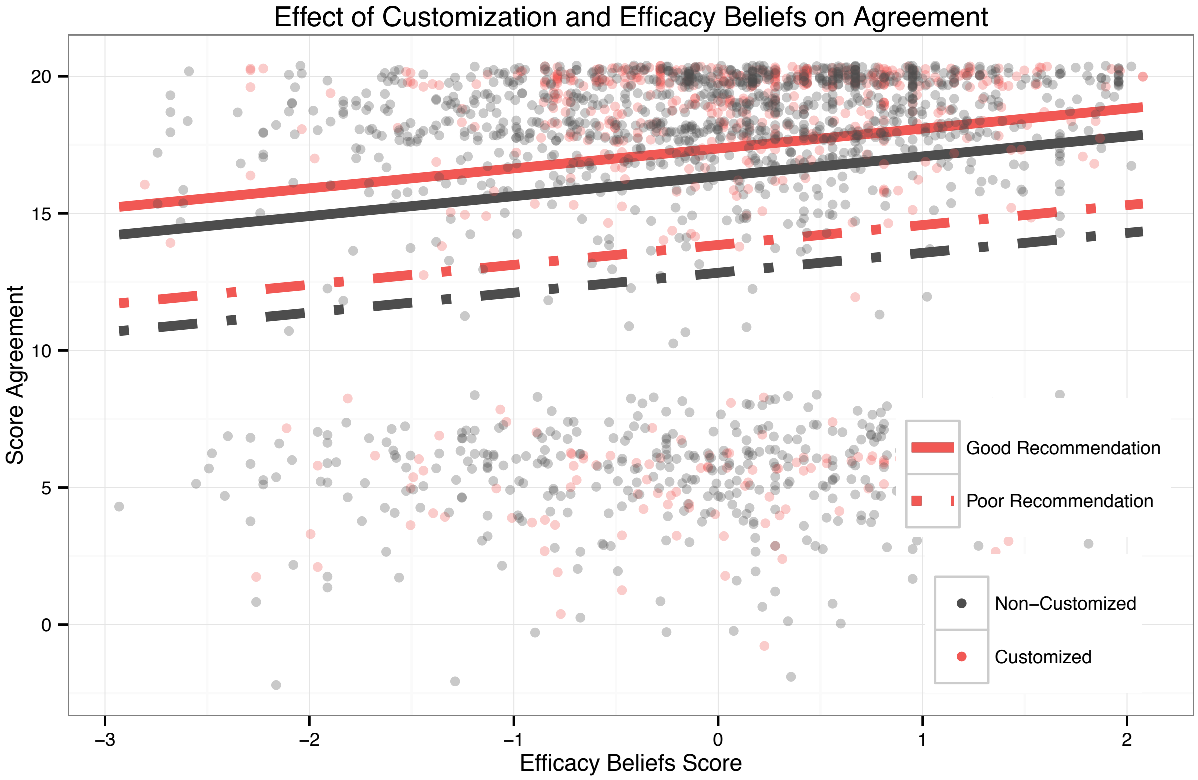


Figure 3.6: Agreement with recommended score.

This finding indicates that having control over an IDA's inner logic does not lead users to inflated expectations for how well the system will work. Subjects who customized the system largely configured it using categories that they believed would work well. Subjects who did not customize the IDA but were assigned configurations that closely matched their ratings of the categories reported the same beliefs in the system's efficacy as the people who had chosen the categories. This suggests that customization does not prompt users to irrationally believe that a system will work better simply because they have influenced it. Rather, users believe a system will work well when they have some understanding of how it works and they believe, based on their domain knowledge of the decision, that its logic is appropriate for the decision at hand.

Agreement. To evaluate the agreement variables, two multilevel models were fit to the data. A multilevel regression model estimated score agreement using the condition the

Table 3.5: Multilevel models of agreement with IDA recommendations.

	Dependent variable:	
	Score Agreement	Winner Agreement
Intercept	16.248^{***}	2.406^{***}
	(1.403)	(0.691)
Customized	1.014^{**}	0.507^{**}
	(0.495)	(0.235)
Poor Recommendation	-3.519^{***}	-1.633^{***}
	(0.235)	(0.126)
Category Ratings	0.114	-0.037
	(0.320)	(0.156)
Efficacy Beliefs	0.726^{***}	0.282^{***}
	(0.145)	(0.075)
Trust Propensity	0.243	0.097
	(0.291)	(0.132)
Round Number	-0.063^{*}	-0.039^{**}
	(0.032)	(0.018)
Random Effects Std. Deviation	2.511	0.499
Log Likelihood	$-6{,}209.821$	-950.162
Log Likelihood χ^2	245.75^{***}	206.27^{***}
Pseudo R^2		0.232
Note:		*p<0.1; **p<0.05; ***p<0.01

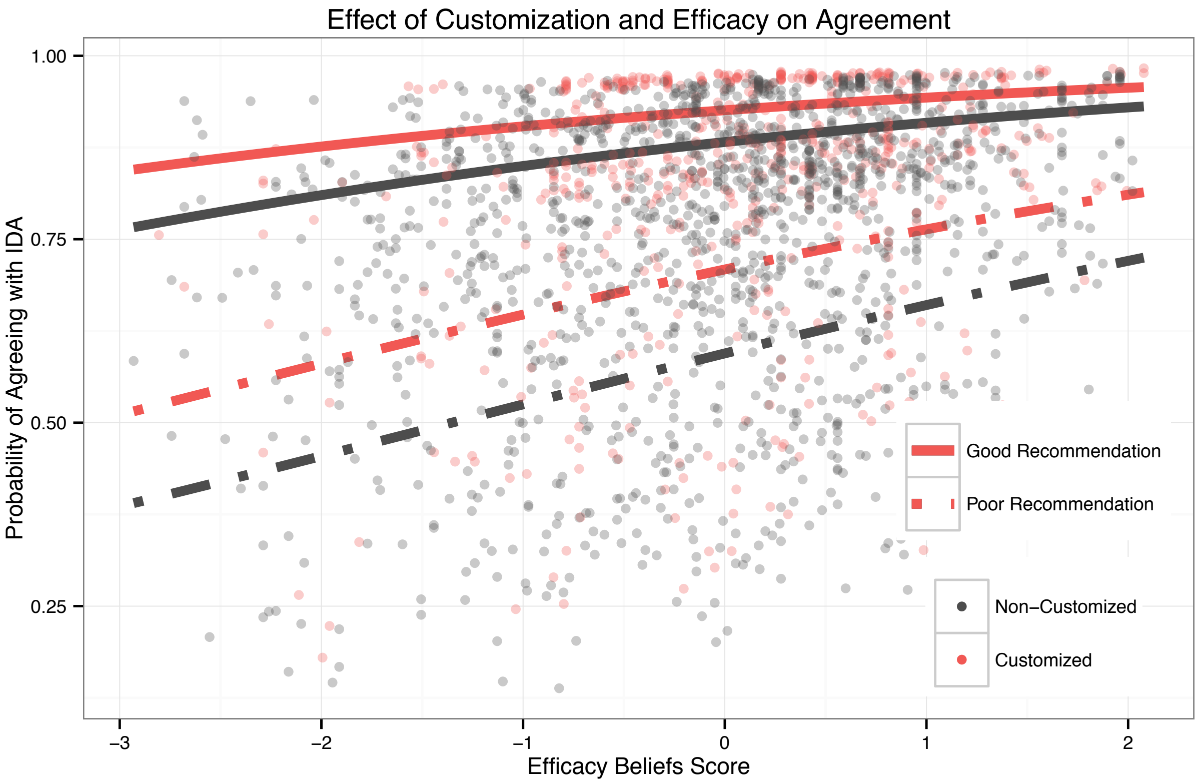


Figure 3.7: Probability of agreeing with IDA's predicted winner.

subject was in, the quality of the recommendation received, the subject's efficacy belief, their category ratings for the emphasized categories, their propensity to trust decision aids, and a random effect for each subject. To evaluate winner agreement, a multilevel logistic regression model was fit using the same independent variables. This model estimates the log odds of a subject choosing the same team to win the game as the IDA. These models are described in Table 3.5 and visualized in Figure 3.6 and Figure 3.7.

H2, the relationship between efficacy beliefs and agreement, is tested by the coefficients for efficacy beliefs in these models. In both models, there was a statistically significant relationship between subjects efficacy beliefs and their agreement with IDA recommendations. When subjects believed the system would work better prior to actually seeing its recommendations, subjects agreed more with the recommendations. This effect can be seen in Figure 3.6 and Figure 3.7. Note that for the model of winner agreement, the R^2 value listed in the table is a Pseudo R^2. This measure of fit was calculated using the method presented by Tjur (2009) that is based on the accuracy of the model at correctly predicting the observed values of either agree or disagree. It should not be interpreted as an indicator of the proportion of variance explained by the model. This Pseudo R^2 will be reported for all logistic regression models in this dissertation.

H3 is a test of the effect of customization on agreement, and was tested using the same models. The statistically significant coefficient for customization indicates the effect that the type of IDA had on agreement. H3 was supported, as subjects who customized the system overall did agree more with the IDA's recommendations than non-customizers.

H3a predicted that some of the effect of customization on agreement would be happen by customization influencing efficacy beliefs, and efficacy beliefs in turn influencing agreement. H3b predicted that customization would also influence agreement for other unobserved reasons. To evaluate H3a and H3b, I conducted a mediation analysis to estimate both the indirect effect of customization on agreement (by way of efficacy beliefs) and the direct effect. An estimate of the effect size of the indirect effect can be obtained by multiplying the coefficients for the effect from treatment to mediator by the coefficient from mediator to outcome (Imai, Keele, & Tingley, 2010). In this study, this means the coefficient for customization in the efficacy beliefs model (Table 3.4), and the coefficient for efficacy beliefs in the agreement models (Table 3.5). Imai et al. (2010) developed a procedure to perform hypothesis testing on this estimate by simulating the potential outcome for each observation. The potential outcome for an observation represents the efficacy beliefs and agreement that would have been observed were the given subject assigned to the opposite condition. Imai et al. showed that the distribution for the potential outcome of an observation can be taken from the observed data under the assumption that the treatment (i.e. customization) was randomly assigned and that there are no unobserved pre-treatment confounding variables. This second assumption is untestable (Imai et al., 2010) and therefore the mediation analysis of this experiment can provide only incomplete evidence of a mediated causal relationship between customization and agreement by way of efficacy beliefs.

Imai et al.'s method for hypothesis testing a mediated causal relationship involves a Monte Carlo sampling of simulated data based on the observed parameter estimates of the models to build a confidence interval about its size and direction. 1000 simulations of the data set are performed and in each simulation the Mediated Effect, the Direct Effect, and the Total Effect of the treatment on the outcome are measured. The average effects from all 1000 simulations are calculated along with a 95% confidence interval to test the hypotheses of mediated and direct effects of customization on agreement. The total effect that results from this analysis can be interpreted as the average change in agreement that subjects would experience had they been assigned to the opposite condition. The mediated effect represents the portion of the total effect that would be the result of a change in efficacy beliefs. The direct effect can be interpreted as the expected change in agreement that would be observed that is not due to a change in efficacy beliefs.

Table 3.6: Mediation analysis for score agreement.

	Estimate	95% CI Low	95% CI High	Sig.
Average Mediated Effect	0.026	-0.108	0.167	
Average Direct Effect	1.031	0.092	2.022	**
Total Effect	1.058	0.089	2.025	**

Table 3.7: Mediation analysis for winner agreement.

	Estimate	95% CI Low	95% CI High	Sig.
Average Mediated Effect	0.02	-0.005	0.009	
Average Direct Effect	0.068	0.008	0.125	**
Total Effect	0.070	0.009	0.129	**

Contrary to expectations, there was not support for H3a. The effect of customization on agreement was not mediated by subjects' efficacy beliefs using either measure of agreement (see Table 3.6 and Table 3.7). This finding provides more evidence for the conclusion

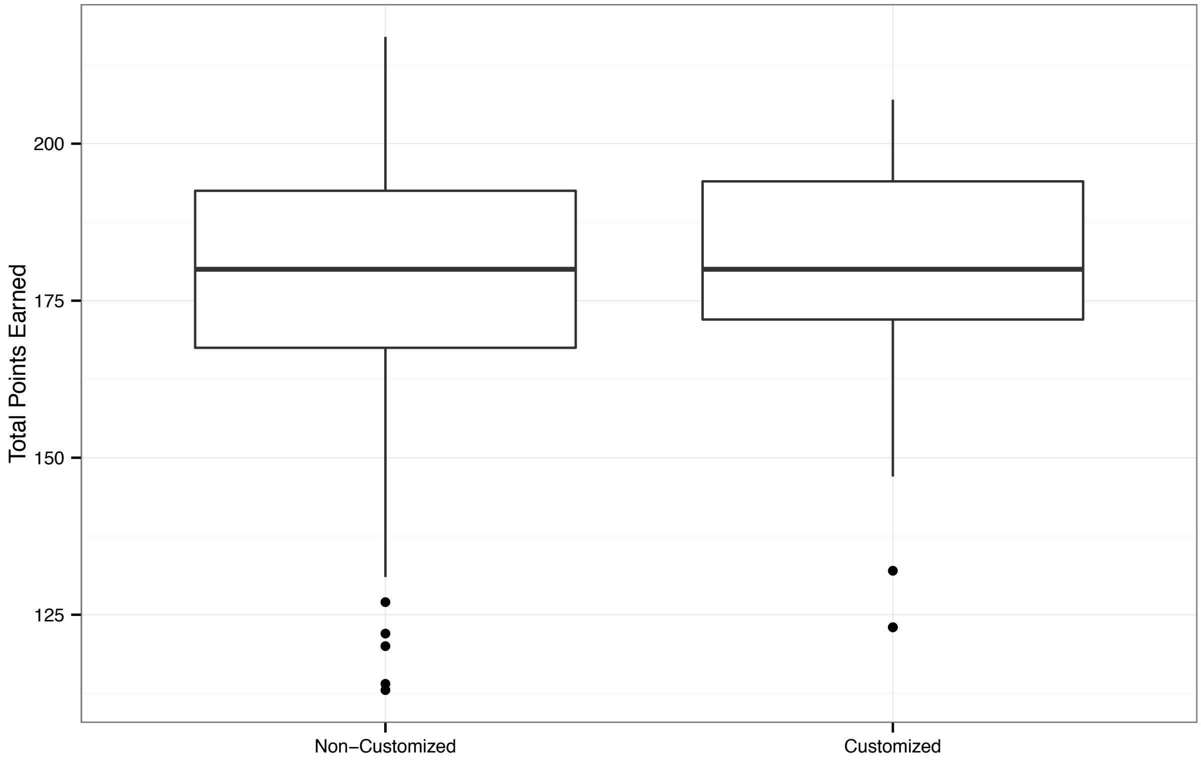


Figure 3.8: Total points earned by subjects in each condition over the 12 rounds.

discussed under H1, namely that customization does not create an irrational or inflated expectation of system efficacy that then leads users to agree with it more often. And while beliefs of high system efficacy do lead to more agreement, these beliefs are not enhanced purely by the act of customizing the system, but rather are determined by users' preferences for how a system can best work to generate recommendations.

The mediation analysis did find support for H3b in that there was a statistically significant ($p < .05$) direct effect of customization on treatment. Subjects who customize the system did agree more with its recommendations than those who did not, but this effect is unrelated to subjects' efficacy beliefs (Table 3.6 and Table 3.7).

Decision quality. H4 predicted that users who customize will make better decisions as determined by the number of points they earn in the game. Figure 3.8 shows the distributions of points earned by subjects in the game. A t-test indicated that there was not a statistically

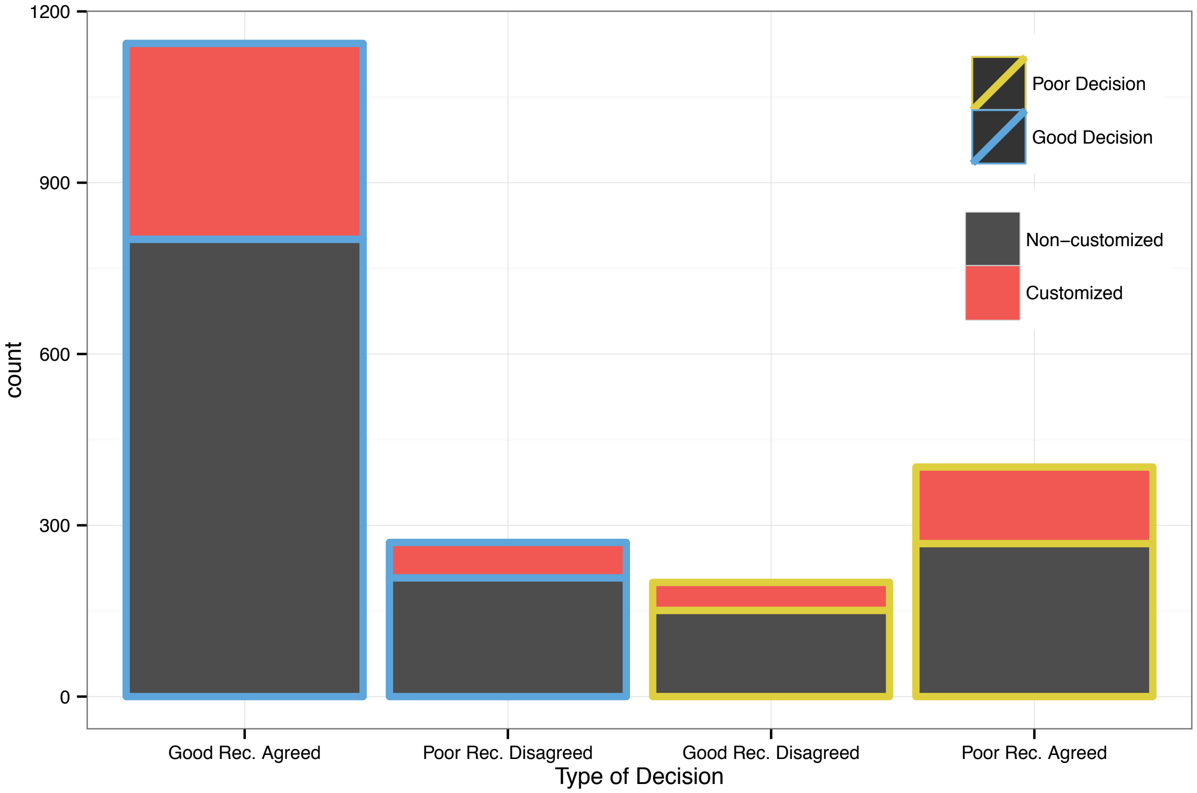


Figure 3.9: Decision quality by recommendation quality.

significant difference between the customization and non-customization groups in terms of how many points they earned ($t(166) = -0.649$, $p = 0.517$). Overall, customization did not lead to better decision making than a non-customizable IDA.

The nature of the decision task allows for different types of "poor" decision making relative to the recommendation that was provided. A poor decision could happen if the user was given a good recommendation but failed to follow it, or if he was given a poor recommendation and agreed with it. Conversely, good decision making can be considered to be when a user agrees with a good recommendation or disagrees with a poor recommendation. Figure 3.9 illustrates the different types of decisions that were made in the study, separated by the condition. Chi-square tests were performed to determine whether there were any differences in the types of decisions that subjects made based on their experiment condition. In terms of any type of good or poor decision, there was not a statistically significant difference between conditions. Subjects in either condition were equally likely to make good decisions. However, some small differences between the conditions were noted when examining the types of decisions. Subjects in the customize condition had a slightly higher proportion (9%) of decisions in which they agreed with a poor recommendation than non-customizers. And conversely, the non-customizers had a slightly higher proportion (9%) of decisions where they disagreed with a good recommendation. These differences were statistically significant ($p < .05$).

Table 3.8: Results of post-test survey.

	Dependent variable:			
	Control	Helpful	Accurate	Important
	(1)	(2)	(3)	(4)
Intercept	2.377***	4.679***	4.426***	4.253***
	(0.360)	(0.442)	(0.333)	(0.400)
Customized System	2.306***	0.536*	0.334	0.583**
	(0.230)	(0.282)	(0.213)	(0.255)
Age	−0.001	0.002	−0.003	0.009
	(0.010)	(0.013)	(0.010)	(0.012)
Gender (Female)	−0.169	0.350	0.159	0.744***
	(0.233)	(0.286)	(0.216)	(0.259)
Automation Trust Propensity	0.157	0.242	0.219*	0.319**
	(0.135)	(0.166)	(0.125)	(0.150)
R^2	0.390	0.052	0.042	0.129
Residual Std. Error (df = 163)	1.337	1.641	1.236	1.485
F Statistic (df = 4; 163)	26.081***	2.238*	1.781	6.047***
Note:				*p<0.1; **p<0.05; ***p<0.01

H5 predicted that users who customize the system would feel more control over it. H6 predicted that users who customize would report that the IDA was more helpful, accurate, and more important to their decision than users who did not customize. H5 and H6 were tested using linear regression models that estimated the response to the survey questions using the condition, demographic information, and automation trust propensity. These models are presented in Table 3.8. H5 was confirmed. The users who customize the system reported that they felt more in control than those who customized the system.

H6 was partially confirmed. Subjects who customized the IDA found the system to be more helpful and more important in their decision making. However, they did not find the IDA's recommendations to be more accurate than the non-customizers. These results support the existing work on customization (Sundar et al., 2012; Hijikata et al., 2012) in that customizable IDAs are better perceived by users than systems that do not have this affordance.

Also noteworthy is that users who had a higher propensity to trust automation reported that the system was more accurate and more important in their decision than those with lower trust in automation. This is an expected finding, but it should be noted as an important individual difference among users. The wide variability in trust in automation that was observed in this subject sample suggests that there may be underlying mental models and attitudes about the efficacy of IDAs that can impact how people use them to make decisions.

To test the hypothesis that greater consistency between configuration and recommendation would be associated with greater agreement (H7) I fit a multilevel logistic regression model similar to the logistic regression model reported in Table 3.5 . This model included the recommendation consistency variable to determine whether the consistency between the configuration and the recommendation would influence subjects' agreement with recommendations, while controlling for the other variables known to affect agreement reported above. The model also tested for an interaction between the condition and recommendation consistency to determine whether any effect of recommendation consistency was equivalent in both conditions of the study (H7a).

This model found that the more consistent a recommendation was with the configuration that had been used, the more likely a subject was to agree with the recommendation. This

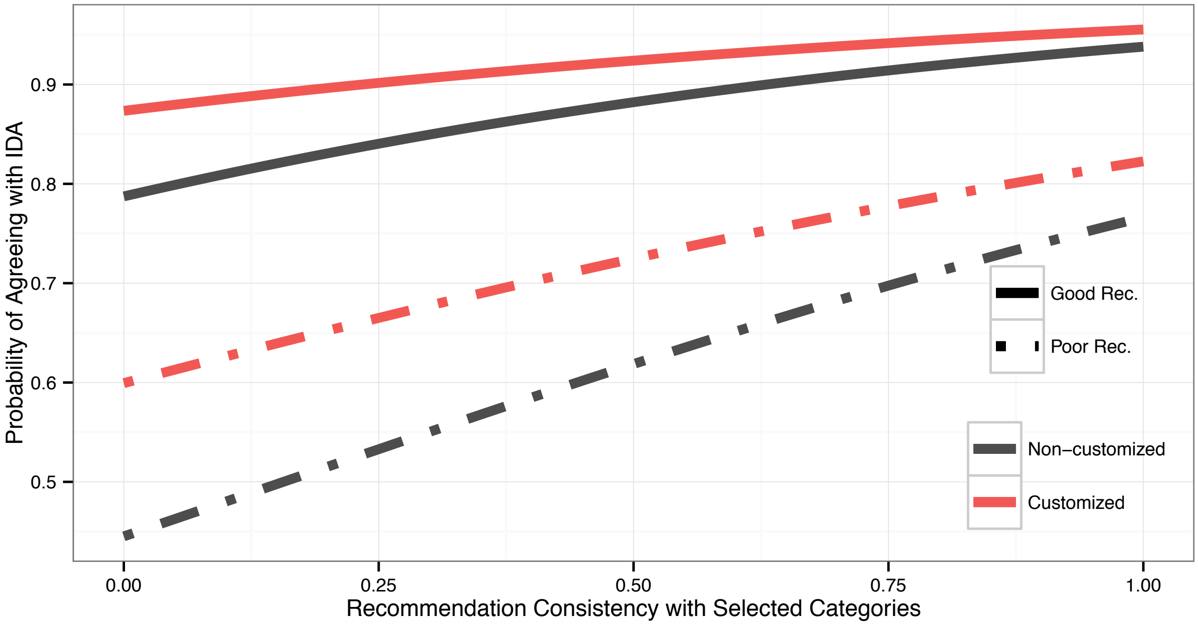


Figure 3.10: Greater consistency between recommendation and configuration led to more agreement.

result supports H7. However, H7a was not supported, as the interaction effect of customization and recommendation consistency was not statistically significant. It does not appear that this tendency to agree with recommendations that matched up well with the configuration that was used is any different when users customize an IDA than when they do not customize.

3.5 Discussion

These results suggest that customization bias as found in my previous work (Solomon, 2014) is not entirely the result of the personalized aspect of the system configuration. In this study where non-customizers were assigned configurations that were believed to match their preferences for how a system should work, subjects reported the same beliefs about the IDA's expected efficacy as users who had the opportunity to actually select the configuration. However, in spite of the non-customizers having the same expectations for the system's

Table 3.9: Recommendation/configuration consistency led to greater agreement.

	Dependent variable:
	Winner Agreement
Intercept	1.309***
	(0.175)
Customized	0.623**
	(0.314)
Recommendation Consistency	1.408***
	(0.236)
Poor Recommendation	$-$1.530***
	(0.128)
Efficacy Beliefs	0.297***
	(0.074)
Customized X Rec. Consistency	$-$0.277
	(0.469)
Random Effects Std. Deviation	1.065
Log Likelihood	$-$930.304
Log Likelihood χ^2	245.98***
Psuedo R^2	0.255
Note:	*p<0.1; **p<0.05; ***p<0.01

efficacy prior to using it, they were less likely to agree with its recommendations. This suggests that there is another cause of customization bias that has not been measured in this study.

This finding has important implications for IDA design. First of all, it highlights the important conceptual difference between personalization and customization, and shows that even if personalization can achieve the same intermediate outcome of finding something that highly matches users' preferences, it may still lead to different outcomes in terms of the decisions that users make in response to recommendations.

Table 3.10: Summary of results.

	Hypothesis	Result
H1	Subjects who customize the IDA will believe that the system has higher efficacy than those who do not customize the system.	Not supported
H2	Subjects' beliefs about the IDA's efficacy will be predictive of their agreement with its recommendations. Subjects will have more agreement when they believe the system has greater efficacy.	Supported
H3	Subjects who customize the IDA will have greater agreement with its recommendations than those who do not customize.	Supported
H3a	The effect of customization on agreement will be partially mediated by efficacy beliefs.	Not supported
H3b	There will be a direct effect of customization on agreement. Subjects who customize will agree more with the IDA for reasons other than an increased belief in its efficacy.	Supported
H4	Subjects who customize the IDA will make better decisions, earning them more points in the game, than those who do not customize.	Not supported
H5	Subjects who customize the IDA will report feeling more control over the system than subjects who do not customize.	Supported
H6	Subjects who customize the IDA will report that the system is a) more helpful to them as they make decisions, b) generates more accurate recommendations, and c) is a more important part of their decision-making process than subjects who did not customize.	Partially supported
H7	Consistency between the recommendations that the IDA gives and the configurations used will lead to greater agreement by subjects.	Supported
H7a	This effect will be stronger for subjects who have customized the IDA than those who did not customize.	Not supported

This is an important conclusion for system designers that relates to the function allocation problem. Systems that automate the information acquisition and information analysis stages of decision making will produce different decision making outcomes than systems that allocate more of those functions to the user, even if the automation performs the task equivalently. Similarly, system designers that try to select "default" settings and configurations that will influence a system's output must consider that even if users were to choose the same configurations on their own, they may interpret the system's output differently than the output generated by the default process.

The consequences of customization bias may be either positive or negative, and this may relate largely to the reliability of the system. In this study, there was no overall impact of customization on decision making quality. However, customizers were slightly more likely to make the poor decision of agreeing with a poor recommendation, and non-customizers made more errors where they disagreed with a good recommendation. The reliability of the system is an important factor for designers to consider when evaluating the potential consequences of customization bias. If a system will produce highly reliable recommendations, then customization will likely be beneficial to decision-making performance as users will be nudged towards agreement. A system that produces unreliable recommendations that require considerable discernment from users would find customization bias to be problematic towards decision making.

This study also provides an evaluation of a different type of system design that has not been examined in the existing literature. Customization of an IDA's algorithm is an emerging area of research in IDA, and this study demonstrates a potential problem that end-user customization can create. However, it also shows that this bias may be countered by using a crowdsourcing or collaborative filtering approach to tailoring the IDA's algorithm that removes the act of customizing from the user. Other users making similar decisions may be leveraged to provide the input to the system, leaving the user and decision maker to interpret the output. Future research should explore this approach further to determine how it may counteract customization bias. A particularly important focus for future research is whether users believe crowdsourcing can be effective in general. Subjects in this study reported that the configurations generated by other users were just as efficacious as those that had bee customized, but then they proceeded to disagree with them more often. Users' mental models of crowdsourcing and its capabilities may largely determine whether they will be receptive to this design. It should be noted that subjects in the non-customizable condition knew that the configurations had been chosen by other users, but did not know that they had been matched to the most similar other user. Whether or not this knowledge would affect users' beliefs of the system's efficacy and agreement with recommendations would make an important contribution in future research.

This study has an additional finding that has important implications for IDA design. When the IDA's recommendations appeared to be consistent with the way it had been configured, subjects were more likely to agree with the system. This effect was the same for customizers and non-customizers. Variation in the consistency between configuration and recommendation was largely a function of random chance. Some categories were always emphasized, and one of the two teams always was stronger in each category, so it always had to appear that the system had been consistent with the configuration to some degree. Nevertheless, when by chance the system appeared consistent, it led users to agreement.

This finding is evidence of a consistency bias where IDA users are more inclined to agree with recommendations when they feel that the recommendations are consistent with the way the IDA is configured. This is a bias because it happened regardless of recommendation quality and because all users received the exact same recommendations. This finding replicates the finding from my earlier work (Solomon, 2014), where I posited that it could be the result of confirmation bias where users made an initial decision in their minds, chose categories that would be consistent with that initial decision, and then were assured it was correct after seeing recommendations that were consistent with the initial decision and configuration. The findings from this study refute that interpretation and suggest that this consistency bias is not an example of confirmation bias. This is because the consistency bias was observed in both conditions of this study, with no statistically significant difference in the size of the relationship. Subjects in the non-customization condition had no opportunity to choose categories and so could not choose categories that were consistent with any initial decision they may have made.

Rather than confirmation bias, this bias may be a conflation of algorithm success with outcome success. When it appeared that the algorithm was able to successfully arrive at a solution that matched its input, subjects may have interpreted this as an indicator of recommendation quality. It is possible that users have a mental model of the IDA algorithm that permits it to find solutions that do not match its input, but when it does find a solution it is an indicator that the solution is reliable.

An important consideration for understanding this bias is that it depends on the transparency of the system. If users know nothing about the configuration of the system, than it is not possible for them to assess whether the recommendations are consistent with its configuration. Consistency bias is therefore highly related to the transparency of the system. One interpretation of this bias may be that users are more inclined to agree with recommendations when they understand how it works, but when the system is not consistent with itself, it is confusing to users who then question whether the system works how they think. Therefore, it is possible that consistency bias is a bias of the transparency of the system. In chapter 5, I will discuss transparency bias and its relationship to consistency bias further.

Another consideration for understanding this bias is that by the design of the study, users generally had high expectations that the system would work well. Therefore, it is possible that this bias would not apply in a situation where users expected the system to perform poorly. If the system was using a configuration that subjects expected to produce poor recommendations, and the system produced poor recommendations, the system would be consistent but it would seem unlikely that users would agree with it. However, this consistency might lead to users having a generally higher expectation of the system because it demonstrates that its logic works as intended, which might lead to greater agreement in the future.

The study was not designed to identify the mechanism behind this bias, but this is an area for future work. In particular, the idea that IDA users may be misled if they feel that a system has worked "as-intended" has design implications. Designing transparent systems that allow users to see the inner logic and understand it may lead to such a bias if users conflate a system working "as-intended" with working effectively to produce good recommendations.

Overall, this study demonstrates that customization bias presents a human factors concern for the design of interactive IDAs. Users may become partial to agreeing with recommendations purely because of their involvement in producing them, regardless of the actual quality of recommendations or even their expectations that the system works well. This creates the potential for decision-making errors and biases that may limit IDA effectiveness. Nevertheless, an important takeaway from this study is that user beliefs about system efficacy are related to decision making. What users believe about how well the system works prior to using it influences how they interpret the output. This issue will be further examined in the next chapter.

3.5.1 Limitations

A major limitation of this study is that the IDA used was only a shell and did not actually use any intelligent technology to produce recommendations. The configurations that were purportedly being used by the system had no effect on the recommendations, and therefore this system only mimics the functionality of an IDA. The manipulation checks used in the study suggested that users were not aware the system was fake, but it nonetheless may have provided a different experience than a true IDA that uses actual intelligent technologies to produce recommendations and, if customizable, is responsive to users. This limitation is addressed in the study reported in chapter 5 which has a similar experiment design but uses an IDA that is actually responsive to its configuration.

Another important limitation for this study, and for all three studies reported in this dissertation, is that there may be important differences between the way people make decisions in a lab setting and how IDA-supported decisions are made in the real world. Some human factors scholars have argued that controlled lab experiments such as this must be complimented by field studies in "naturalistic" settings (Klein, 2008) where people have real pressure and consequences for their decisions. Klein argues that studying decision making through observational field studies allows human factors researchers to better understand cognitive and decision-making processes as they actually happen in real-world and that this provides a useful basis for system design. In this study, I offered incentives to subjects to make good decisions. And since the subjects were recruited from a workforce who in many cases use participation in studies as a means of income (Ross, Irani, Silberman, Zaldivar, & Tomlinson, 2010), the decision task has some "naturalistic" validity. However, there are undoubtedly differences between an unfamiliar baseball decision task worth a few extra dollars and high pressure decisions like those in a hospital or on a battlefield, and these differences may strongly affect how people use and trust an IDA. Furthermore, the design of the study tried to limit the IDA as the *only* source of information about the decision task, when in more natural settings, an IDA may be one of many information tools being used by a decision maker. I cannot know whether the effect I have reported here would be applicable in situations where users have a variety of tools and information resources.

Another limitation is that this study focused on novice users using a tool for the first time, and these users were limited in their ability to learn and adjust their decision making over time because they were not shown their scores until after all rounds had been played. While this provides some internal validity to the study, it does so at the expense of some external validity since in the real world, users can evaluate their own decision performance over time and use that to influence their future decisions and future interactions with IDAs. It is not clear from this study whether these results would apply to more experienced users, or how these biases might change over time. This is important because many IDAs in real-world settings are used primarily by people with high expertise and a lot of experience. It is possible that these biases only affect novice users and that with experience, users adjust their behavior to optimize decisions. However, it is also possible that these biases become stronger, as experts begin to use more shortcuts or heuristic decision making. Additional research is needed to evaluate how these biases affect more experienced IDA users.

CHAPTER 4

CAUSAL EFFECT OF EFFICACY BELIEFS AND EXPECTATIONS BIAS

Because IDAs are used for difficult decisions with high uncertainty, and because the systems themselves are often not highly transparent, users may have difficulties forming well-calibrated beliefs about a system's efficacy. Users' perceptions of an IDA's reliability may not match its actual reliability, and this can be a source of automation bias. In the study presented in the previous chapter, there was a relationship between users' efficacy beliefs and their agreement with recommendations. When users expected that a system would work well, based at least partially on how it was configured, they were more likely to agree with recommendations regardless of their quality.

The evidence from that study however can only suggest an association between efficacy beliefs and agreement with recommendations. Efficacy beliefs were not randomly assigned, and the only randomly assigned variable in that study (customizability of the IDA) did not have an effect on efficacy beliefs. Although efficacy beliefs were measured before agreement, it cannot be ruled out that agreement actually caused efficacy beliefs. It is possible that subjects decided whether they would agree with recommendations before they actually saw them and this decision influenced their efficacy beliefs. Or perhaps more likely, the relationship between efficacy beliefs and agreement is spurious, with some unobserved third variable causing users to both have high beliefs of efficacy and to agree with recommendations. The design of the previous study, and its analysis which suggests a causal relationship, assume no unobserved variables that cause a spurious relationship, but this assumption cannot be tested under the design of that experiment. For that reason, in this chapter I present a study that is designed to estimate the causal effect of efficacy beliefs on agreement with IDA recommendations.

Establishing that efficacy beliefs cause agreement makes an important contribution to

IDA research. If agreement with recommendations, and consequently users' calibration of trust and decision making, can be influenced by their efficacy beliefs, than efficacy beliefs present a target for system designers to engineer users' decision making. Finding ways to give users clear and precise expectations about how well a system works and what quality of expectations can be expected can lead to better calibrated trust in IDAs and better decisions by their users.

If efficacy beliefs prior to seeing recommendations cause users to agree with recommendations regardless of their quality, it represents a bias that is caused by users' expectations of the IDA's efficacy. But because expectations were merely observed and not randomly assigned, we cannot know whether the bias is truly a bias caused by expectations or if that relationship is a by-product of some other unobserved bias. In this chapter I have designed a study to identify whether expectations cause a bias in agreement. This study provides evidence that the expectations users have about an IDA's efficacy bias their subsequent agreement with its recommendations.

4.1 Methods

To study whether users' beliefs of system efficacy cause them to agree with recommendations, I conducted an experiment using a randomized encouragement design. Under this design, subjects were encouraged by a randomized treatment variable which is described below to have either high or low efficacy beliefs. They then used an IDA to get recommendations and made their decision. This design uses the encouragement variable as an instrumental variable in order to estimate the causal effect of efficacy beliefs on agreement. An instrumental variable is a variable which has no correlation with the dependent variable other than through its correlation with the explanatory variable (i.e. efficacy beliefs). Through this method, variation in the explanatory variable which is caused by the instrument (which has been randomly assigned in this case) can be considered to be random variation with regards to the dependent variable. Therefore, correlation between the random variation in the explanatory variable and the dependent variable can be interpreted as a causal relationship between the explanatory and dependent variables (Angrist & Pischke, 2008).

In this study, the same fantasy baseball task was used to implement this randomized encouragement design as was used in the customization bias study. In this study, only the non-customizable version of the IDA was used. This was done to prevent customization bias from creating unwanted noise in the data set. As in the previous study, the IDA provided 8 good recommendations and 4 poor recommendations, scoring an average of 15 points. The games used in the task, the screening quiz, the pre-test survey and post-test survey, and the scoring procedure and incentive were the same as in the customization bias study.

Encouragement of efficacy beliefs. Subjects were assigned to one of two conditions. In the *high efficacy* condition, subjects were told that the system's average performance was 18 points. They were also told that the data about Major League Baseball used by the IDA was comprehensive and contained no known errors. In the *low efficacy* condition, subjects were told that the system's recommendations score 12 points on average. They were also told that its data set contained errors and omissions. This information was presented to subjects in the instructions, but the interface of the system also contained a reminder of this information. In order to proceed to the baseball prediction game, as in the previous study, subjects had to pass a quiz on the instructions that verified whether they understood the average quality of the system's recommendations.

These differences in the instructions were intended to encourage subjects towards a believing the IDA had high or low efficacy in producing accurate recommendations about the outcomes of baseball games. In addition to these differences in instructions, there was also a difference in the configuration that was purportedly used by the system. As in the customization bias study, subjects rated each of the 27 categories according to how well it would inform a computer in predicting the outcome of baseball games. Using the ratings obtained about these categories in pilot testing, I created configurations that used categories that were either highly rated or poorly rated on average by subjects. I created 8 configurations, with 4 using mostly poorly rated categories and the other 4 using only highly rated categories. In the high efficacy condition, the IDA was presented as using one of the configurations with highly rated categories, and in the low efficacy condition the IDA used only configurations that were poorly rated.

The goal of these conditions was to randomly set subjects' beliefs of system efficacy to either a high or low level. I used three different approaches to encourage these efficacy beliefs (expected performance, data quality, configuration quality) in order to create a strong and effective instrument. A disadvantage of simultaneously using three approaches is that the effect of any of the three cannot be identified as they are confounded with the condition assignment. However, as the primary purpose of this study is to estimate a causal effect, a strong instrument is necessary. Pilot tests that used only the data quality and expected performance approaches produced only a moderate difference in efficacy beliefs and thus may have resulted in a weak instrument.

A total of 93 subjects were recruited from Amazon Mechanical Turk. Subjects were paid $2 for participation and an average of $2.25 in bonus payments for decision performance. Subjects took the same baseball knowledge screening quiz as in the customization bias study in Chapter 2. 31 subjects failed this quiz and were screened out of further participation. The final data set included 62 subjects, with 32 in the high efficacy condition.

This study has three hypotheses.

- **H1.** Encouragement of high or low efficacy beliefs will cause subjects to have high or low efficacy beliefs.

- **H2.** Efficacy beliefs will be associated with agreement.

- **H3.** Efficacy beliefs will cause agreement as assessed by two-stage least squares regression.

4.2 Results

Table 4.1: Descriptive statistics and factor analysis of dependent variables.

Statistic	N	Mean	St. Dev.	Factor Loading	Std. Err.
Likert measure	744	4.824	1.395	1.000	-
Expected score	744	13.743	4.210	3.013	0.121
Expected probability of correct winner	744	75.882	14.961	11.253	0.435
Agreement	744	15.145	8.967		
Winner Agreement	744	0.794	0.404		

Factor Analysis Fit Indices	
Comparative Fit Index	1.000
Root Mean Square Error	0.000

4.2.1 Descriptive results

A confirmatory factor analysis was performed to obtain factor scores for the efficacy beliefs variable using the three different measures. The factor loadings are reported in Table 4.1.

Figure 4.1 and Figure 4.2 show subjects' efficacy beliefs and winner agreement over the course of the 12 rounds of the experiment. I fit multilevel regression models to the data to determine whether there was any trend in how users made decisions or their efficacy beliefs as a result of repeatedly playing the baseball prediction game. These models found no statistically significant trend. Subjects did not adjust their agreement with recommendations or their beliefs of its efficacy as a linear function of how many rounds of the game had been played.

4.2.2 Manipulation check

To determine whether the encouragement had the intended effect on subjects' beliefs about the IDA's efficacy, I fit a multilevel regression model that estimated the efficacy beliefs factor score with the encouragement variable, propensity to trust decision aids, and a random effect

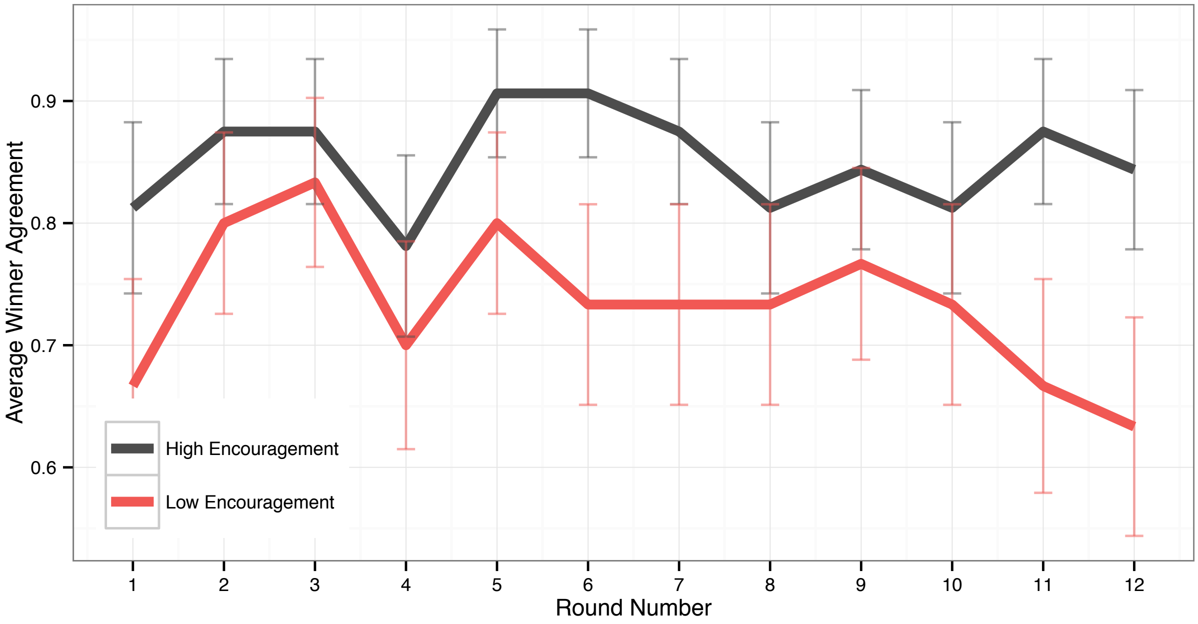


Figure 4.1: Winner agreement over the course of the 12 rounds.

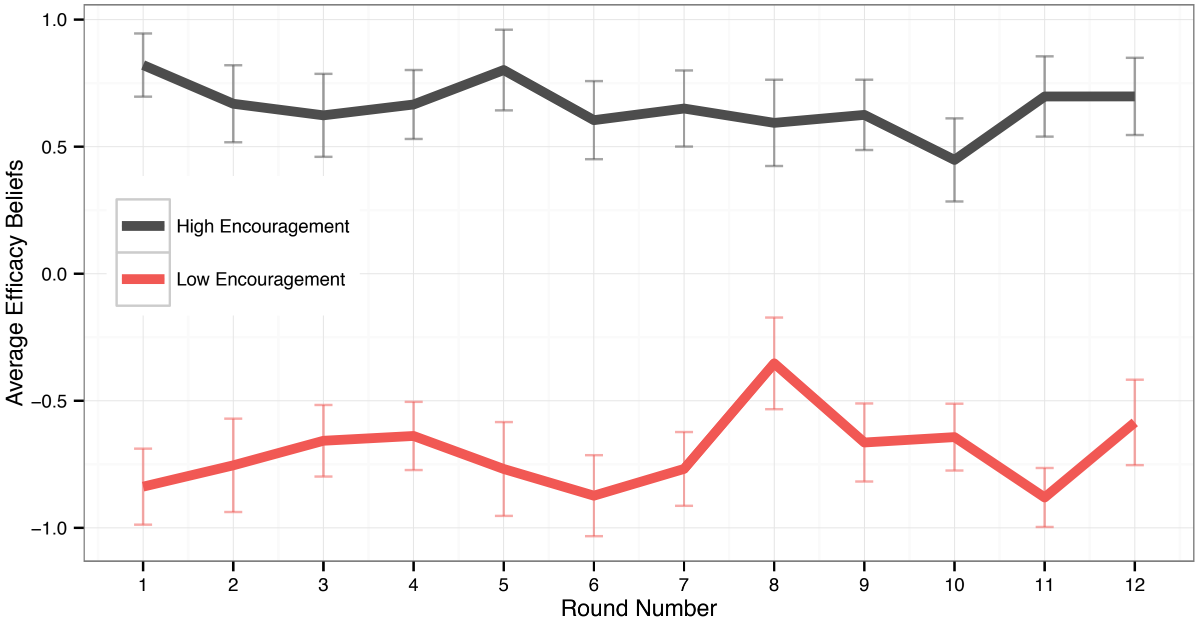


Figure 4.2: Efficacy Beliefs over the course of the 12 rounds.

Table 4.2: Efficacy variables by experiment condition.

Statistic	High Encouragement		Low Encouragement	
	Mean	St. Dev.	Mean	St. Dev.
Likert	5.505	1.067	4.097	1.337
Score	16.464	3.340	10.842	2.903
Probability	83.529	12.638	67.725	12.791
Efficacy Beliefs Factor	.658	0.853	-.702	.850

Table 4.3: Effect of encouragement on efficacy beliefs.

	Dependent variable:			
	Efficacy Factor	Likert	Score	Probability
Intercept	-0.662^{***}	4.149^{***}	10.923^{***}	68.337^{***}
	(0.106)	(0.136)	(0.423)	(1.707)
Encouraged	1.253^{***}	1.272^{***}	5.405^{***}	14.179^{***}
	(0.149)	(0.192)	(0.598)	(2.412)
Trust Propensity	0.362^{***}	0.461^{***}	0.735^{*}	5.499^{***}
	(0.102)	(0.131)	(0.409)	(1.651)
Random Effects Std. Dev.	0.547	0.686	2.215	8.981
Log Likelihood	−763.979	−1,078.090	−1,723.692	−2,710.091
Log Likelihood χ^2	61.176^{***}	47.885^{***}	59.624^{***}	41.291^{***}
Note:				*p<0.1; **p<0.05; ***p<0.01

for each subject as regressors. I fit three additional models using the separate efficacy beliefs variables (Likert Scale, Expected Score, and Expected Probability of Correct Winner). These models are detailed in Table 4.3. All four models offer the same conclusion that subjects in the High encouragement condition had higher beliefs about the system's efficacy than subjects in the Low encouragement condition. Figure 4.3 illustrates the distributions of efficacy beliefs in each condition. From these analyses it is clear that the encouragement variable was successful in its intended effect to manipulate subjects' beliefs about the system's efficacy.

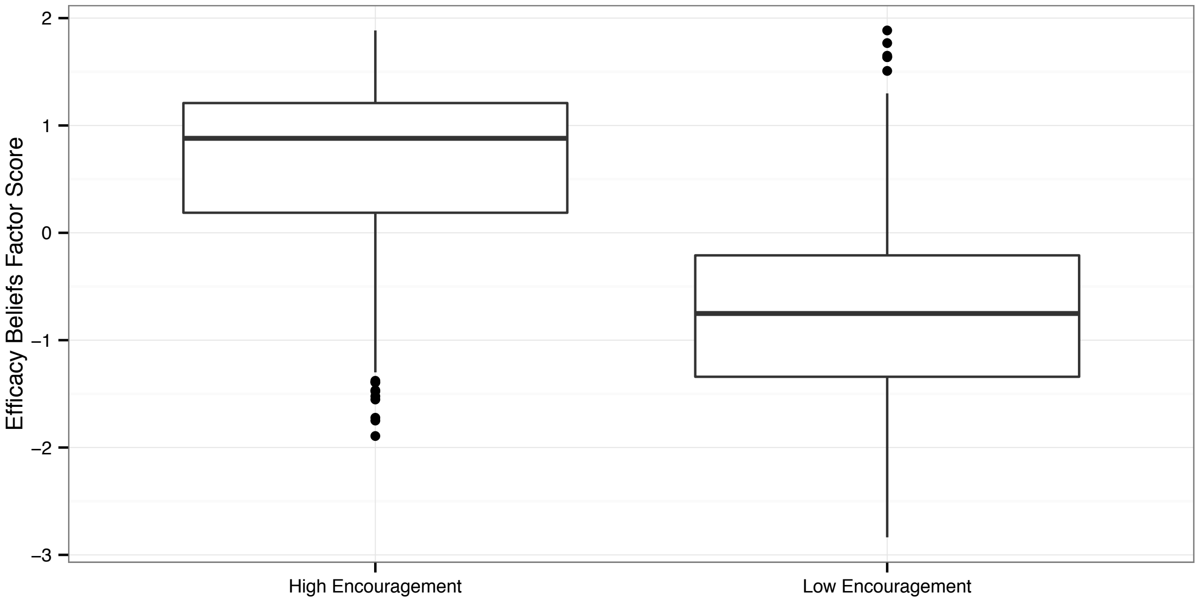


Figure 4.3: Distributions of efficacy beliefs by encouragement condition.

Table 4.4: Effect of efficacy beliefs on agreement.

	Dependent variable:	
	Agreement	Winner Agreement
Intercept	16.591***	2.308***
	(0.775)	(0.221)
Efficacy Beliefs	1.315***	0.321**
	(0.380)	(0.128)
Poor Recommendation	−4.226***	−1.554***
	(0.519)	(0.213)
Trust Propensity	−0.898	0.297
	(1.053)	(0.238)
Random Effects Std. Dev.	5.622	1.029
Log Likelihood	−2,533.185	−329.493
Log Likelihood χ^2	72.853***	66.118***
Pseudo R^2		0.225
Note:		*p<0.1; **p<0.05; ***p<0.01

4.2.3 Estimate of effect of efficacy beliefs on agreement

Table 4.4 describes models that estimate agreement with recommendations. The models show a statistically significant relationship between efficacy beliefs and agreement for both measures of agreement. When subjects believed that the system would work well prior to receiving recommendations, they agreed with those recommendations more regardless of their quality. Another noteworthy result from these models is that the individual difference of generally having trust in automated decision aids did not have a statistically significant relationship with agreement when controlling for efficacy beliefs. This suggests that the individual variation in underlying trust is well captured by the efficacy beliefs measures. When users trust decision aids, they also expect them to perform well but base their decisions more on the expectation of system efficacy than on their underlying trust of decision aids in general.

Table 4.5: Two-stage least squares estimates of causal effect of efficacy beliefs on agreement. Efficacy beliefs are instrumented by encouragement in this model.

	Dependent variable:	
	Agreement	Winner Agreement
	(1)	(2)
Intercept	16.578***	0.870***
	(0.487)	(0.017)
Efficacy Beliefs	2.77***	0.087***
	(1.210)	(0.032)
Poor Recommendation	$-$4.298***	$-$.228***
	(1.233)	(0.034)
Random Effects Std. Dev.	5.879	0.135
R^2	0.053	0.084
Wald χ^2	13.09***	53.46***
Note:	*p<0.1; **p<0.05; ***p<0.01	

4.2.4 Causal effect of efficacy beliefs on agreement

Encouragement designs such as this one can allow for causal inference about the actual treatment, in this case efficacy beliefs, even though only the encouragement variable has been reliably randomized. The encouragement design uses the randomized encouragement variable as an instrumental variable that creates some random variance in efficacy beliefs. The random variance in efficacy beliefs that is caused by the encouragement variable can be used to estimate the causal effect of efficacy beliefs on agreement (Angrist & Pischke, 2008).

I used two-stage least squares as the method for estimating the causal effect of efficacy beliefs with the instrumental variables method. Two-stage least squares performs two regressions. In the first stage, the explanatory variable (efficacy beliefs) is regressed on the instrument (encouragement), as well as the recommendation quality variable as an exogenous covariate. In the second stage, agreement is regressed on efficacy beliefs and recommendation quality, with the fitted values from the first stage replacing the observed values.

Because of the repeated measures of the study, the data were treated as panel data for the purposes of two-stage least squares estimation. Since the encouragement variable was assigned at the subject level, rather than at the round level, random effects for each subject were estimated rather than fixed effects in each stage of the two-stage least squares. Fixed effects cannot be estimated because these effects are colinear with the encouragement variable. The model was run using Stata's *xtivreg* command which performs a generalized two-stage least squares regression that includes the random effects. Robust standard errors were calculated using a bootstrapping method that randomly resampled the data 1000 times to re-estimate the standard errors. Bootstrapped standard errors were used because Stata's *xtivreg* command assumes constant variance if regular standard errors are used. The random effects in this model necessitate an additional assumption, which is that any individual-level effects are not correlated with any of the instruments or covariates in the model (Mundlak, 1978). This assumption can reasonably be made for these data because both the encourage-

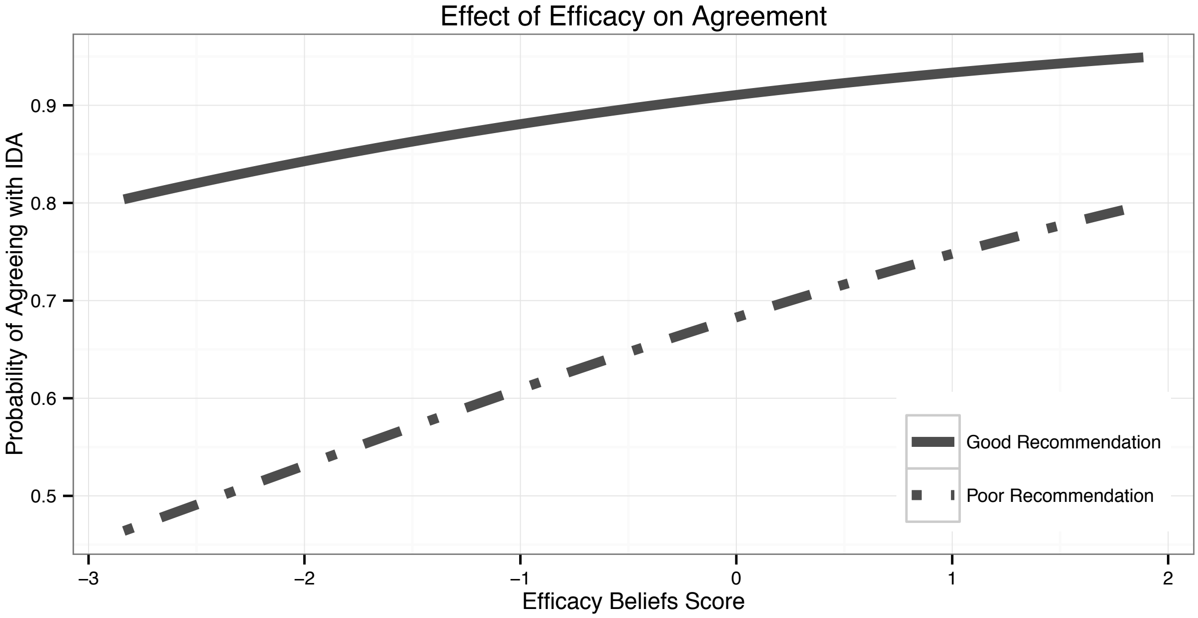


Figure 4.4: Association between efficacy beliefs and agreement.

ment variable and recommendation quality were randomly assigned.

A separate model was fit for each version of agreement. For the model with the binary Winner Agreement dependent variable, I followed Angrist and Pischke's (2008) suggestion and used a Linear Probability Model that treats the binary variable as continuous rather than a non-linear transformation such as the logistic regressions used in other parts of this dissertation. Angrist and Pischke argue that these non-linear models create additional complexity when used with instrumental variables and that this additional complexity is often not justified because Linear Probability Models perform well at estimating marginal effects, even if their predicted values are imprecise.

The results of the two-stage least squares are reported in Table 4.5. For both measures of agreement, the analysis showed statistically significant effects of efficacy beliefs on agreement. These estimates suggest that believing an IDA has high efficacy prior to seeing its recommendations causes users to agree with those recommendations, even when controlling for the quality of the recommendation. This effect is illustrated in Figure 4.4

Two-stage least squares has two primary requirements in order to be able to interpret a causal effect. The first requirement is that the instrument is strongly correlated with the explanatory variable. The analyses presented in the previous section offer evidence of a strong effect of the encouragement on efficacy beliefs. In addition, Stock and Yogo (2005) developed criteria for identifying weak instruments. Under their criteria, a model with one endogenous explanatory variable and one instrument should have an F-statistic of greater than 16.38 in the first stage regression in order to be assured of only a minimal amount of bias due to a weak instrument. The F-statistic from the first stage regressions (which is the same regression for both models) is 75.08. Therefore I can conclude that the instrument is strong, satisfying the first requirement.

The second requirement is that the instruments have no effect on the outcome variable other than through the explanatory variable. This can be tested when a model is overidentified, meaning that there are more instruments than explanatory variables (Sargan, 1958). However, this model is exactly-identified with one instrument and one explanatory variable, and therefore this assumption cannot be explicitly tested. For this reason, the interpretation of the two-stage least squares model of a causal effect of efficacy on agreement is valid only inasmuch as this assumption is valid. While this is a limitation that should be considered in the interpretation of these results, I argue that this assumption is reasonable in this circumstance. One of the encouragement variable's manipulations was a system-estimation of its own efficacy that was merely communicated to users. The other two manipulations were variations of how the system purportedly worked to produce recommendations. It is difficult to conceive of any way these randomly assigned manipulations could affect agreement other than by affecting subjects' beliefs about the system's efficacy. For this reason, the most likely interpretation of the relationship that has been found between efficacy beliefs and agreement is that efficacy beliefs cause some agreement with IDA recommendations

4.2.5 Decision Making

Table 4.6: Decision-making quality.

	Dependent variable:
	Number of Points Earned
Intercept	16.572^{***}
	(0.327)
Efficacy Beliefs	0.608^{**}
	(0.240)
Trust Propensity	-0.730^{*}
	(0.428)
Poor Recommendation	-6.199^{***}
	(0.400)
Random Effects Std. Dev.	1.816
Log Likelihood	−2,300.606
Log Likelihood χ^2	211.51^{***}
Note:	*p<0.1; **p<0.05; ***p<0.01

I examined how subjects' efficacy beliefs impacted their decision making quality by fitting a multilevel model that estimates the number of points earned from a round of the game with the efficacy beliefs factor score, recommendation quality, and trust propensity as regressors along with random effects for each subject. This model is described in Table 4.6.

Subjects made better decisions when they had higher beliefs about the IDA's efficacy. This can likely be explained by the fact that high efficacy beliefs caused agreement with recommendations, and since there were more good recommendations than poor ones, and since good recommendations suggested the best possible decision that could be made, subjects who believed the system would work well tended to follow those good recommendations. Although this effect is statistically significant, it is not a particularly large effect. The difference between the lowest efficacy beliefs score and the highest is only about 3 points, which

Table 4.7: Summary of results.

	Hypothesis	Result
H1	Encouragement of high or low efficacy beliefs will cause subjects to have high or low efficacy beliefs.	Supported
H2	Efficacy beliefs will be associated with agreement.	Supported
H3	Efficacy beliefs will cause agreement as assessed by two-stage least squares regression.	Supported

is fairly small given the overall variance on decision quality.

Another noteworthy finding is that subjects with higher propensity to trust decision aids actually made worse decisions than those with lower levels of trust. This finding may illustrate an individual difference in automation bias. People who have higher propensity to trust decision aids may not have scrutinized recommendations carefully, causing a particularly high rate of error where they agreed with a poor recommendation.

4.3 Discussion

This study has demonstrated that the beliefs IDA users have about the efficacy of a system prior to seeing its output has an effect on their decision making after they see its recommendations. When users expect the system to produce good recommendations, they interpret the recommendations the system produces as being more trustworthy. In this study, all subjects saw identical recommendations but were encouraged to have varying beliefs about the efficacy of the system, and these beliefs changed the level of agreement with the same recommendations.

Because recommendations were identical, this finding that efficacy beliefs influenced agreement is evidence of another decision making bias. Efficacy beliefs were measured prior to subjects seeing the recommendations, making them expectations about the system that were formed based on subject's knowledge of the system and of the statistics about the teams playing in the game. The greater agreement by those with greater efficacy beliefs is an *expectations bias.*

An important implication for the design of IDA from this finding is that when making decisions, evaluating the efficacy of the system is a part of the decision making process. An IDA's recommendations are information that must be evaluated, and beliefs about the process that produces recommendations can lead people to interpret the same information in different ways. This has relevance to the debate over transparency in IDAs. Transparent IDA's are well understood to provide a better user experience (Herlocker et al., 2000), however their effect on decision making is less conclusive, with some evidence suggesting a negative impact (Ehrlich et al., 2011). This study offers some evidence for why transparency has not demonstrated a clear impact on decision making performance. If transparent systems provide information about how a system works, this information must be evaluated by users. Evaluating the system's inner logic and forming an expectation about its efficacy may distract users from evaluating recommendations. Transparency may therefore present somewhat of a paradox because users, who are seeking advice from an automated system because they face a difficult decision with high uncertainty, may have high uncertainty about what constitutes an effective process for generating recommendations. In many cases users are likely better qualified to evaluate recommendations directly (without knowledge of any process that produced them) than to evaluate a computational process for producing recommendations.

Calibration of users' expected reliability of a system and its actual reliability should be an important goal for system designers, because this calibration can lead to better decisions (Dzindolet et al., 2003, 2002). Poor decisions are certain to happen if users believe a system to be either much more reliable or much less reliable than it actually is. Therefore, IDA designs should consider ways to help users calibrate their efficacy beliefs.

Another contribution of this study is that it demonstrates that efficacy beliefs are malleable, and that the design of the system can influence users' efficacy beliefs, at least for new users like the subjects in this study. The encouragement variable manipulated three aspects of the system. The system itself stated its expected efficacy by indicating its average performance within the interface. Also, subjects were given information about the quality of the data that it uses to make predictions. And the configuration that the system used was either one that might be expected to work well or one that used categories most people do not think could work well for predicting game outcomes. Because the intent of the encouragement variable was to randomly manipulate efficacy beliefs, these three design features were varied to high or low degrees simultaneously in an effort to maximize random variation in efficacy beliefs. It is clear that at least one of these design features is responsible for the observed differences in efficacy beliefs between conditions. However, the study design makes it impossible to determine the specific effect of each individual feature on efficacy beliefs. Nevertheless, all three are reasonable causes of efficacy beliefs that can be incorporated into a system design. Future work can explore how these features and others influence users' expectations about IDA output. A particularly useful future study would examine the concept of data quality more precisely to determine how users evaluate the quality of data that are input into IDAs and how that affects their expectations of their efficacy. This dissertation has primarily focused on customization of IDA logic, but customization of data is another way to make systems customizable. Evaluating whether customization of data affects efficacy beliefs, and how customization and general attitudes about data affect decisions and trust in the system, would make a valuable continuation of the work presented here.

The result that high efficacy beliefs improved decision making warrants some additional discussion as well. Since high efficacy beliefs cause users to agree with recommendations, high efficacy works to improve decision making when systems are largely reliable and users otherwise commit omission errors by not following good recommendations. However, had the IDA in this study been far less reliable, it is possible that high efficacy beliefs could have hurt decision making. Again, this illustrates the importance of calibration between users' beliefs and actual system performance. Without this calibration, IDAs can lead users to poor decision making. Designing to allow users to form accurate beliefs about an IDA's efficacy can have important benefits for making IDAs that are effective at improving user decision making.

4.3.1 Limitations

This study was designed to identify the causal effect of efficacy beliefs on agreement, and offers only very limited insight into how users naturally form expectations of the system's efficacy. The encouragement variable manipulated three aspects of the system (configuration, past performance, data quality) and its instructions, but these aspects were intentionally confounded with each other. Therefore I cannot assess how much each specific aspect contributed to the variance in efficacy beliefs. Since the study found that expectations can have a strong effect on users' behavior, it is important to explore aspects of the system design or characteristics of users that influence efficacy beliefs so it can be determined how best to engineer agreement with recommendations.

CHAPTER 5

CUSTOMIZATION, NON-CUSTOMIZATION, OR BOTH IN AN IDA FOR EXERCISE DECISIONS

Customization in IDA designs can create a difficult socio-technical challenge for designers. Customizable IDAs may be able to take advantage of users' expertise, their local knowledge of their decision scenario, as well as personal knowledge of their preferences and decision making styles to improve the IDA's algorithm and provide better recommendations. However, as I will show in this chapter, there are decision-making consequences to this design, and these consequences may not be justified by an improvement in the IDA's algorithm.

In this chapter, I will present a study that compares an IDA design utilizing customizable recommendation logic to a design that does not give users any control over the recommendation. Additionally, this study will evaluate a design that provides users with both customized and non-customized recommendations simultaneously. This study will evaluate and compare these different designs regarding their effect on users' decision making. The purpose of this evaluation is to produce insights that can help IDA designers determine how best to incorporate customization, if at all, in an IDA design. This study will also extend my work on customization bias and replicate previous findings in a different decision context using a different type of IDA, notably that will actually adjust recommendations in response to users' configuration decisions.

In this study, I will also describe a test of the effect that customization has on the transparency of an IDA. Transparency in IDAs is typically thought of as a function of providing explanations in the interface about how recommendations were generated (see section 5 of Chapter 2 above). However, customization may make a system more transparent without providing explanations because it gives signals about what aspects of the decision are important to the algorithm. In this study I will show that customization leads to a moderate increase in the transparency of the IDA. I will also show an association between transparency and users' agreement with recommendations. Users who feel the system is more transparent are more likely to agree with it, even when controlling for recommendation quality and customizability of the IDA. This association suggests another important decision making bias that results from IDA use.

5.1 Research Questions

RQ1. *Can customization lead IDA users to make better decisions even if recommendations are no better than those provided by a non-customizable IDA?*

One argument that can be made for making IDAs customizable is that users provide expertise, local knowledge and situational awareness about the decision scenario and their own preferences that is impossible or difficult for the system to obtain from any other source but the decision maker. Therefore, by incorporating a customizable algorithm into an IDA design, designers allow the system to capture information that may lead to better recommendations for specific decisions. However, because customization places additional demands on users, it creates a different decision process compared to systems that completely automate the acquisition and analysis of information. Users must think about how to configure the IDA, spend time and effort doing so, and interpret both how their input has influenced the system's output as well as evaluate the output and its appropriateness for the decision. Even if customization leads to better recommendations, these additional demands placed on the user may affect their decisions. For example, the process of thinking about how to configure the system might lead users to new insights about the decision, helping them make better decisions in a way that is not related to the actual recommendations provided. Or, it may create fatigue that leads users to insufficiently evaluate all alternatives or take cognitive shortcuts when selecting an action. These are examples of mechanisms by which customization might affect decision making that are independent of the recommendations that the IDA produces.

In this study I will conduct an experiment that compares decision making as supported by both customizable and non-customizable IDAs. I have designed a study to seek out evidence that customization can lead users to make better decisions. The mechanism by which I suspect this could happen is elaboration over the decision that may be structured within the decision process by the design of the IDA. This mechanism is a cognitive process for which there is not a clear or direct measure at this time. Therefore, I will first determine whether customization has any affect on decision quality in this study so that I can determine whether development of a measure of customization elaboration is a worthwhile research direction.

The hypotheses related to this research question are:

- **H1** Users who use the IDA will make better decisions than others who make the decision unaided.

- **H2** Users who see customized recommendations will make better decisions than users who see only non-customized recommendations.

RQ2. *Can customization bias be observed in an IDA that is truly customizable?*

A major limitation of the studies I have presented on customization bias (Chapter 3, (Solomon, 2014)) is that the IDA used in those studies was merely a shell that included a customizable interface, but the system's recommendation logic was not truly customizable. Although manipulation checks from those studies indicated that users did perceive to have some control over the IDA when customizing it, those results may have been tempered by the fact that subjects might not always have been able to recognize a clear connection between their actions and the system's output. Therefore, those results may be only a manifestation of the Illusion of Control (Langer, 1975). However, recent work on this theory has demonstrated that the Illusion of Control can only be observed in situations when people have little to no actual control (Gino et al., 2011), as was the case in my previous studies.

Gino et al. found that in situations where there is a high degree of actual control, people tend to underestimate their control. If customization bias from my previous studies was the result of the Illusion of Control, then in a context where users have a high degree of actual control over the recommendations, it is plausible that customization bias would not be observed or that customizers may be biased against following their custom recommendations. For this reason, it is critical to evaluate customization bias in a context where users have a high degree of actual control over the recommendations that are produced by the IDA. The hypothesis (**H3**) is that users who see only custom recommendations will have more agreement than users who see only non-custom recommendations.

RQ3. *Can an IDA design utilizing both customized and non-customized recommendations reduce customization bias or improve decision making over either approach individually?*

Customization bias may have positive or negative effects on IDA effectiveness. For example, in a system where users have too little trust and often ignore good recommendations, customization may help users make better decisions by increasing the likelihood that they follow the system's advice. However, customization may also create automation bias where users are too trusting of the system and follow poor recommendations. Because customization may have negative effects on decision making, developing design interventions that can reduce customization bias is an important contribution to IDA research.

One potential intervention on customization bias is for the IDA to simultaneously show both customized recommendations as well as recommendations produced by an algorithm that is not affected by users' inputs. Non-customized recommendations may provide users with a contrasting perspective or a "second opinion" about what decision should be made. This additional perspective may trigger users to scrutinize their recommendations more closely and base their decision on a broader set of information. If users consider more alternatives, they may be more inclined to accept some alternatives that were not generated by their customized algorithm.

The hypotheses are:

- **H4** Users who see both custom and non-custom recommendations will make the best decisions overall.
- **H5** Users who see both custom and non-custom recommendations will have less agreement with recommendations than users who see only custom recommendations.

RQ4. *Can customization create transparency in an IDA?*

Transparency is well understood to be important for the user experience of IDAs (Cramer et al., 2008; Wang & Benbasat, 2007). IDA research has often used explanations as the design mechanism to create transparency (Herlocker et al., 2000). Not all types of explanations are effective however (2009), and some work has found that explanations need to be tailored to individual users to be effective (Tintarev & Masthoff, 2007). Tailoring explanations to individual users may be a difficult technical challenge, and I am not aware of any work that has reported succeeding at this. And other work has found that some users' decision making is inhibited by explanations (Ehrlich et al., 2011).

Customization, however, may provide an opportunity to add transparency without needing to provide explanations. A customizable system, by virtue of giving users controls that describe some effect on the algorithm, give users a signal about what is important to the algorithm. This signal may help users understand the system and give it transparency. In this study that compares customizable and non-customizable versions of an IDA, I will measure differences in how well users feel they understand the algorithm and the logic behind the recommendations. I hypothesize that customizing the system will make users feel the system is more transparent (**H6**) and that the more transparent users feel the system's logic is, the more they will agree with recommendations (**H7**).

5.2 Exercise Recommender

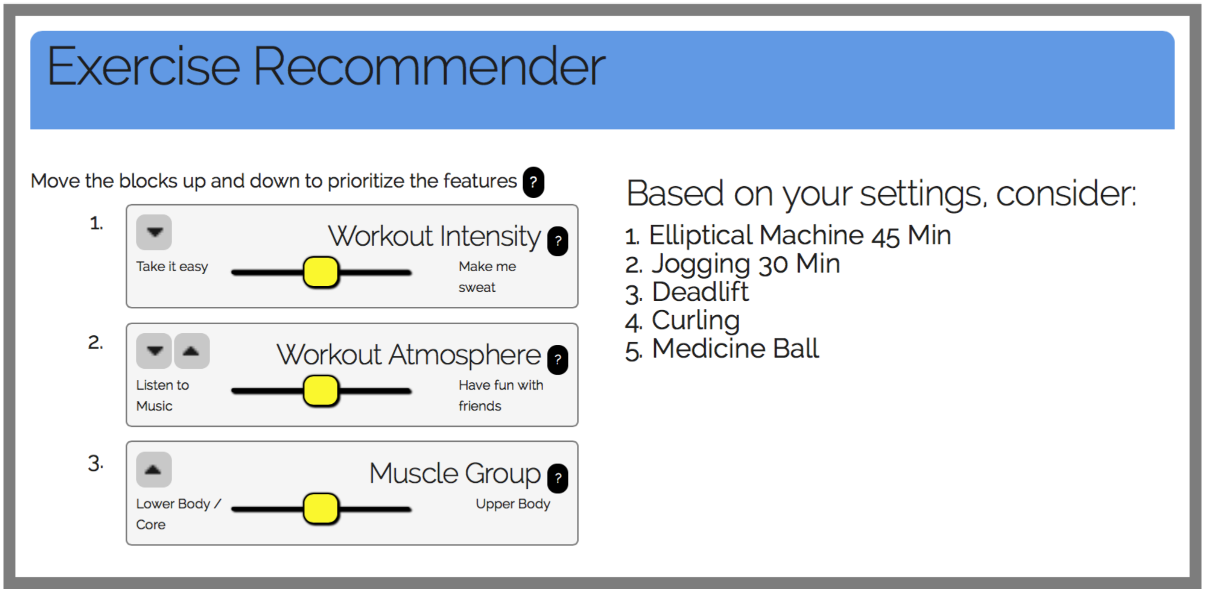
To answer these research questions, I built an IDA that makes recommendations to users about fitness or exercise activities. In this section I will describe this system and its design in detail. Figure 5.1 and Appendix G show the interface for this system.

5.2.1 IDA Recommendation Data

I created a list of 50 exercise activities by referencing existing catalogs and selecting exercises or activities that are well known. In selecting exercises, I tried to create a balance between having diversity in the types of activities that could be chosen –so that the activities could be differentiated from each other by decision makers–, and creating an exhaustive list of activities that would be burdensome for subjects to browse when participating in the study.

After creating an initial list of activities, I then chose a set of attributes by which these activities could be evaluated. These attributes are:

- Cardio - The amount of aerobic exercise required by an activity.
- Intensity - How physically or psychologically intense an activity is. Low intensity activities can be thought of as relaxing.
- Group - How many people are ideally needed for an activity. The highest value means at least 15 people
- Lower Body/Core - The degree to which an activity provides exercise for the legs, abdomen, or lower back.
- Upper Body - The degree to which an activity provides exercise for the arms, neck, shoulders, or upper back.
- Convenience - The resources required for an activity. Inconvenient activities require a lot of money, equipment, time, or other resources.



Exercise Recommender

Move the blocks up and down to prioritize the features ?

1. Workout Intensity ?

2. Workout Atmosphere ?

3. Muscle Group ?

Based on your settings, consider:

1. Elliptical Machine 45 Min
2. Jogging 30 Min
3. Deadlift
4. Curling
5. Medicine Ball

Figure 5.1: Exercise Recommender interface.

- Difficulty - The amount of skill or experience required to perform an activity. Easy activities can be completed by a novice. Difficult activities require training or expertise to perform optimally.

- Fun - Whether an activity is enjoyable or not.

I developed this list of attributes for two purposes. One purpose was to create a set of preference profiles that represent a set of things that a user of an IDA such as this might care about when deciding on an activity. The second purpose was to create attributes for the IDA's content-based recommender system to use when making recommendations. As discussed in Chapter 2, content-based recommender systems make recommendations based on explicitly-known attributes of the items in the system's catalog. By creating this list of attributes, I was able to give the IDA explicit content information about the exercises in order to make recommendations.

I chose attributes that would create some diversity within the framework of *search* attributes versus *experience* attributes. Nelson (1970) developed this framework, which has become widely adopted in marketing of consumer goods, that distinguishes between attributes of products for which information can be obtained through search compared to things for which information can only be obtained through experience. For example, a mattress has some search characteristics such as its size, price, or brand name. This information can be obtained easily through a search. However, a mattress also contains experience attributes (e.g. comfort) which for any particular person may only be obtained by experiencing the mattress. The distinction between these two types of attributes is important for recommender systems (Ochi, Rao, Takayama, & Nass, 2010). Users may be able to reasonably expect a system to have and use information about search attributes of the recommendable items, but experience attributes may be, or at least appear to users, to be less compatible with automated recommendations. Since users may care about many experience attributes, but may expect the system to maintain and use search attributes, I included attributes that fit across the continuum of this framework. The *Fun* attribute is a strong example of an experience attribute, whereas the muscle groups worked by an activity can be easily obtained through a search or through basic knowledge of an activity. Other attributes like the intensity or difficulty may be understood through search but likely require some experience as well to obtain complete information. IDA users, particularly of customizable systems that demand user input, may find it challenging to coerce the system into considering both search and experience characteristics. Yet this challenge is inherent in many contexts of IDA-supported decision making, and for this reason I chose to diversify the attributes that go into the preference profiles and into the system's content-based recommendations as an acknowledgement of this challenging aspect of IDA-supported decision making.

5.2.1.1 Crowdsourcing Exercise/Attribute Evaluations

Because the attribute list contained both search and experience attributes, and because I was unable to locate a reliable source of objective evaluations of exercises for all search attributes, I conducted a survey to crowdsource the evaluation of all exercises against all attributes. This crowdsourcing serves as a way to bootstrap the IDA by obtaining some initial content ratings by which to make recommendations.

For the survey, 112 crowd workers were enlisted through Amazon Mechanical Turk, and each was paid $0.50 for completing the survey. The survey asked them to rate 40 exercise activities that were divided into 4 groups of 10, with each group being rated according to a different attribute. For example, a subject would rate one group of ten exercises on the *Fun* attribute, then another group of 10 on the *Difficulty* attribute and so on. The survey was run until at least 10 ratings had been recorded for each of the 400 exercise/attribute pairs (50 exercises by 8 attributes). Workers were given the description for each attribute that is listed above. The ratings for each attribute were assessed on a 10-point scale. Appendix D shows an example of this survey.

The survey contained several measures intended to ensure high data quality. Workers had to answer three "attention check" questions where they were given specific instructions in the question prompt to verify that they were reading the prompts. Any worker who failed any of these attention check questions was removed from the final data set. Also, each subject was shown two "repeats" of questions they had previously answered to determine whether they would be consistent about their ratings. Any worker whose answer to a repeated question deviated by more than one point from their original answer was removed from the final data set. The 10-point scale also included an additional option to indicate that the worker was not familiar with the activity in question, and exercises that frequently received this response were removed. After all data cleaning, 83 workers and 44 exercises remained in the data set, for a total of 3320 ratings.

5.2.1.2 Latent Attributes

Table 5.1: Latent attributes from exercise survey.

Attribute	Workout Intensity	Workout Atmosphere	Muscle Group
		Factor Loadings	
Cardio	0.625	0.343	-0.213
Convenience		-0.659	
Difficulty	0.520		
Fun		0.709	0.196
Group		0.534	-0.207
Intensity	0.812		0.111
Lower Body	0.531		-0.415
Upper Body			0.823
		Factor Fit Measures	
Sum-of-Squared Loadings	1.609	1.345	0.991
Proportion of Variance	0.201	0.168	0.124

In some initial usability testing of the Exercise Recommender and the study task, I determined that eight attributes were too many for users to simultaneously consider in an explicit way both for using and configuring the Exercise Recommender and for making a decision that matched the preference profile for all attributes. To respond to this problem, I conducted an exploratory factor analysis on the ratings survey data to derive a smaller number of latent factors.

In this factor analysis, I created an exercise-by-attribute matrix, and in each cell of the matrix I inserted the median rating for that combination. The intent of the factor analysis was to derive some latent factors that represent combinations of the explicit attributes such that the correlations between the attributes could be translated into a meaningful construct or label. In an exploratory factor analysis using varimax rotation such as this one, the researcher specifies a number of latent factors for the algorithm to find such that the variance between the different factors is maximized. I initially specified four factors, and found four factors with sufficient factor loadings. However, upon examining the specific correlations within these factors and the list of exercises that would score high or low on these factors, the fourth factor (which by design of the factor analysis process explains the least variance) did not appear to have an identifiable theme. I was unable to put a clear label to this fourth factor, which would make it confusing to users of the Exercise Recommender to understand. The first three categories did have a more clear theme however. I then ran a second factor analysis searching for only three factors, the analysis returned three factors with the same theme and general correlations as the original analysis with four factors. This analysis is described in Table 5.1. I used this second analysis to generate factor scores using Bartlett's method (DiStefano, Zhu, & Mindrila, 2009) for each exercise/latent attribute combination.

I added the label *Workout Intensity* to the first factor. The intensity attribute had the highest loading, followed by cardio, lower body, and difficulty. Activities that scored high on this factor were things like Climbing Stairs, Jumprope, Basketball, Jogging for one hour, and Boxing. Activities like Stretching, Bowling, Golf and Yoga scored low on this factor. The Exercise Recommender describes this factor as "How much will the activity make you work hard, breathe hard and sweat." Users can specify a level for this latent attribute between "take it easy" and "make me sweat."

I gave the second factor the label *Workout Atmosphere*. This factor had high loadings for the *Fun* and *Group* attributes, as well as strong negative loading for *Convenience.* This factor appears to identify activities that are fun to do in a group of people such as recreational activities (Snorkeling, Whitewater Rafting, Square Dancing) and team sports (Soccer, Ultimate Frisbee). Activities that are typically done by oneself in a gym scored very low in this factor, such as Lunges, Squats, Planks, and Bicep Curls. To help users understand this latent attribute, the high level of this attribute was labeled in the Exercise Recommender interface as "have fun with friends" and the low level as "listen to music." The description of the latent attribute within the interface was "Fun social recreation activities versus solitary workouts."

The strongest factor loadings for the final latent attribute were the *Upper Body* and *Lower Body/Core* attributes, which had opposite signs. For this reason, I labeled this latent attribute as *Muscle Group* and labeled the high level as "upper body" and the low level as "lower body/core."

5.2.2 Customizing Recommendations

One of the most important goals for the design of this system was to allow users to customize the recommendations. Users can customize recommendations by specifying what level of each of the three latent attributes they prefer. For example, if users want an exercise that will be intense and also fun, they might set *Workout Intensity* and *Workout Atmosphere* to the higher settings. And if they prefer a core exercise, they could set *Muscle Group* to a lower setting. The Exercise Recommender also allows users to prioritize the attributes. The recommender algorithm described below can give more weight to items that match some attributes than matching on others. To facilitate this in the user interface, users can move the blocks that contain each attribute up and down through the list and rearrange their order. The attribute at the top of the list is given the highest weight in the algorithm, making activities that match closely on this attribute more likely to appear in the recommendations.

The recommender algorithm is based on a formula for weighting queries in recommender systems presented by Schafer, Konstan and Reidl (2004). This formula calculates a similarity score between a user's query and the items in the system's catalog. A user's query consists of two values for each of the three attributes for which users can specify their preferences. The first value is the level that the user specifies for that attribute. These values are on a scale between -2 and 2, which approximately matches the range of the centered factor scores stored for each activity/attribute pair within the system's database. The second value is the rank (1, 2, or 3) that the attribute is set to as priority for that attribute. This second value is subtracted from 4 in order to give the highest ranked attribute a value of 3 and the lowest a value of 1.

Taking this query, the Exercise Recommender calculates the similarity between this query and every activity in the system's database using Equation 5.1.

$$Similarity(Activity, Query) = 1 - \sqrt{\frac{\sum\limits_{a \in attributes} w_a^2 (1 - d_a)^2}{\sum\limits_{a \in attributes} w_a^2}} \tag{5.1}$$

In this equation, d represents the degree of match between the query and the activity's rating (factor score) for that attribute. d is calculated as $d = \left| QueryLevel - Rating \right| / 3$. Dividing by 3 sets d on a scale between 1 and 0 which is required by the equation. w is the weight that the attribute is being given based on the priority ranking. For example, if a user set the *Muscle Group* attribute to level 2, and the activity being calculated has a score of 1 for that attribute, d would be calculated as $\left| 2 - 1 \right| / 3 = 0.333$. If that attribute was ranked as most important, w would equal 3 in the equation.

After all similarity scores have been calculated, the activities are then sorted by these scores and the top five are returned to be presented in the user interface. The Exercise Recommender was built as a web application that was accessed through a web browser. I built and deployed the system using the web.py[1] framework, HTML5, JQuery 1.10.2 and

[1] http://webpy.org

jQueryUI 1.11.4 [2], and the Skeleton CSS framework [3]. These technologies enabled the system to be usable both on a traditional desktop web browser as well as other devices with smaller screens and touch devices.

5.3 Methods

I conducted an experiment in which subjects would use a version of the Exercise Recommender to help them make a decision about which exercise activity would be most appropriate given their preferences for the kind of activity they want. The study compared decision making using different versions of the Exercise Recommender, as well as a control condition of unaided decision making, in order to answer the Research Questions described above.

5.3.1 Decision Task

In order to evaluate the quality of decisions made by users of the Exercise Recommender, I created a decision task in which subjects were given a set of preferences for some attributes of exercise activities and were asked to choose an activity that most closely matched those preferences. An important feature of this task is that subjects were not choosing activities that they would prefer to do outside of the context of the study on their own. Instead, they were choosing exercises that matched an assigned preference profile.

Under this method, rather than having each subject supply their own preferences as would be most naturalistic, preferences are assigned to subjects by the researchers. I gave subjects an incentive in the form of additional payment to make decisions that match the assigned preferences rather than their own personal preferences. By assigning all users the same preferences, decision making about what might otherwise be a horizontally differentiated set of preferences can be made vertically differentiated and objectively evaluated. This

[2] http://jqueryui.com

[3] http://getskeleton.com

approach draws on methodology from experimental economics, and specifically on Induced Value Theory (Smith, 1976). Smith showed that in the absence of other incentives, subjects in a decision making experiment can be assigned preferences for some decision alternatives over others by varying a financial reward from the experiment from choosing those more valued alternatives. For example, if subjects are told that if the outcome of a game is A, they will earn 1 point and if it is outcome B, they will earn two points, and the points are later exchanged for real cash, subjects will actually prefer the game to have outcome B over A, and will make decisions that they believe will result in B.

The decision task I created for this study makes use of Induced Value Theory by giving subjects a set of preferences represented as point values that can be earned from a chosen exercise activity. Subjects were shown five attributes of exercises. These attributes were five of the eight attributes that were evaluated in the exercise rating survey described above. The attributes that made up a preference profile were *Cardio*, *Convenience*, *Fun*, *Difficulty*, and *Group*. For each of these attributes, subjects were told they prefer either a high or a low level, and assigned a point value for that attribute. The ratings given to the activities for these attributes establish a ground truth. For each attribute, a median split of the ratings for that attribute determined whether the exercise would be classified as “high” or “’low” for that attribute. When a subject selected an activity, that activity’s ground truth high/low ratings were compared to the subject’s preference profile. Subjects earned the specified number of points if their selected exercise had the same high/low level as their preferred level from the profile.

For example, if a subject had the following preference profile:

- High in Fun- 30 points
- Low in Cardio- 20 points
- Low in Group- 15 points

- High in Convenience- 10 points
- Low in Difficulty- 5 points

and the subject chose Bowling, the subject would earn 55 points because Bowling is rated as High in *Fun* (30 points), Low in *Cardio* (20 points), High in *Group* (0 points, not matched), Low in *Convenience* (0 points, not matched) and Low in *Difficulty* (5 points).

Previous work on IDAs in e-commerce has used a similar approach to evaluating decision making (Pereira, 2001) that appropriates points to users based on the match between their preference for attributes and the item they have selected. There are a few important differences between Pereira's method and my own. Pereira did not offer an incentive for subjects to earn more points. I consider this incentive to be critical for a study conducted online by Mechanical Turk workers, who otherwise have great incentive to perform the task quickly without trying exceptionally hard to make good decisions. The second difference is that Pereira asked subjects to create their own preference profiles prior to using a decision aid by selecting their preferred values for each attribute then weighting the importance of the attributes. This method creates some external validity for the decision task because subjects are making decisions for their own preferences. However, there are some serious threats to internal validity with this approach that make it inappropriate for my study. First, it eliminates the possibility to evaluate whether subjects improve their decision making over time because a subject can have only one profile. Second, it creates a potential selection bias. Since subjects choose their own preferences, it is possible that some subjects have preferences that are inherently a better match to the limited catalog of options than others. This means that some subjects may have more options available to them than others that would give high scores, and the maximum possible score may be different for each subject. For these reasons, I chose to assign users the same preferences, leaning on Induced Value Theory to provide assurance that the assigned profiles do in fact create a true preference for selecting activities that match them.

Along with other colleagues, I have used this approach based on Induced Value Theory to assess decision making crowdfunding systems (Wash & Solomon, 2014; Solomon, Ma, & Wash, 2015). That research took advantage of another important characteristic of Induced Value Theory, which is that it can be used to induce preferences in order to evaluate strategic decision making in games played between multiple parties. The application of this approach in this study about the Exercise Recommender is more simple, as it involves no strategic behavior. Subjects' payouts are determined entirely by their own decisions. This is a useful feature of this task design because it removes some external incentives that can creep into induced value experiments, such as learning and negotiation effects that can come from repeated games (Andreoni, 1988). Instead, the quality of subjects' decisions can be quite objectively measured as the number of points earned from their activity selection.

To create the ten preference profiles, I first clustered the exercise activities according to their ratings for the five attributes in the preferences using hierarchical agglomerative cluster analysis. I created ten clusters, then chose one activity from each cluster to set as a "top choice." I then set the preferred level of each attribute in the profile to equal the rating of the chosen exercise, so that every preference profile had at least one activity that would score the maximum of 80 points. The cluster analysis created diversity in the preference profiles, and ensured that all types of activities that could be chosen were represented among the possible best choices.

5.3.2 Experiment Conditions

There were four conditions of the study, which each relate to different versions of the Exercise Recommender. In the *Unaided Decision* condition, subjects did not use the Exercise Recommender to help them make a decision. In the *Custom Only* version, subjects could customize the Exercise Recommender's recommendation algorithm as described above, and were shown only recommendations that were produced by their customized algorithm. In the *Non-Custom Only* condition, subjects used a version of the Exercise Recommender that was not customizable, but provided recommendations of equivalent quality to users. These non-customized recommendations will be described in detail below. In the *Both algorithms* version, the Exercise Recommender displayed recommendations produced by the users customized algorithm, as well as recommendations produced by the non-customized algorithm.

Non-customized recommendations. Customization may affect decision making if users are effective at customizing the algorithm and therefore improve the recommendations that are produced. An important goal of this experiment, and in understanding customization bias and decision making, is to account for the effect that recommendation quality might have. Recommendation quality can be easily measured in this design due to the nature of the decision making task, and therefore analyses can account for it. However, it was nevertheless desirable that recommendation quality be as close to constant as possible between conditions. Because recommendation quality is a dependent variable and cannot be assigned to subjects who customize the IDA, this is not a straightforward manipulation in the experiment design.

To balance the quality of recommendations between conditions, the non-customizable algorithm used a pool of recommendations that had been produced in a pilot study by users of the Custom-only system. In this pilot, 37 subjects participated in the the *Custom Only* condition of the study. When the Exercise Recommender needed to display some non-customized recommendations to a user, it randomly sampled a recommendation set from this pool. The Exercise Recommender selected a recommendation set that had been produced by a pilot subject using the same preference profile as the current subject requesting the recommendations. This method of giving non-customized recommendations resembles collaborative filtering in that the system gives recommendations to a user based on the actions of other similar users.

The Exercise Recommender did not give an explanation for how the non-customized algorithm worked, other than to indicate to users in the *Both algorithms* condition that they were not affected by the user's input. There were several reasons for not explaining the non-customized recommendations. First, as a design choice for an IDA there is little justification in the literature on explanations for giving explanations in regards to decision making. Explanations do not have a clear benefit to decision making, and in some cases may harmful to decision making (Ehrlich et al., 2011). A second reason is that if any explanation were to be given, it would create a conflict for experimental validity. For example if users in the non-custom only condition were told that other similar users had customized the IDA, it would reveal the nature of the experiment to those users and create a potential for demand effects. If a false or deceptive explanation were given, it would violate the established norms of experimental economics upon which the decision task is based. Furthermore, different explanations may result in different effects, as has been found in several previous studies (Lim et al., 2009; Tintarev & Masthoff, 2008, 2012), which means that the specific explanation chosen could potentially have a direct effect on the results of the experiment. Evaluating specific explanation designs is not one of the research questions for this study, although it is an important area for future research that builds off of the experiment protocol I have developed for this study.

5.3.3 Procedure

I recruited subjects from Amazon Mechanical Turk to participate in a study about Exercise Choices. I offered $2.00 as a guaranteed payment for participation, and told them in the recruitment post that that they could earn a bonus payment depending on their decisions in the study, with the average bonus payment expected to be $3.00. After enrolling in the study, subjects were assigned to a condition and given instructions about their task. These instructions can be seen in Appendix F. These instructions described the decision task, including how many points would be earned and descriptions of the attributes they would be given preferences for. After viewing the instructions, subjects took a quiz that evaluated their understanding of the key points of the instructions. In particular, this quiz required subjects to demonstrate that they understood how their scores were calculated, understood the Exercise Recommender and the way it produced recommendations, and verified that subjects understood their incentive to make good decisions for their preference profile. If subjects answered any question incorrectly, they were shown the correct answers and then redirected to the instructions for review. They then had to retake the quiz with the questions and answers slightly altered and reordered. They had to pass the quiz before moving on, but were free to take it as many times as needed. The median number of quiz attempts was 2.

After completing the quiz, subjects were shown a page with their preferences. In the unaided condition, they were also shown a menu where they could select an exercise for the given preference profile. In the three Exercise Recommender conditions, after viewing their preferences they had to load the Exercise Recommender into a frame on the screen. They then used the system to get recommendations, and once recommendations had been generated the menu appeared where they could make their decision. Recommendations could only be produced one time per round, so that all decisions were based on using the IDA to produce one set of recommendations, rather than allowing for repeated requests for recommendations.

After making their decision, subjects were shown their score, and reminded that the maximum possible score was 80 points so they could gage how close to the best possible decision they had made. Subjects then proceeded to the next round where a new set of preferences was shown. Subjects used the same version of the Exercise Recommender (or used no IDA) for all rounds of the study. After finishing the ten rounds of the decision task, subjects were directed to a post-test questionnaire. The order of preference profiles that subjects made decisions for was randomized for each subject.

A total of 155 subjects completed the study. After an initial analysis of the data, I removed two subjects who completed the study from the same IP address, and five subjects who completed the decision-making portion of the study in less than 3 minutes. In pilot testing I determined that three minutes was too fast for subjects to have been seriously looking at their profiles and making thoughtful decisions, so I removed four subjects who completed the study in less than three minutes. 48% female with a median age category of

26-34. Subjects took an average of 27 minutes to complete the study.

5.3.4 Measures

Decision quality.

I measured decision quality using the number of points earned by the chosen activity for the subject's preference profile.

Recommendation quality.

The Exercise Recommender gave five recommendations in each round, plus an additional five to subjects in the Both Algorithm condition. I used three different measures to assess the quality of recommendations that a subject received in a given round. *Average Recommendation Quality* was the average number of points that all recommendations shown would earn. *Weighted Recommendation Quality* gave extra weight to recommendations higher on the list. The average was calculated by adding copies of each recommendation score to the vector, with the number of copies determined by the activity's rank in the recommendation list. The top activity was repeated 5 times, while the fifth activity appeared only once. The third measure is the score of the best recommendation on the page.

Post-test measures.

The post-test questionnaire assessed several additional measures. Appendix H shows the full post-test questionnaire. This questionnaire assessed subjects' knowledge about fitness and exercise activities, their user experience with the Exercise Recommender, their perception of transparency of the Exercise Recommender, their propensity to trust decision aids, and demographic information. This questionnaire also contained questions about whether they looked at, considered, trusted or ignored recommendations from the Exercise Recommender.

Table 5.2: Percentages of people answering yes to these questions about the Exercise Recommender.

	Custom Only	Noncustom Only	Both
I used the Exercise Recommender to help me make my decision	92%	66%	94%
I could configure the Exercise Recommender to adjust the recommendations it gave me	73%	3%	77%
The Exercise Recommender gave me some recommendations that I had no control over	41%	18%	74%
The Exercise Recommender gave me only recommendations that I had no control over	11%	45%	3%

Table 5.3: Questionnaire questions about recommendations. Answered on 5-point Likert scale (Strongly Disagree to Strongly Agree). Standard deviations in parentheses.

	Custom Only	Noncustom Only	Both
Looked at recommendations	4.216	4.026	4.429
	(0.584)	(0.944)	(0.502)
Trusted recommendations	3.297	2.921	3.229
	(0.996)	(1.148)	(0.843)
Ignored recommendations	2.189	2.579	2.229
	(0.811)	(1.056)	(0.843)
Looked at custom recommendations	4.054	-	3.971
	(0.815)	-	(0.664)
Looked at non-custom recommendations	-	3.552	3.857
	-	(0.921)	(0.879)

5.4 Results

5.4.1 Questionnaire

Table 5.2 describes subjects' answers to questions about the Exercise Recommender. This table illustrates that subjects in the Both Algorithms condition understood the nature of their recommendations, as they found the system to be configurable, and that some but not all of their recommendations were not influenced by their configuration.

Table 5.3 describes the questionnaire questions about attention to the Exercise Recom-

Table 5.4: Self-reported user experience variables.

	Custom Only	Noncustom Only	Both
Useful	3.459	3.000	3.571
	(1.016)	(1.040)	(0.884)
Accurate	3.432	3.000	3.324
	(0.987)	(1.040)	(0.843)
Easy to use	4.135	4.421	4.118
	(0.918)	(0.793)	(1.094)
Configurable	3.541	1.973	3.029
	(0.931)	(1.000	(0.937)

mender's recommendations. There were no statistically significant differences between any conditions on any of these measures. Almost all subjects reported that they looked at and considered the Exercise Recommender's suggestions, with an average higher than the "agree" point and only a few subjects reporting any disagreement with the statement about considering recommendations. Therefore, the data in the three Exercise Recommender conditions can be largely interpreted as IDA-supported decision making. Furthermore, subjects in the Both Algorithms condition indicated equivalent amounts of looking and considering both types of recommendations. The small difference in means between looking at custom and non-custom recommendations within the Both Algorithms group was not statistically significant. This suggests that subjects were aware of all recommendations and looked over them before making decisions, and that there is no difference between any of the Exercise Recommender conditions in terms of seeing and considering recommendations.

Table 5.4 describes responses to questions about the user experience of the Exercise Recommender. ANOVA and Tukey Honest Significant Difference tests found that subjects in the Non-custom condition thought the IDA was less useful than subjects in the Both Algorithm condition. Also, subjects in the Non-custom condition found the system to be less configurable than the other two conditions with the IDA. No other comparisons showed statistically significant differences between conditions.

The questionnaire asked three questions to assess the transparency of the Exercise Rec-

Table 5.5: Self-reported understanding of IDA logic.

	Custom Only	Noncustom Only	Both
Understood why	3.919	3.368	3.829
	(0.682)	(0.942)	(0.664)
Made sense	3.784	3.289	3.629
	(0.672)	(1.011)	(0.770)
Logic was clear	3.784	3.211	3.600
	(0.787)	(1.094)	(1.094)

ommender. Subjects stated on a 5-point Likert Scale whether they understood why the IDA gave certain recommendations, whether those recommendations made sense, and whether the system's logic was clear. Table 5.5 described these results. ANOVA and Tukey's HSD tests were run to determine any differences between conditions on these items. The Non-custom Only condition was rated as having less transparency on all three measures than the Custom Only condition, and ($p < .05$), and the Non-custom Only condition was also rated as less transparent then the Both Algorithm condition on the "understood why" item. This finding is not surprising as the Non-custom recommendations were given no explanation about their process, however it is noteworthy in that it offers some evidence that users perceive the act of customizing specific inputs as a form of transparency. It should be noted however that the effect sizes in these tests are relatively small, and all means sat between the "neutral" and "agree" points on the scale, suggesting that subjects did not find the IDA highly transparent overall.

5.4.2 Recommendation Quality

Table 5.6 describes the means and standard deviations of the three recommendation quality measures within each condition. I conducted ANOVA's on each of these measures by condition. These tests found no statistically significant differences between conditions on the measures of average recommendation quality, weighted average recommendation quality, and number 1 recommendation score. There was a statistically significant difference ($p < .05$)

Table 5.6: Means and tandard deviations of recommendation quality within conditions.

	Non-custom only	Custom only	Both algorithms
Average Rec. Quality	45.013 (15.238)	46.486 (15.210)	46.279 (14.843)
Weighted Average Rec. Quality	43.757 (16.250)	45.014 (16.088)	44.710 (15.999)
Top Rec. Score	64.854 (15.546)	65.897 (15.210)	72.500 (11.959)

on the measure of best recommendation score, which evaluated how many points the best recommendation (appearing anywhere on the page) would earn. Further analysis using a multilevel regression model indicated that the Both Algorithms condition had a higher average for the best recommendation that appeared on the page. This can likely be explained by the fact that the Both Algorithms condition showed ten recommendations compared to five in the other IDA decisions. Figure 5.4 shows the distribution of this measure. In most rounds, the system gave at least one recommendation that scored 80 points, which was the maximum possible in every round.

Overall, the study design was successful at providing recommendations of equivalent quality to subjects in all three IDA-supported conditions. Customized recommendations were equivalent in both conditions that used them, as were non-customized recommendations.

5.4.3 Decision Making

To evaluate decision making, I fit a multilevel linear model that included dummy variables for each recommendation type, the interaction between recommendation types, and random effects for each subject to account for the repeated observations per subject. The model also included a fixed effect for each profile. I used fixed effects because I was interested in the specific effect of each profile in this study, and not any effects that might generalize to a broader population of preference profiles for exercise activities. However, in a post-hoc test I checked this model (and all others in this chapter) using random effects for the profiles, and there was no substantive difference that would alter any conclusions. The intercept of this model represents a decision made in the Unaided condition in profile number

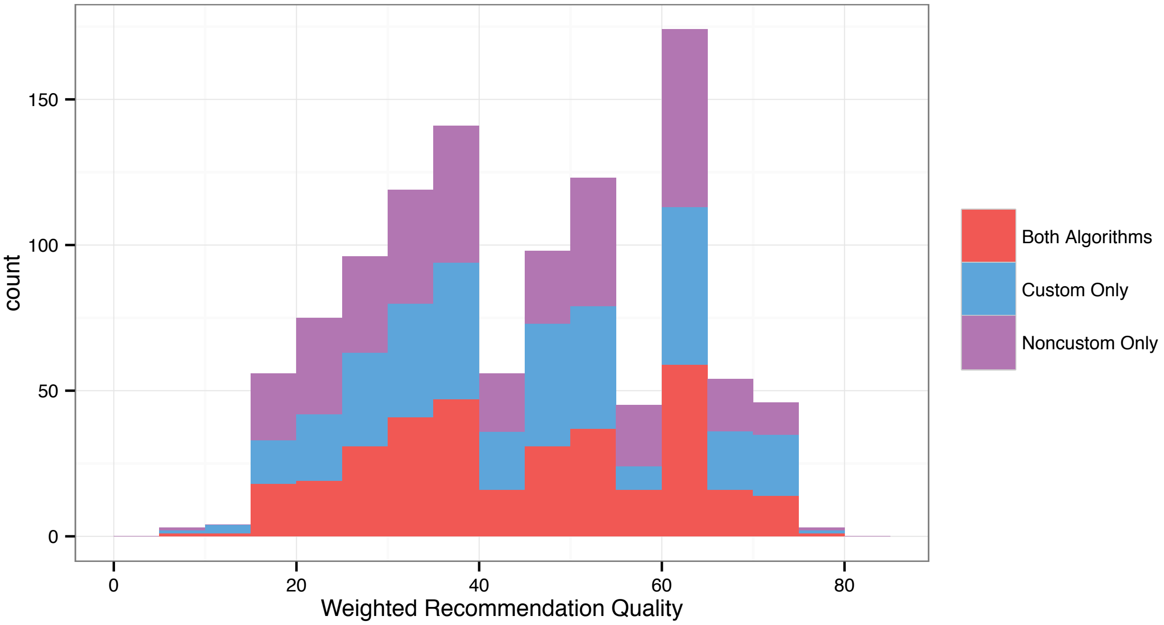


Figure 5.2: Distribution of weighted recommendation scores.

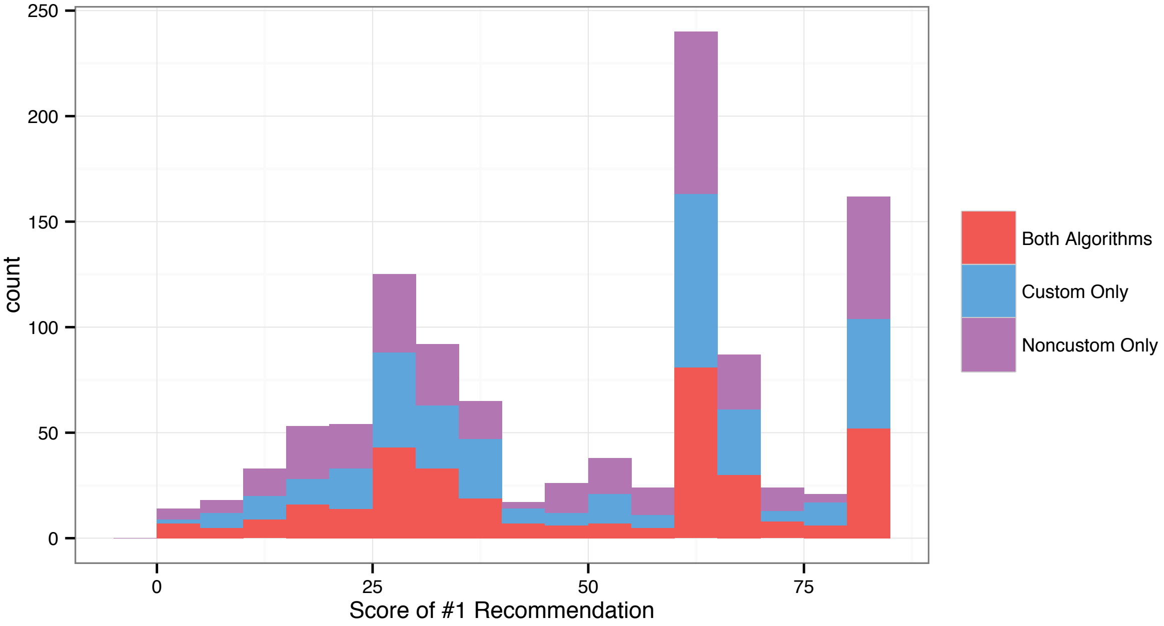


Figure 5.3: Distribution of recommendation scores listed as #1 on the page.

Table 5.7: The effect of each recommendation type on decision quality.

	Dependent variable
	Points
Intercept	59.302***
	(1.815)
Customized Recs. Shown	0.462
	(1.492)
Non-customized Recs. Shown	−0.500
	(1.482)
Both Recommendation Types	−0.811
	(2.125)
Profile 2	−7.230***
	(2.209)
Profile 3	3.074
	(2.209)
Profile 4	−14.662***
	(2.209)
Profile 5	−10.541***
	(2.209)
Profile 6	4.527**
	(2.209)
Profile 7	−4.932**
	(2.209)
Profile 8	−16.385***
	(2.209)
Profile 9	−8.108***
	(2.209)
Profile 10	−7.838***
	(2.209)
Random Effect Standard Deviation	2.370
Log Likelihood	−6,444.970
Log Likelihood χ^2	0.8921
Note:	*p<0.1; **p<0.05; ***p<0.01

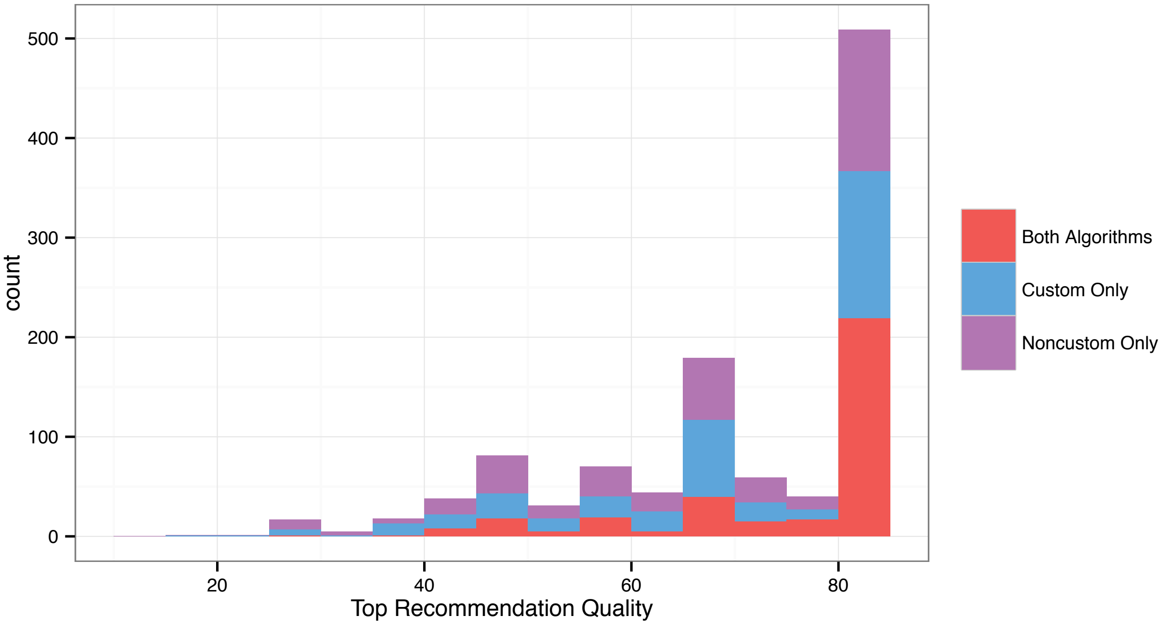


Figure 5.4: Distribution of scores of best recommendation on the page.

1. The coefficient labeled "Both Recommendation Types" represents the interaction effect between seeing customized and non-customized recommendations. This model is described in Table 5.7.

The model suggests that the Exercise Recommender in general did not have an effect on the average decision quality for subjects, as there were no statistically significant differences between the IDA-supported decisions and the Unaided Decision condition. While some profiles appeared to be more challenging than others, the Exercise Recommender in any form did not help subjects make better decisions on average. I conducted a repeated measures ANOVA on this model to test whether there were any differences between any conditions in decision quality. This test found no statistically significant differences between the conditions of the study ($F(3, 144) = 0.287$) on decision quality. For this reason, there is no support for H1 (the IDA will improve decision making), H2 (customization will improve decision making), or H4 (showing both types of recommendations will lead to the best decisions).

Although the Exercise Recommender did not improve decisions on average, it did influ-

Table 5.8: Decision making within only the IDA-supported conditions.

	Dependent variable:
	Points earned from chosen activity
Intercept (Both condition)	28.994^{***}
	(4.358)
Custom Only Condition	5.442
	(4.553)
Non-custom Only Condition	7.327
	(4.464)
Avg. Recommendation Quality	0.530^{***}
	(0.075)
Custom x Rec. Quality	−0.089
	(0.093)
Non-custom x Rec. Quality	−0.141
	(0.092)
Profile 2	−1.235
	(2.554)
Profile 3	2.858
	(2.513)
Profile 4	−3.869
	(2.792)
Profile 5	−3.886
	(2.611)
Profile 6	2.945
	(2.510)
Profile 7	−0.895
	(2.517)
Profile 8	-7.568^{***}
	(2.594)
Profile 9	−1.076
	(2.665)
Profile 10	0.358
	(2.763)
Random Effects Standard Deviation	1.774
Log Likelihood	−4,726.461
Log Likelihood χ^2	267.16^{***}
Note:	*p<0.1; **p<0.05; ***p<0.01

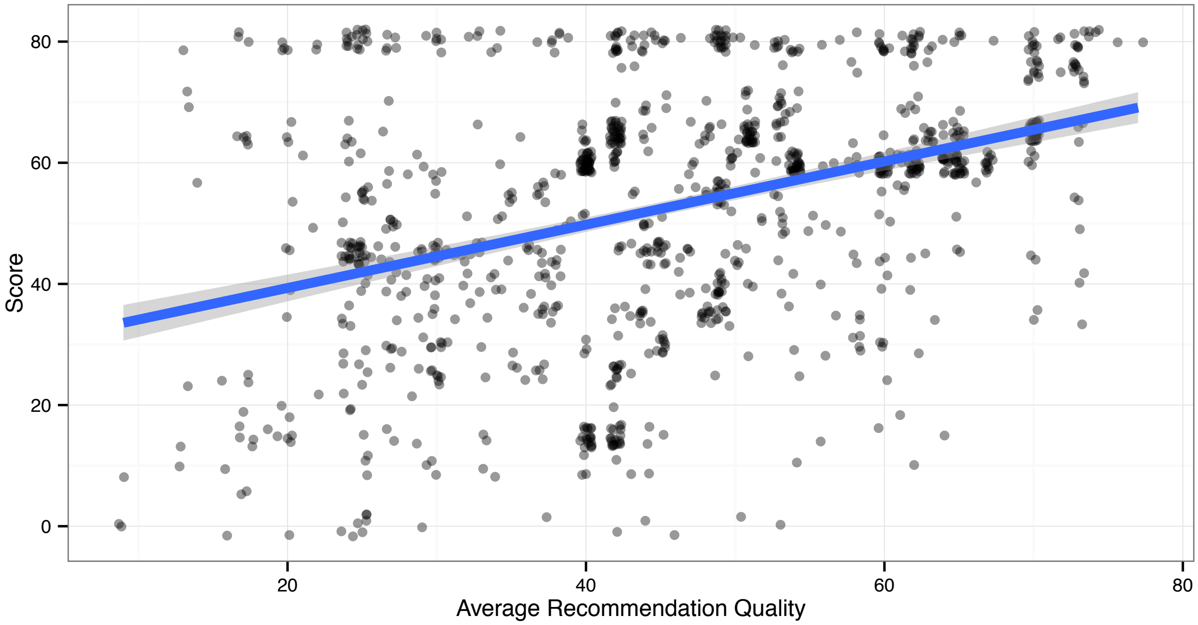


Figure 5.5: Effect of recommendation quality on decision quality

ence decisions. The quality of recommendations given by the IDA had a strong influence on the quality of decisions for users in all three IDA-supported conditions. I fit a multilevel regression model to the data from the three IDA-supported conditions and included the average quality of recommendations on the page as a covariate. This model is described in Table 5.8. The results were similar for all measures of recommendation quality. Figure 5.5 illustrates the relationship between recommendation quality and decision quality. For every additional point of recommendation quality, subjects could expect to earn an additional half a point from their decision.

This finding suggests that the IDA changed the decision process for subjects who used it when compared to those who made unaided decisions. Evaluating recommendations made by the IDA is an important part of decision making, even though it may not actually improve overall decisions.

Table 5.9: Model of agreement with one of the recommendations in Custom only and Non-custom only conditions.

	Dependent variable:
	Logodds of agreement
Intercept (Custom only)	-2.316^{***}
	(0.555)
Non-custom only	-0.536^{**}
	(0.270)
Trust Propensity	0.496^{***}
	(0.158)
Avg. Rec. Quality	0.055^{***}
	(0.008)
Profile 2	1.441^{***}
	(0.409)
Profile 3	0.641
	(0.395)
Profile 4	0.237
	(0.416)
Profile 5	1.025^{**}
	(0.404)
Profile 6	0.407
	(0.390)
Profile 7	1.228^{***}
	(0.414)
Profile 8	2.328^{***}
	(0.459)
Profile 9	1.685^{***}
	(0.420)
Profile 10	0.713^{*}
	(0.412)
Random Effects Standard Deviation	0.801
Log Likelihood	-393.757
Log Likelihood χ^2	137.78^{***}
Psuedo R^2	0.289
Note:	*p<0.1; **p<0.05; ***p<0.01

5.4.4 Agreement

Table 5.9 describes a model that estimates the likelihood of agreeing with one of the recommendations given by the IDA by subjects in the Custom only and Non-custom only conditions (H3). There was a statistically significant difference between these two conditions in agreement with recommendations. Subjects who customized the system were more likely to follow one of its recommendations than subjects who did not customize the recommendations. These results support H3, and offer further support that customization creates a decision making bias by IDA users.

Presenting both customized and non-customized recommendations did lead to a statistically significant reduction in customization bias, which means there is no support for H5. In a model that estimated the likelihood of agreeing with custom recommendations in the Custom Only and Both Algorithms condition, there was no statistically significant difference between the conditions in agreement with the custom recommendations.

Recommendation quality also had an effect on agreement with recommendations. Subjects were more likely to agree with recommendations when they had higher quality. This relationship held true for all measures of recommendation quality. Likewise, subjects who reported a higher degree of overall trust in decision aids were more likely to agree with recommendations. These findings are not surprising, but they are noteworthy because I observed customization bias while controlling for these other important factors. A subject who customized the IDA would be more likely to agree with one of them than someone else who has the same propensity to trust decision aids and received recommendations of equal quality.

The Both algorithms condition offers a within-subjects version of this test of customization bias, as subjects saw both types of recommendation simultaneously. As can be seen in Figure 5.6, subjects were much more likely to agree with their customized recommendations than the non-customized recommendations. Subjects chose an activity from the customized

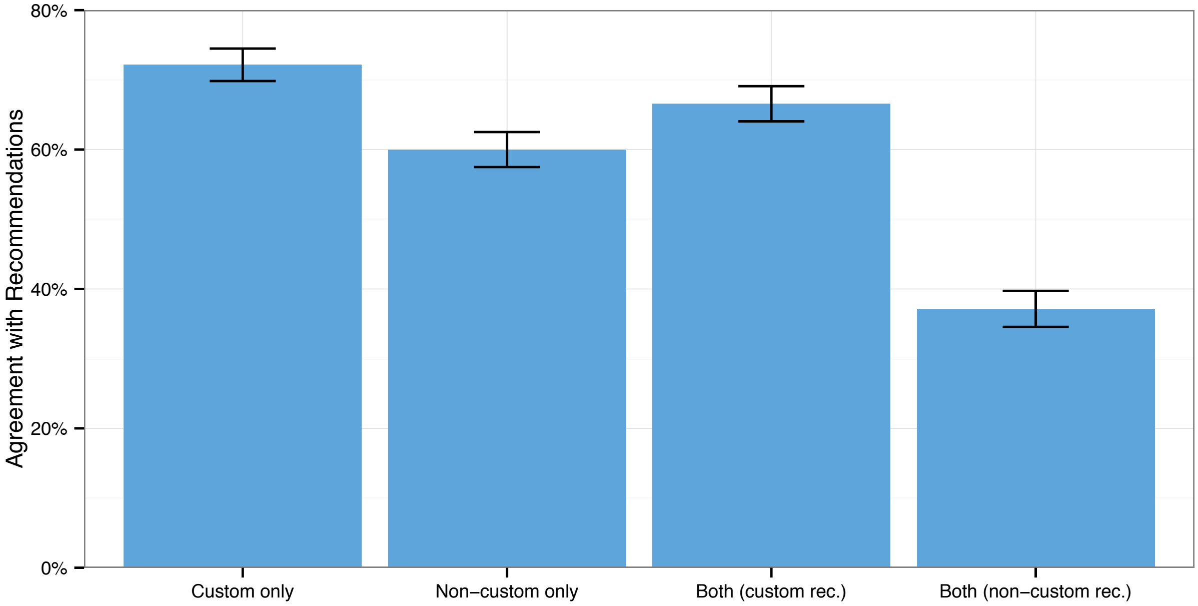


Figure 5.6: Agreement with recommendations.

list in 67% of all rounds, compared to choosing an activity on the non-customized list only 37% of the time. A paired Mann-Whitney U-test confirmed this difference was statistically significant ($p < .001$). It should be noted that in 27% of rounds, the chosen activity appeared on both lists at the same time. When subjects chose an activity that appeared on only one list, 79% of the time it was a custom recommendation and 21% it was from the non-custom list.

Figure 5.7 shows how subjects in each of the IDA conditions adjusted their agreement with the IDA over the course of the ten rounds of the experiment. In this figure, the bolded line represents a moving average smoothed using LOESS (Local Regression) smoothing. The light dotted lines are the actual averages for each condition at each round number.

This figure illustrates that each condition had a distinct pattern of agreement over the ten rounds. Subjects in the Non-custom condition maintained a stationary amount of agreement, with the average barely moving at all and not trending in any direction. Subjects in the Custom only condition started at higher agreement than the Non-custom subjects, and their

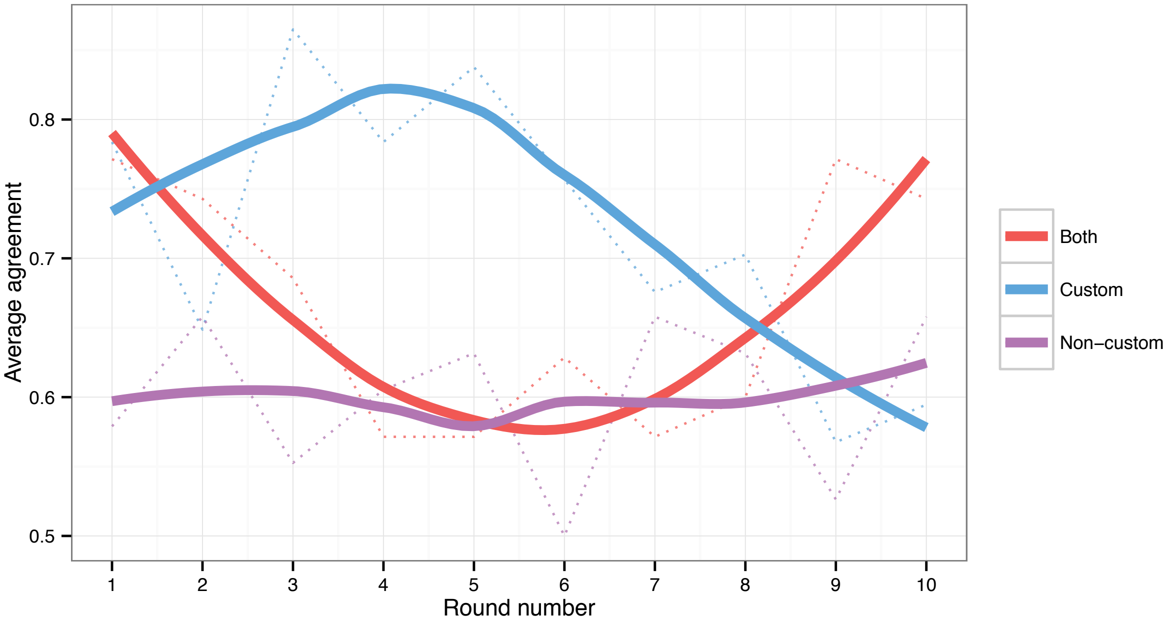


Figure 5.7: Change in average agreement over the 10 rounds of the experiment.

Table 5.10: Transparency measures by condition. Items are on a 5-point scale.

	Non-custom only	Custom only	Both algorithms
1. I understood why the suggestions were made	3.368 (0.942)	3.919 (0.682)	3.829 (0.664)
2. I thought the suggestions made sense	3.289 (1.011)	3.784 (0.672)	3.629 (0.770)
3. The logic behind the recommendations was clear	3.211 (1.094)	3.784 (0.787)	3.600 (0.775)

agreement increased for the first several rounds before beginning a gradual decline. Subjects in the Both Algorithms condition had an opposite pattern from the Custom only group. Their initial agreement was also high, but they immediately showed a decline in trust. However, in the later rounds they regained that trust and by the end of the experiment, these subjects were showing the most agreement with the IDA.

5.4.5 Transparency

Table 5.10 shows the responses to questions related to the transparency of the Exercise Recommender. I conducted an ANOVA for each item followed by Tukey's Post-Hoc tests to test for pairwise differences between conditions. All three ANOVA tests indicated that the model was statistically significant ($p < .05$). Subjects reported less *understanding why the suggestions were made* (Item 1) than both of the other two conditions ($p < .05$) according to the Tukey's test with Honest Significant Differences to account for the multiple comparisons. For items 2 and 3, only the difference between the Non-custom only and the Custom only conditions was statistically significant ($p > .05$). These results support H6 and suggest that customization can serve to increase the transparency of an IDA. These differences are not particularly large however, suggesting that customization may only have a modest effect on transparency.

Table 5.11: Relationship between perceived transparency and agreement with recommendations.

	Dependent variable:		
	Logodds of agreement		
	(1)	(2)	(3)
Item 1 (understood why)	0.401*** (0.143)		
Item 2 (made sense)		0.314** (0.134)	
Item 3 (logic clear)			0.335*** (0.124)
Random Effects Standard Deviation	0.834	0.851	0.837
Log Likelihood	−594.605	−595.604	−594.737
Log Likelihood χ^2	155.85***	153.87***	155.59***
Pseudo R^2	0.260	0.260	0.260
Note:			*p<0.1; **p<0.05; ***p<0.01

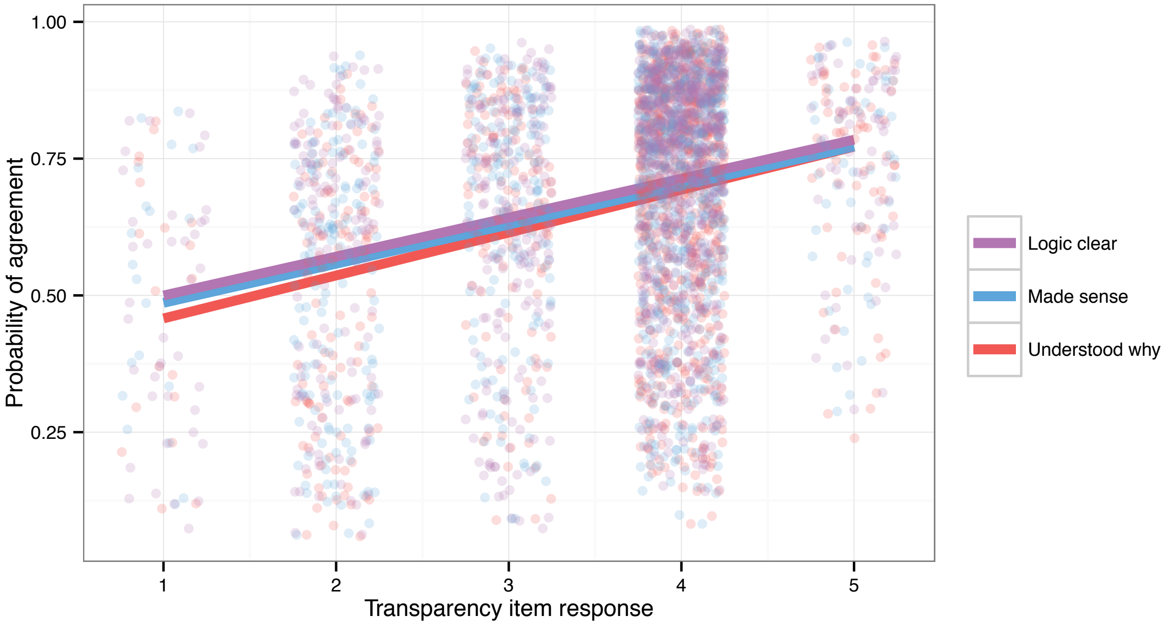


Figure 5.8: Effect of transparency on agreement with recommendations.

These items are not tests of subjects' actual understanding of how the Exercise Recommender works, rather a measure of their perception of its transparency. To evaluate how subjects' perception of transparency affected their agreement with recommendations, I fit three models that estimated the log odds of agreement with recommendations with one of the transparency items as an explanatory variable. These models also included the experiment condition and average recommendation quality as covariates, along with fixed effects for the profiles and a random effect for each subject. I fit a separate model using each of the three items, rather than including them all in one model, because the three items are highly correlated with each other and would create multicollinearity. Similarly, propensity to trust decision aids was moderately correlated with the three transparency items ($r = 0.448$, 0.449, and 0.427) and so it was not included in these models even though in previous models I found it to be a strong estimator of agreement.

Table 5.11 describes these models. Note that this table excludes the other covariates in the model for brevity, as those relationships are reported in Table 5.9. These models indicate that holding all other variables constant, subjects who reported perceiving higher transparency of the Exercise Recommender were more likely to agree with recommendations. The coefficient of each of the three transparency items was statistically significant, even when controlling for the condition the subject was in. From these models, we can conclude that there is support for H7 in that there is an important relationship between users' perception of transparency of an IDA and their likelihood of following its recommendations.

This relationship is important because it is observable after controlling for recommendation quality and for customization. This suggests that a perception of transparency creates an additional decision making bias. Two users with the same preferences and given equivalent recommendations should have equivalent agreement with the recommendations. Yet this model demonstrates that those users who felt the recommendations were more transparent were more likely to follow them.

5.5 Discussion

To summarize the findings of this study:

- There were no differences in decision making quality (points earned) between any of the four conditions. The Exercise Recommender did not help subjects make better decisions, nor was there any effect of the type of recommendation displayed (or the combination of both types).

- Subjects who customized the IDA were more likely to choose one of its suggestions than those who did not customize, even when controlling for the quality of the recommendations. This illustrates that customization bias extends to systems with high controllability.

- The three different IDA interfaces led to different patterns in agreement over time. Non-custom only recommendations led to a stationary pattern, where subjects did

Table 5.12: Summary of results.

	Hypothesis	Result
H1	Users who use the IDA will make better decisions than others who make the decision unaided.	Not supported
H2	Users who see customized recommendations will make better decisions than users who see only non-customized recommendations.	Not supported
H3	Users who see only custom recommendations will have more agreement than users who see only non-custom recommendations.	Supported
H4	Users who see both custom and non-custom recommendations will make the best decisions overall.	Not supported
H5	Users who see both custom and non-custom recommendations will have less agreement with recommendations than users who see only custom recommendations	Not supported
H6	Customizing the system will make users feel the system is more transparent	Supported
H7	The more transparent users feel the system's logic is, the more they will agree with recommendations	Supported

not change their average agreement over time. Custom only recommendations led to initial growth in agreement, followed by a slow decline. Showing both types of recommendation led subjects to slowly reduce their agreement in early rounds, then increase their agreement later on.

- Subjects who used the Non-custom only version of the IDA perceived less transparency than those who had customized the IDA.
- Subjects who reported that they better understood the logic behind the IDA's recommendations were more likely to agree with the recommendations, regardless of which condition they were in.

The lack of any difference in decision quality between the Unaided and the IDA-supported groups was surprising, although it is consistent with many applied studies of IDA effectiveness (Bright et al., 2012). Demonstrating the efficacy of IDAs for improving aggregate measures of decision making has proven to be a challenge, and the results of this study are further evidence that decision aids will not automatically lead to better decisions.

What is apparent, both from this study and other work on decision aids, is that IDAs do alter the decision making process, even though they may not improve it. The strong effect of recommendation quality on decision quality shows that when given a decision aid, users do use it to inform their decisions. Subjects in all the IDA-supported conditions relied on the decision aid to make decisions, and therefore the quality of recommendations was a strong determinant of their score for the round. By giving suggestions for the decision, IDA users must evaluate those suggestions as part of making their decision. Evaluating these suggestions may alter the overall decision process. For example, it may lead to users to evaluate just the suggested alternatives and not consider others, using the IDA as an initial filter. In this case, suggested alternatives may be given greater scrutiny than those alternatives would receive in unaided decision making, and this could make users better discerning of the quality of those suggestions. However, if the system does not recommend the best possible options, users may not consider them, and therefore overall the system would not be optimally effective in relation to unaided decision making. A limitation of this study is that it did not explore decision making strategies or recommendation evaluation process in detail. That is an important direction for future research that would be a valuable contribution to understanding IDA effectiveness and informing design. This study offers support for a hypothesis that IDA-supported and Unaided decision making involve separate processes. Introne and Iandoli (2014) found that there was no relationship between subjects performance in a IDA-supported decision task and the same individual's performance when making the decision unaided. This suggests that there may be separate skills and separate processes that determine whether a person will make good evaluations of recommendations than whether they will make good decisions. The fact that IDA-supported users relied heavily on the recommendations, yet performed no better than unaided subjects, further supports this finding. In order to build more effective IDAs, future research is needed to better understand the processes involved in evaluating IDA output and how that fits in to the overall decision-making process.

This study offers more evidence that the act of customizing an IDA algorithm biases users to accepting its recommendations. Again, subjects who customized the IDA were more likely to follow one of its suggestions than those who received only non-custom recommendations, even when accounting for the quality of the recommendations. This is an important addition to the findings discussed in my previous work (Solomon, 2014) and in Chapter 3, as it shows that this customization bias can happen when using a more responsive IDA than was used in those studies. It also gives evidence against the Illusion of Control as a mechanism of customization bias. As the Illusion of Control can only be observed in cases where there is little actual control (Gino et al., 2011), this study refutes that theory as an explanation. In fact, this study, along with the findings of the previous studies, suggest that customization bias is most likely to happen when the IDA appears to be truly responsive to users' input. This suggests that customization bias may be the result of users feeling that they have successfully exerted control over the system, or that they were able to successfully produce reasonable-looking recommendations.

The different patterns of agreement shown between conditions over time illustrates another difficulty with designing effective IDAs. As customization led to an inverted U-shape pattern, with subjects initially increasing their agreement, followed by a decline later on. If customization were to be used as an effort to increase agreement with recommendations, the effect may be short-lived. It is unclear why users of the system with both recommendation types showed the opposite pattern, where they initially decreased their agreement but later began to agree more often with the IDA. This difference is unclear, but noteworthy because in both conditions subjects customized the system. It suggests that the declining pattern of agreement with customized recommendations can be altered by different system designs.

An important finding from this study is the observed relationship between transparency and agreement with recommendations. The more that users believed that they understand how the system produces recommendations, the more likely they were to agree with them. This is another example of a decision making bias by IDA users. This relationship held true even when accounting for recommendation quality and for the customizability of the system (which I have shown affect both agreement and transparency). Therefore, I argue that this is evidence of a transparency bias. Some users may have been actively trying figure out how the IDA produces recommendations, whereas others may not have tried to actively see into the system or given any thought to how recommendations are produced. Some other possibilities are that different users bring different mental models and experience using these kinds of systems, and they may use that experience or those mental models to infer how the system works, with this experience also making them more or less trusting of the system. The correlation that was found between propensity to trust automated decision aids and perceived transparency offers some evidence that could support this explanation.

It is also possible that users' pre-disposition to trusting or distrusting decision aids may lead them to effectively decide whether they will agree with the IDA before they receive the recommendation, and their evaluation of recommendations could consist of finding reasons to "rationalize" this decision. If they can find a reason to justify their decision, they may claim that the system's recommendations are more clear.

Overall, this finding is consistent with existing literature on transparency (Sinha & Swearingen, 2002; Tintarev & Masthoff, 2008; Cramer et al., 2008). However, an important addition here is that unlike previous studies, transparency was not created by using explanations. In fact there was little effort to create transparency in the design of the IDA, suggesting that there could be an important individual difference in how people interpret the logic behind an IDAs recommendations. Some people may feel they can easily understand recommendations without explicit explanations, whereas others may need explanations or other indicators within the design.

And unlike previous work, the effect of transparency in this study can be considered to be a bias because it was observed even when holding the quality of recommendations and

and the experiment condition constant in the statistical model. Although this is only evidence of an association between transparency and agreement, there is some intuition to a hypothesis that greater transparency would cause greater agreement with recommendations. If users believe they understand why a recommendation was given, it may provide confidence that the recommendation is not the output of a random process but that it represents true insight into the decision. However, a limitation of this study is that it was not designed explicitly determine whether transparency causes agreement. An alternate explanation may be that subjects who agreed more often with recommendations may have tried to justify this agreement when prompted about transparency, even if it did not affect their decisions during the experiment. An important area for future research will be to determine whether transparency causes agreement and whether variations in the design of an IDA affect agreement by way of altering its transparency. If transparency causes users to follow recommendations, it offers designers another target for engineering agreement in IDAs.

As discussed in chapter 3, there is apparent similarity between consistency bias and transparency bias due to the fact that consistency bias depends on the system having some transparency. It is possible that these biases are both manifestations of a single underlying effect. One important consideration however is that in this study, there was an association between transparency and agreement *within* the non-custom only condition. Since consistency bias as described in chapter 3 requires that users have knowledge about the configuration of the system, the finding from the present study would appear to be distinct from that effect because subjects in the non-custom only condition had no information about the configuration of the system.

5.5.1 Limitations

One limitation of this study is that unlike most recommender systems, subjects were not selecting items that match their own preferences but rather to match an assigned preference profile that included only a small number of attributes. Therefore, this decision task may engage a different type of decision making process than what is done by users in naturalistic settings. This study used a decision task that would naturally have a lot of horizontal differentiation in that different users would have widely varying preferences. Some work on recommender systems (Häubl & Murray, 2003) has shown that these systems can not only find items that match users existing preferences, but that the system itself can persuade users to form new preferences. The design of this study does not allow for that aspect of interaction with IDAs because preferences were pre-determined and fixed.

Another limitation of this study is that users were shown their scores between rounds, and then in a post-test were asked about the transparency of the system. One explanation for the transparency bias result is that users saw how well they had performed, and then based their reported understanding of the system's logic on the score they had received. In this case, the transparency measures may primarily reflect users' interpretations of their performance rather than the transparency of the system.

CHAPTER 6

CONCLUSIONS

I have shown that IDAs can impact the decision making process by providing recommendations to users that require some evaluation. IDA users must assess recommendations and determine whether the actions they suggest will meet the goals of their decision task. Giving users good recommendations is a critical factor to obtaining agreement with recommendations, but there are also several biases that affect agreement and are unrelated to recommendation quality. Users are biased towards agreeing with recommendations when they have customized the IDAs logic, even when their customization has had no actual impact on the recommendations. Users are also biased towards agreement when prior to seeing recommendations they have high expectations of the system's efficacy at producing recommendations. The consistency between recommendations and the IDAs configuration also creates a bias, with users being more inclined to agree with recommendations when they appear to be consistent with the way the IDA was configured. And there is also a bias of transparency, with users who reported higher understanding and clarity regarding the way the IDA produced recommendations also being more likely to agree with recommendations than those who reported lower understanding. In this section I will discuss the implications of these findings both for IDA design and for future research on IDAs and computer-supported decision making.

6.1 IDA-Supported Decision Making

When Intelligent Decision Aids provide specific recommendations to decision makers action and alternatives that should be considered, they place a demand on users to evaluate these recommendations before making a decision. I have presented evidence that this requirement

Table 6.1: Decision biases in IDA-supported decision making.

Bias	Description	Evidence
Customization Bias	Users agree more with recommendations when they customize the IDA	Table 3.5, Table 5.9, Figure 3.6, Figure 3.7, Figure 5.6
Expectations Bias	Users agree more with recommendations when they expected it to work well before seeing recommendations	Table 4.5
Consistency Bias	Users agree more with recommendations when they appear to be consistent with the IDAs configuration	Table 3.9, Figure 3.10
Transparency Bias	Users agree more with recommendations when they feel that the IDA's logic is more clear and understandable	Table 5.11

of evaluating recommendations is an important part of IDA-supported decision making that can be influenced by the design of the system such that different system designs can lead to different evaluations of the same recommendations. In particular, the process of generating recommendations and users' beliefs about the IDA's efficacy and their involvement in that process can lead users to respond in different ways to recommendations.

This requirement to evaluate recommendations does not have a clear placement within Parasuraman et al.'s (2000) framework of automation in decision support that was described in Chapter 2 (see Figure **??**). Evaluating recommendations involves some analysis of information and may fit there. However, when considered within the domain of human-machine cooperation, evaluations of recommendations must be performed by the entity (user or machine) that is most responsible for the action selection stage. It makes no sense, for example, for a system to provide recommendations to a user about actions to select if the system will be the one selecting the action. In designs that have a low level of automation at the action selection stage, which is to say systems such as IDAs where a human is primarily responsible for choosing the action, evaluating recommendations should be considered a distinct and critical stage in the decision process (see Figure 6.1).

The results of these studies provide some evidence that this recommendation evaluation

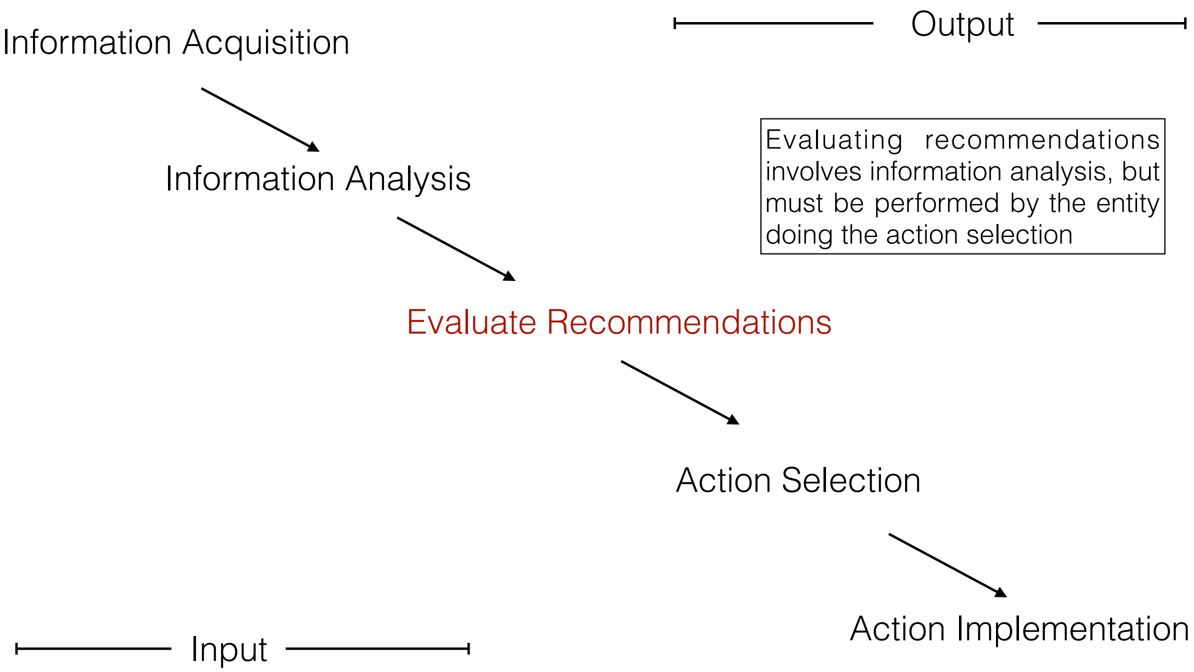

Information Acquisition
Output
Evaluating recommendations involves information analysis, but must be performed by the entity doing the action selection
Information Analysis
Evaluate Recommendations
Action Selection
Action Implementation
Input

Figure 6.1: IDA-supported decision making involves evaluating recommendations, which does not fit neatly into Parasuraman et al.'s framework.

stage is critical. First of all, in all three experiments, the quality of recommendations was one of the most important factors that affected decision performance. The better the recommendations that subjects received from the IDA, the better the decisions that they made. This shows that the recommendations do in fact matter. An argument could be made that recommendations provide just an additional piece of information to consider, and that users might learn something useful from any type of recommendation, even poor ones. While this is possible, the evidence from the studies above suggests that users do tend to follow recommendations, even though, as in the case of the Exercise Recommender study, they may not necessarily benefit from using recommendations.

The transparency bias observed in the Exercise Recommender study showed that users were more likely to agree with recommendations when they had a better idea of how the system's logic produced those recommendations, i.e. when they felt the system was more transparent. And in the baseball studies, users were more likely to agree with recommendations when prior to actually seeing the recommendations, they felt that the logic that would be used was efficacious for producing good recommendations (expectations bias). These findings suggest that evaluating recommendations is a challenging stage and that many factors, from individual differences in trust of automated aids to the design of explanations, customization, or other features unrelated to actual recommendation quality can impact how users evaluate IDA recommendations. In other words, while providing good recommendations is important, it may not always be sufficient to enable good decision making. Users must learn how to calibrate their trust of IDA recommendations with their actual reliability and learn how to identify recommendation quality. An important conclusion from my studies here is that this task is difficult and prone to bias. Characteristics of the recommendations, the system, or the users themselves can lead people to perform poorly at identifying recommendation quality, ultimately leading to poor decisions. An important area for future work will be to develop theory about how users evaluate the quality of recommendations and how these evaluations can be engineered through the design of the system so that users are able to make good decisions about when to follow or not follow recommendations.

6.2 Customization as an IDA Design

The findings from these studies offer some practical design advice regarding the use of customization in IDAs. In Chapter 2, I outlined an argument for using customization in IDAs that was based on the benefits of personalization, transparency, and theories of function allocation. I argued that customization could benefit decision making by improving recommendations, helping users build situational awareness, and increasing the transparency of the IDA. However, I also argued that customization may create some hazards for decision makers, notably that it may make users biased towards their recommendations, it may enable confirmation bias by letting people produce recommendations that support a decision they have already made, or that it may require so much skill to produce recommendations that the users most in need of the decision aid may not possess.

Because recommendation quality is so critical to decision quality in IDA-supported decision making, customization can be very beneficial to systems for which users can improve recommendations through customization. If users' custom recommendations are better than personalized recommendations achieved through intelligent algorithms or collaborative filtering, than customization will benefit decision making for that system. An important part of a user-centered design process for IDAs that are considering using customization is to evaluate how well users perform at producing recommendations compared to the best non-customizable alternatives. If users are unable to produce recommendations of equivalent quality to non-customizable alternatives, than customization will likely be a poor design for that IDA. The importance of recommendation quality in many cases will outweigh any other benefits that may come from customization, such as improved user experience and user acceptance of recommendations, and therefore customization will be very harmful to system effectiveness if it leads to suboptimal recommendations. Likewise, the potential detriments of customization such as biased decision making may be tolerable if overall users are much more effective at producing recommendations in a customizable system than the best non-customizable alternative.

Customization bias is an important consideration for system designers, and should be carefully considered alongside evaluations of recommendation quality as well as evaluations of trust calibration. Systems for which users' reliance on recommendations is inconsistent with the quality of the recommendations that it produces should carefully consider how customization bias might affect the IDA's effectiveness. In a system where users over rely on recommendations and are prone to making commission errors, which is to say prone to following poor recommendations, customization bias should be a strong concern as it is likely to only exacerbate the problem of following poor recommendations. However, systems that suffer from an under-reliance could benefit from customization bias. If users are largely ignoring good recommendations and making errors of omission, the bias created by giving users some control over the system's logic could lead to better decision performance. It is possible even that giving only the perception of control could be beneficial in such cases. Adar, Tan, and Teevan (2013) describe the benefits of "benevolent deception" in user interfaces using examples such as crosswalk signal buttons or thermostat controls that do not actually do anything to the system but which give users a sense of control that benefits them in the context in which the system is being used. Customization may be able to provide a similar benefit to users by increasing buy-in to recommendations to good recommendations, even if users do not change the recommendations or are unable to improve them. In cases where more agreement is needed, this could benefit IDA effectiveness.

Another important consideration for designing customizable IDAs is the feedback that users receive about how their customization has impacted recommendations. In these studies, there was no effort made to provide explicit feedback to users about how their actions had affected the algorithm or the recommendations. However, there was some evidence that users sought out this information and used it in their decision making. The consistency bias described in Chapter 3 showed that when the system appeared to give a recommendation that was consistent with how the IDA had been configured, users were more likely to agree with recommendations. And this was true for users of both the customizable and non-customizable IDAs. It suggests that when users believe that a configuration has been successful, users will be inclined to follow recommendations, including poor recommendations. When users customize an IDA, they are aware of the configuration and may seek evidence in the recommendations that the configuration has been successful or that the recommendations are consistent with the configuration. If they do not appear consistent, users may believe that the system does not work well and potentially ignore good recommendations. Or if the recommendations appear consistent with the configuration, users may misinterpret the success of the algorithm as a good recommendation. System designs should seek out ways to communicate to users how the recommendations tie back specifically to the configurations so users of customizable IDAs can determine the extent of their impact.

6.3 Implications for IDAs in the Wild

Industries and societies are advancing efforts to use artificial intelligence and automation in knowledge work (Carr, 2014), and these efforts mean that IDAs will impact an increasing number of decisions. Designing IDAs that are effective at leading users to make better decisions than they would unaided should be the primary goal for system designers. I have presented several findings in this dissertation about how IDAs affect decision making in a laboratory context. An incorrect diagnosis by a doctor, a poor investment by a banker, or a missed security threat by an analyst could have far more severe consequences than the few dollars that that good decisions were worth to subjects in these studies. Yet if applied to real world setting, we can see that there are very important considerations for IDA-supported decision making that relate to the findings I have presented here.

For example, the clinical IDA DxPlain (Barnett et al., 1987) assists clinicians in making diagnoses by extracting information from a patient's health record and using an extensive database and artificial intelligence to recommend possible diagnoses. It allows users to customize the algorithm by emphasizing certain aspects of the health record, similar to how the customizable IDAs in this dissertation allowed for emphasizing certain categories of baseball statistics or certain attributes of exercise activities. If this affordance of customization creates a bias towards following its recommendations, than patients could be adversely affected by any poor recommendations. Errors in the database or technical limitations of the artificial intelligence could lead to poor diagnoses that are followed as a result of customization bias. And since evaluating the accuracy of medical diagnoses is slow, difficult and expensive, it could be difficult to ever attribute poor health outcomes to the design of DxPlain if in fact such problems existed.

In e-commerce, giving users greater control over the algorithms that provide recommendations could increase sales by eliciting customization bias. If users feel they have greater control over how a system like Amazon or Netflix provides recommendations, it could lead users to trust those recommendation more frequently and purchase more recommended items.

These results may have some application to other types of intelligent systems besides IDAs. Social networking sites like Facebook use algorithms to control and filter what content users see from within their network, and users have widely varying understandings of how this algorithm works and their own role in influencing it (Rader & Gray, 2015). Facebook recently made changes to the news feed interface to allow users to customize what content they see, which Facebook states was intended to better personalize the new feed[1]. The results of this dissertation suggest that as long as users do a good job of personalizing their news feeds for themselves, Facebook may see an increase in engagement with its content that exceeds even any increase that would follow more personalized content. In the studies I have presented here, recommendations were personalized equally as well for users of both customizable and non-customizable systems, yet users of the customizable system followed its recommendations more frequently and closely. If this effect extends to consumption of content on Facebook, it would follow that users will consume more content from their newsfeed because they have customized the algorithm. Systems that enable users to collect and analyze extensive data about themselves like fitness trackers could allow users to configure how data are collected and analyzed. Customization bias might lead users to believe the device is more accurate than it really is because of their customization. Perceived accuracy in quantified self systems or consumption of content in social media may be different from agreement with explicit decision recommendations provided by IDAs. But if the interfaces for interaction with the underlying logic or algorithms are similar, it is reasonable to suspect that other types of intelligent systems can be affected by the biases I have observed in these studies.

The importance of users' beliefs about an IDA's efficacy prior to using the system has implications for many types of IDAs. If clinicians, for example, are not willing to follow a good recommendation based on a belief that the system that produced it uses poor logic or has erroneous information, than the system may be harming clinical decision making and

[1] http://newsroom.fb.com/news/2015/07/updated-controls-for-news-feed/

harming patients. Similarly, investors using an IDA may be persuaded by recommendations to make an investment because they believe in the process that the system used, such as a particular type of statistical model that the investor believes is powerful. This dissertation did not seek to understand clearly how users form their efficacy beliefs, and rather focused on the consequences of those beliefs on agreement with recommendations and on decision making. But it is clear that these consequences are meaningful and therefore an important direction for future research will be to understand how users come to form efficacy beliefs and how appropriate efficacy beliefs– ones that are consistent with the actual efficacy of the IDA– can be engineered.

An important limitation of of the studies in this dissertation is that all users were inexperienced with the IDA they were using. These results may only apply to new users and as they gain experience, potentially over years of using the system, the biases I have observed here may disappear or change.

These findings present another potential challenge for IDAs in application. It is not clear whether the skills that make for a good decision maker in a given context, such as domain knowledge, experience, and sound decision processes, will necessarily translate into skill in customizing an IDA well or into interpreting recommendations. It appears that some degree of literacy with IDAs may be required, and this may place an undesireable burden on users to develop this literacy. For example, the skills and capabilities required to become a good doctor may not be the same capabilities that will enable a person to effectively use a clinical IDA. And developing this literacy might require time and effort that might otherwise be spent sharpening their skills in their practice. This is another important consideration for future research that follows from my work in this dissertation. How difficult is it for people to learn to control complex algorithms given an affordance of customization, and what are the consequences in practice if substantial literacy is required in order for customizable IDAs to be effective?

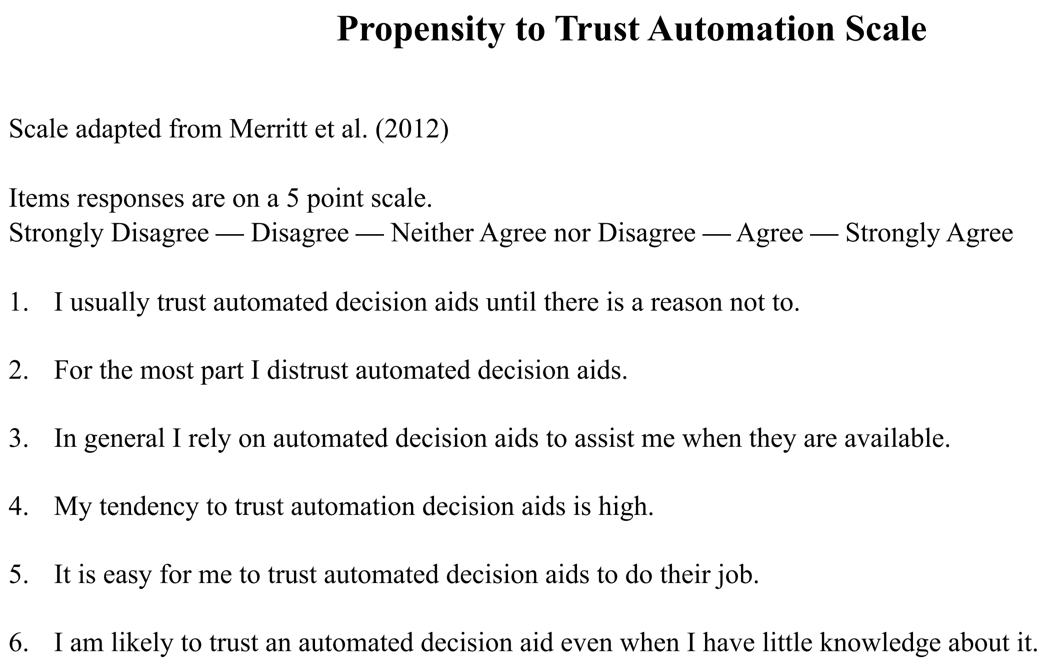
6.4 Conclusion

Intelligent Decision Aids have tremendous potential to improve decision outcomes in many different contexts, but they also present difficult socio-technical challenges as users must learn to interact effectively with powerful but often complex and opaque technologies. Evaluating the recommendations that are produced creates a new type of uncertainty for decision makers, and I have found that users often do not make correct evaluations of IDA recommendations and frequently make biased decisions to follow or not follow recommendations. The design of the IDA, and particularly the use of customization to afford user increased control over these complex technologies, can influence how users navigate this challenge. I believe that IDA designs must account for the way that interaction with the system affects decisions and seek designs that help users learn how to identify recommendation quality and make good decisions about following good recommendations and ignoring poor recommendations.

APPENDICES

APPENDIX A

TRUST PROPENSITY SCALE



Propensity to Trust Automation Scale

Scale adapted from Merritt et al. (2012)

Items responses are on a 5 point scale.
Strongly Disagree — Disagree — Neither Agree nor Disagree — Agree — Strongly Agree

1. I usually trust automated decision aids until there is a reason not to.

2. For the most part I distrust automated decision aids.

3. In general I rely on automated decision aids to assist me when they are available.

4. My tendency to trust automation decision aids is high.

5. It is easy for me to trust automated decision aids to do their job.

6. I am likely to trust an automated decision aid even when I have little knowledge about it.

APPENDIX B

BASEBALL KNOWLEDGE SCREENING QUIZ

1. How many innings are there in a typical Major League Baseball game?
   - a. 7
   - b. 3
   - c. 9
   - d. 14

2. Which of the following best describes a designated hitter?
   - a. A player who bats in place of his team's pitcher
   - b. A player who is substituted into the lineup when a team really needs a hit
   - c. The player who has the best batting average on the team.
   - d. A player who bats in place of any other outfield player on his team.

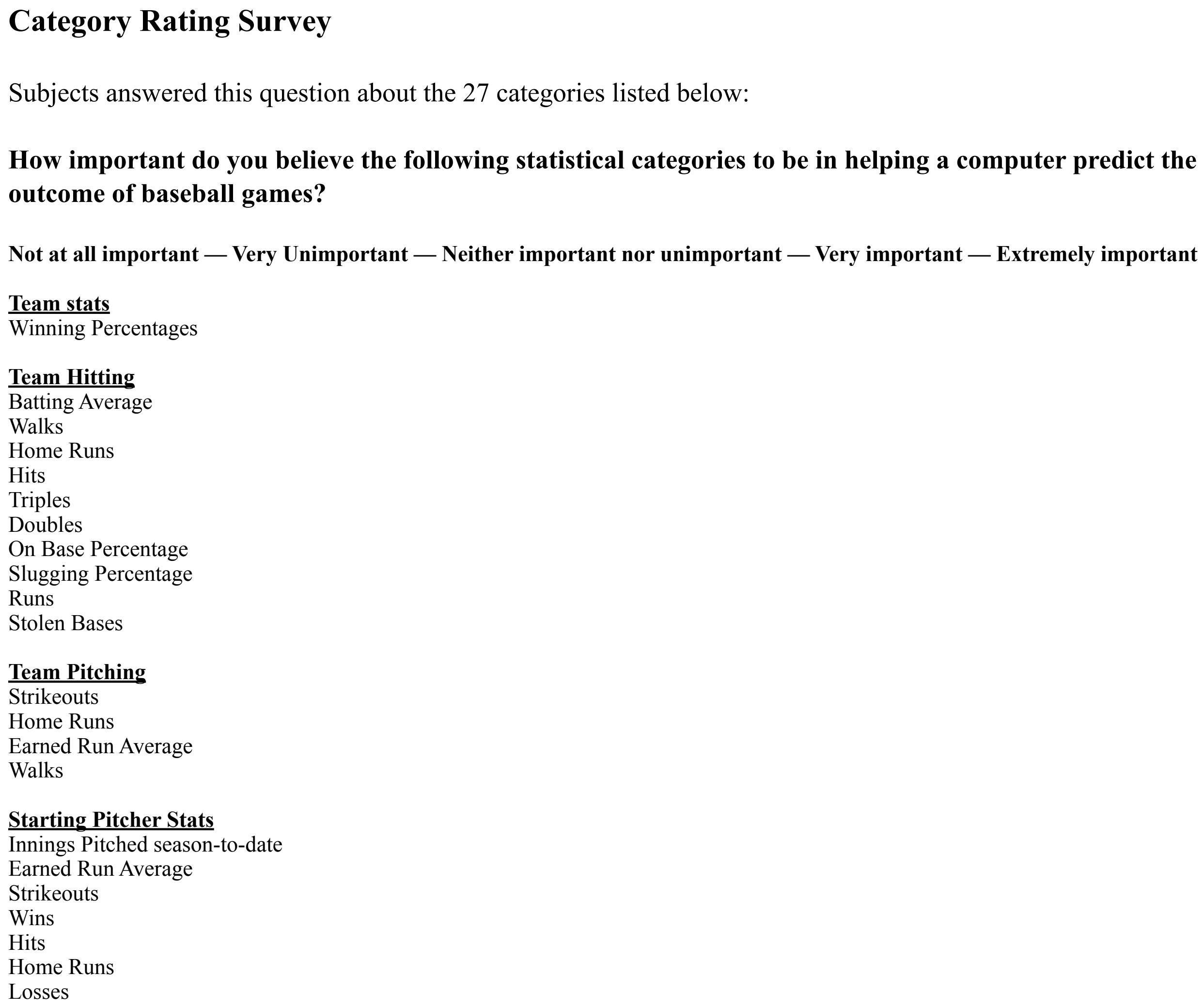
3. Which of the following pitchers has a better Earned Run Average (ERA)?
   - a. Pitcher A 3.13
   - b. Pitcher B 6.49

4. Which of the following best describes a triple in baseball?
   - a. When a player gets three hits in a game
   - b. When a player gets a hit and makes it safely to third base
   - c. When a player gets three hits in one at bat
   - d. When a pitcher throws three strikes to the same batter

5. Which of the following best describes a walk in baseball?
   - a. When a pitcher is replaced and walks to the dugout
   - b. When a batter must walk back to home plate because his hit has gone foul
   - c. When a batter can round the bases at a slow pace because he has hit the ball over the fence.
   - d. When a pitcher throws four balls to a batter so he can walk to first base

APPENDIX C

CATEGORY RATINGS SURVEY



Category Rating Survey

Subjects answered this question about the 27 categories listed below:

How important do you believe the following statistical categories to be in helping a computer predict the outcome of baseball games?

Not at all important — Very Unimportant — Neither important nor unimportant — Very important — Extremely important

Team stats
Winning Percentages

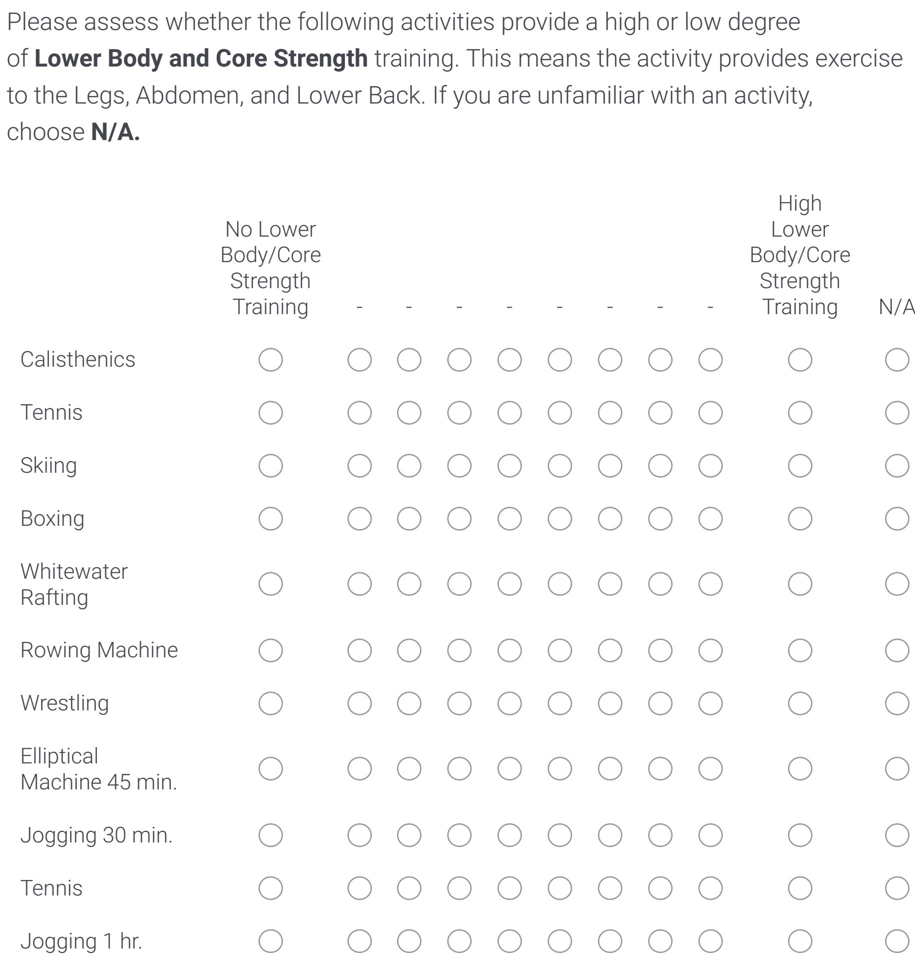
Team Hitting
Batting Average
Walks
Home Runs
Hits
Triples
Doubles
On Base Percentage
Slugging Percentage
Runs
Stolen Bases

Team Pitching
Strikeouts
Home Runs
Earned Run Average
Walks

Starting Pitcher Stats
Innings Pitched season-to-date
Earned Run Average
Strikeouts
Wins
Hits
Home Runs
Losses

APPENDIX D

EXERCISE RECOMMENDER SEED DATA SURVEY EXAMPLE



Please assess whether the following activities provide a high or low degree of **Lower Body and Core Strength** training. This means the activity provides exercise to the Legs, Abdomen, and Lower Back. If you are unfamiliar with an activity, choose **N/A.**

	No Lower Body/Core Strength Training	-	-	-	-	-	-	-	-	High Lower Body/Core Strength Training	N/A
Calisthenics	○	○	○	○	○	○	○	○	○	○	○
Tennis	○	○	○	○	○	○	○	○	○	○	○
Skiing	○	○	○	○	○	○	○	○	○	○	○
Boxing	○	○	○	○	○	○	○	○	○	○	○
Whitewater Rafting	○	○	○	○	○	○	○	○	○	○	○
Rowing Machine	○	○	○	○	○	○	○	○	○	○	○
Wrestling	○	○	○	○	○	○	○	○	○	○	○
Elliptical Machine 45 min.	○	○	○	○	○	○	○	○	○	○	○
Jogging 30 min.	○	○	○	○	○	○	○	○	○	○	○
Tennis	○	○	○	○	○	○	○	○	○	○	○
Jogging 1 hr.	○	○	○	○	○	○	○	○	○	○	○

APPENDIX E

EXERCISE ACTIVITIES AND LATENT FACTOR SCORES

	Activity	Workout Intensity	Workout Atmosphere	Muscle Group
1	Basketball	1.469	0.352	-0.125
2	Bench Press	0.889	-0.526	2.324
3	Bicep Curls	-0.701	-1.250	2.246
4	Biking 1 Hr	0.406	0.438	-1.375
5	Bowling	-2.752	1.129	0.706
6	Boxing	1.324	0.072	1.383
7	Calisthenics	0.800	-0.804	-1.163
8	Canoeing	-0.074	0.620	0.860
9	Curling	-0.990	-1.043	0.871
10	Deadlift	0.402	-0.846	-1.541
11	Diving	-1.369	0.142	0.204
12	Dumbell Fly	0.418	-1.040	1.469
13	Elliptical Machine	0.283	-1.085	-0.776
14	Golf	-1.368	0.011	-0.265
15	Hiking	1.057	0.462	-0.441
16	Inline Skating	0.191	1.157	-1.875
17	Jogging 1 Hr	1.082	-1.003	-1.199
18	Jogging 30 Min	0.114	-1.325	-1.793
19	Jumprope	1.346	-1.406	0.111

Table E.1 Exercise ratings factors scores

Table E.1 (cont'd)

20	Lunges	-0.209	-1.781	-1.337
21	Medicine Ball	-0.708	-0.819	0.301
22	Mountain Climbing	0.141	1.775	1.370
23	Paddle Boarding	0.086	0.212	1.367
24	Pilates	-1.040	0.323	0.334
25	Planks	-0.379	-1.496	1.024
26	Plyometrics	0.979	-0.589	1.491
27	Pushups	1.038	-1.170	1.897
28	Rock Climbing	1.485	0.967	0.871
29	Rowing Machine	0.330	-0.883	0.831
30	Scuba Diving	-0.695	2.038	0.530
31	Shoulder Press	-1.644	-0.618	1.454
32	Skiing	0.954	0.793	0.185
33	Snorkeling	-1.354	1.937	-1.083
34	Soccer	1.056	1.274	-1.564
35	Square Dancing	-0.804	1.294	-2.106
36	Squats	0.513	-1.651	-1.293
37	Stairs 30 Min	1.396	-2.013	-1.953
38	Stretching	-3.328	-1.691	-0.534
39	Surfing	0.469	1.746	-0.446
40	Swimming	1.022	0.542	0.425
41	Table Tennis	-1.659	0.851	-0.113
42	Tennis	0.327	1.028	-0.292
43	Ultimate Frisbee	0.097	0.485	-0.027
44	Volleyball	0.547	1.974	-0.324

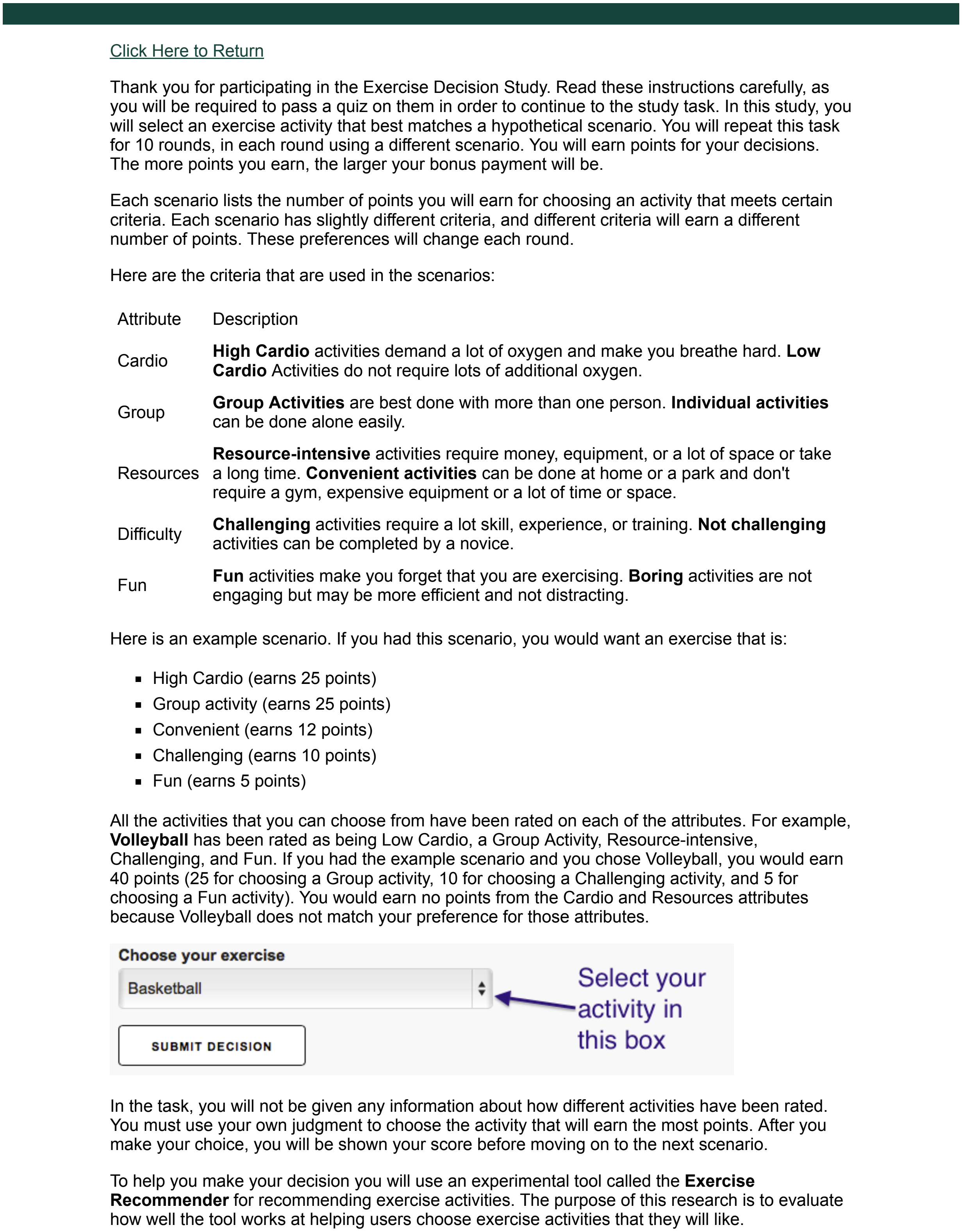
Table E.1 (cont'd)

45	Walking 1 Hr	-2.241	0.998	-1.415
46	Wallsits	0.052	-1.532	-0.985
47	Whitewater Rafting	0.477	1.555	1.246
48	Wrestling	1.006	-0.086	0.724
49	Yoga	-1.085	-1.140	0.115
50	Zumba	0.647	1.624	-0.314

APPENDIX F

EXERCISE RECOMMENDER STUDY INSTRUCTIONS

Figure F.1: Exercise recommender instructions



Click Here to Return

Thank you for participating in the Exercise Decision Study. Read these instructions carefully, as you will be required to pass a quiz on them in order to continue to the study task. In this study, you will select an exercise activity that best matches a hypothetical scenario. You will repeat this task for 10 rounds, in each round using a different scenario. You will earn points for your decisions. The more points you earn, the larger your bonus payment will be.

Each scenario lists the number of points you will earn for choosing an activity that meets certain criteria. Each scenario has slightly different criteria, and different criteria will earn a different number of points. These preferences will change each round.

Here are the criteria that are used in the scenarios:

Attribute	Description
Cardio	**High Cardio** activities demand a lot of oxygen and make you breathe hard. **Low Cardio** Activities do not require lots of additional oxygen.
Group	**Group Activities** are best done with more than one person. **Individual activities** can be done alone easily.
Resources	**Resource-intensive** activities require money, equipment, or a lot of space or take a long time. **Convenient activities** can be done at home or a park and don't require a gym, expensive equipment or a lot of time or space.
Difficulty	**Challenging** activities require a lot skill, experience, or training. **Not challenging** activities can be completed by a novice.
Fun	**Fun** activities make you forget that you are exercising. **Boring** activities are not engaging but may be more efficient and not distracting.

Here is an example scenario. If you had this scenario, you would want an exercise that is:

- High Cardio (earns 25 points)
- Group activity (earns 25 points)
- Convenient (earns 12 points)
- Challenging (earns 10 points)
- Fun (earns 5 points)

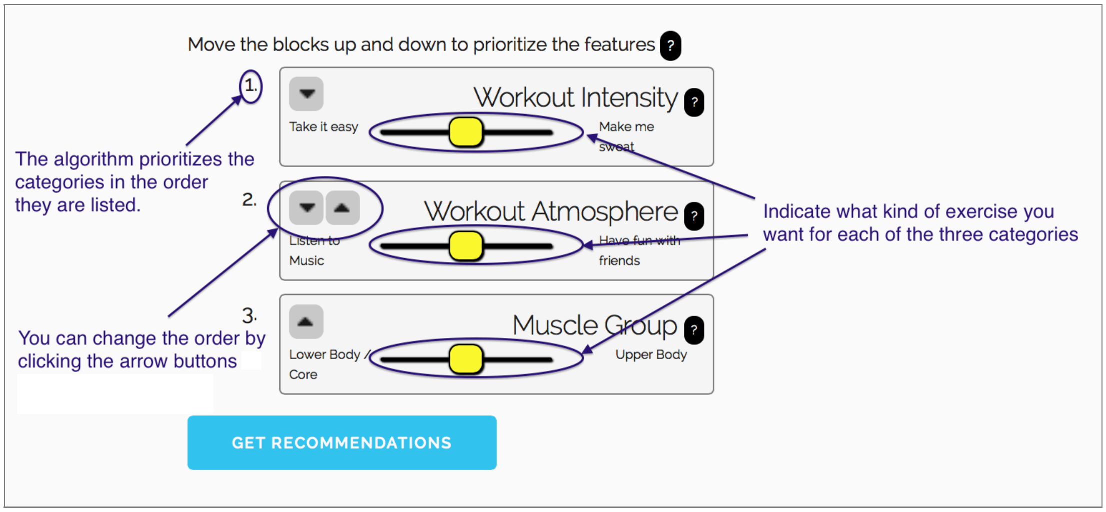
All the activities that you can choose from have been rated on each of the attributes. For example, **Volleyball** has been rated as being Low Cardio, a Group Activity, Resource-intensive, Challenging, and Fun. If you had the example scenario and you chose Volleyball, you would earn 40 points (25 for choosing a Group activity, 10 for choosing a Challenging activity, and 5 for choosing a Fun activity). You would earn no points from the Cardio and Resources attributes because Volleyball does not match your preference for those attributes.

In the task, you will not be given any information about how different activities have been rated. You must use your own judgment to choose the activity that will earn the most points. After you make your choice, you will be shown your score before moving on to the next scenario.

To help you make your decision you will use an experimental tool called the **Exercise Recommender** for recommending exercise activities. The purpose of this research is to evaluate how well the tool works at helping users choose exercise activities that they will like.

Figure F.1 (cont'd)

This system will provide some recommendations for exercises that you can select. **You do not have to choose one of the activities. You are free to choose any activity even if the Exercise Recommender has not suggested it.**



Move the blocks up and down to prioritize the features ?

The algorithm prioritizes the categories in the order they are listed.

You can change the order by clicking the arrow buttons

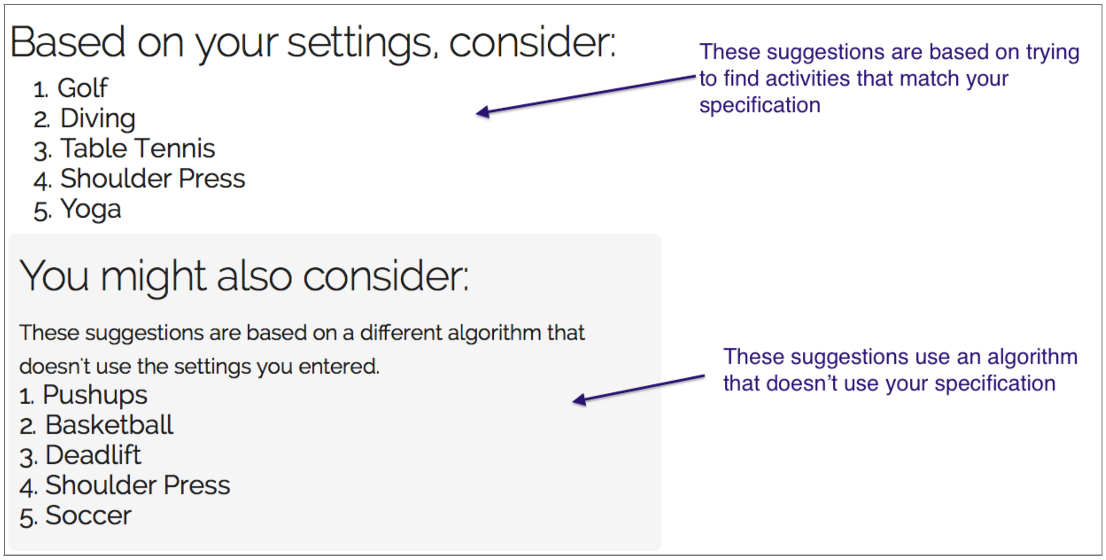
Indicate what kind of exercise you want for each of the three categories

You will be able to adjust the settings of the Exercise Recommender to help its algorithm suggest activities for you. You can adjust three settings:

- **Workout Intensity** is how much will the activity make you work hard, breathe hard and sweat.
- **Workout Atmosphere** looks for either Fun social recreation activities or solitary workouts depending on your setting
- **Muscle Group** looks for activities that focus more on upper body muscles or an lower body and core muscles.

The system will prioritize each setting category in the order they are listed. If Muscle Group is listed at number 1, the Exercise Recommender will try hardest suggest activities that match your specification for Muscle Group. You can adjust the priority order by clicking the arrow buttons to move the tiles up and down, or by dragging tiles up and down.

Figure F.1 (cont'd)



Based on your settings, consider:

1. Golf
2. Diving
3. Table Tennis
4. Shoulder Press
5. Yoga

These suggestions are based on trying to find activities that match your specification

You might also consider:

These suggestions are based on a different algorithm that doesn't use the settings you entered.

1. Pushups
2. Basketball
3. Deadlift
4. Shoulder Press
5. Soccer

These suggestions use an algorithm that doesn't use your specification

The system will provide two sets of recommendations. One set is based on your configured settings and the priority order you have specified. These suggestions apear on top.

The second set of recommendations is based on a different experimental algorithm. This algorithm does not use your settings to produce its suggestions. Both of the algorithms can produce good recommendations, but sometimes they might produce different results. You should evaluate and consider all recommendations when making your decision. Again, you do not have to choose an activity that has been suggested by the Exercise Recommender.

Important Points:

- Each round, after looking over your preferences, click the link that says "Load Exercise Recommender" to open up the system to help you make your decision. You must load the Exercise Recommender and generate recommendations before you can select an activity.
- The Exercise Recommender does not have any information about your points. Its suggestions are based entirely on the settings you enter or on an intelligent algorithm that does not know either your settings or your point preferences.
- You can only generate recommendations one time per round. Be sure your settings are correct when you click the "Generate Recommendations" button.
- Reminder- You are free to choose an activity that has not been suggested. Your objective is to choose the activity that **you** believe will score the most possible points, regardless of what the system suggests. The system works well, but it may not always suggest the best possible activity.

Be sure you fully understand the instructions. Before beginning the task, you must pass a quiz on these instructions

Click here to continue to the quiz

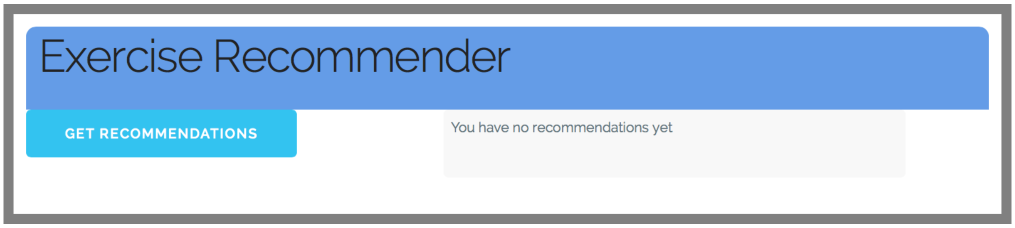
If you have already passed the quiz, click here to return to the game

APPENDIX G

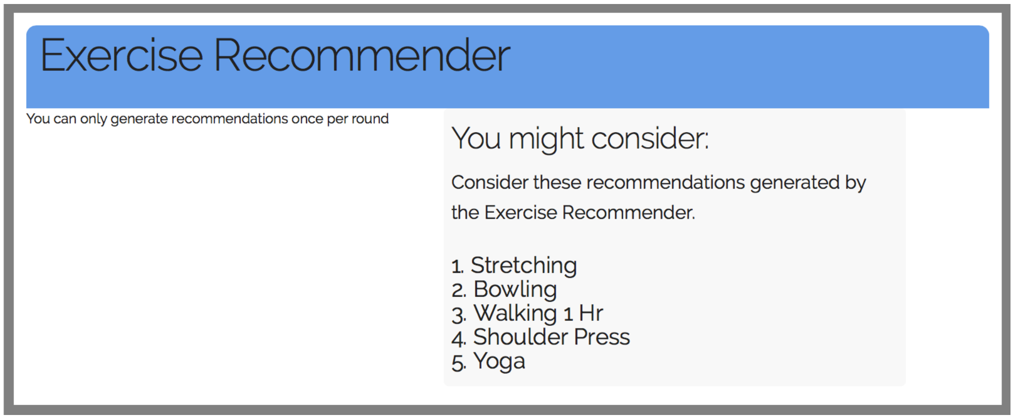
EXERCISE RECOMMENDER INTERFACE

Figure G.1: Exercise recommender interface

Non-custom only interface



You have no recommendations yet



You can only generate recommendations once per round

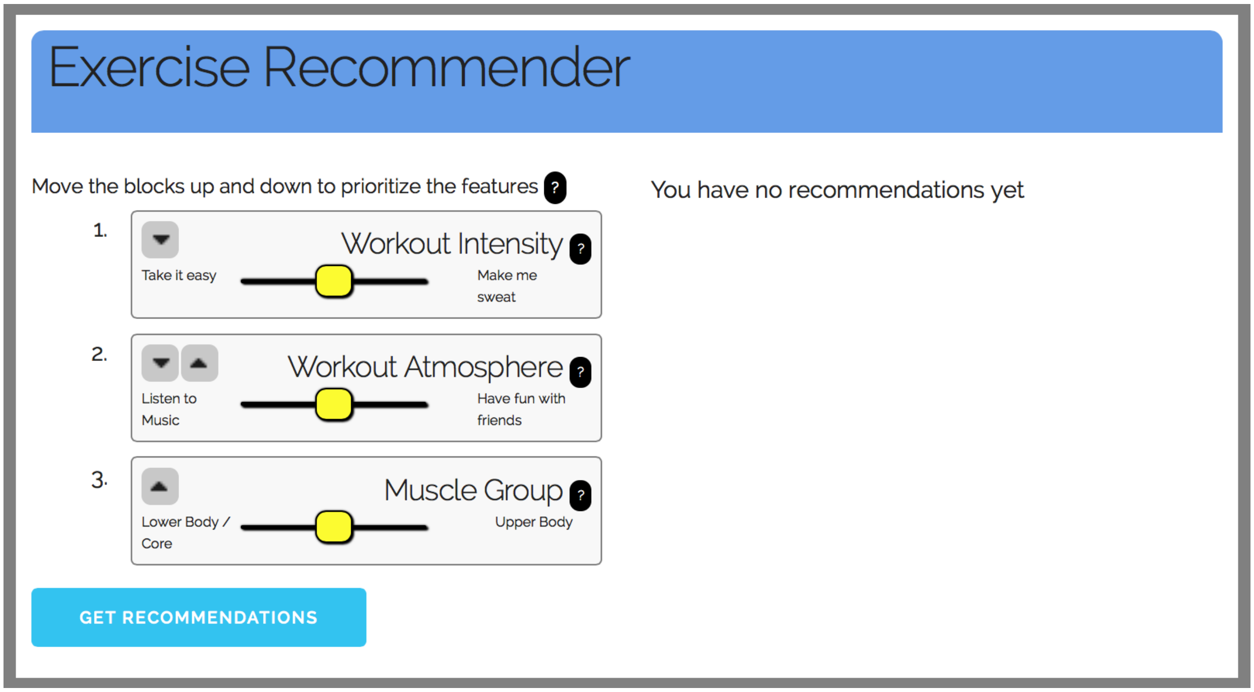
You might consider:

Consider these recommendations generated by the Exercise Recommender.

1. Stretching
2. Bowling
3. Walking 1 Hr
4. Shoulder Press
5. Yoga

Figure G.1 (cont'd).

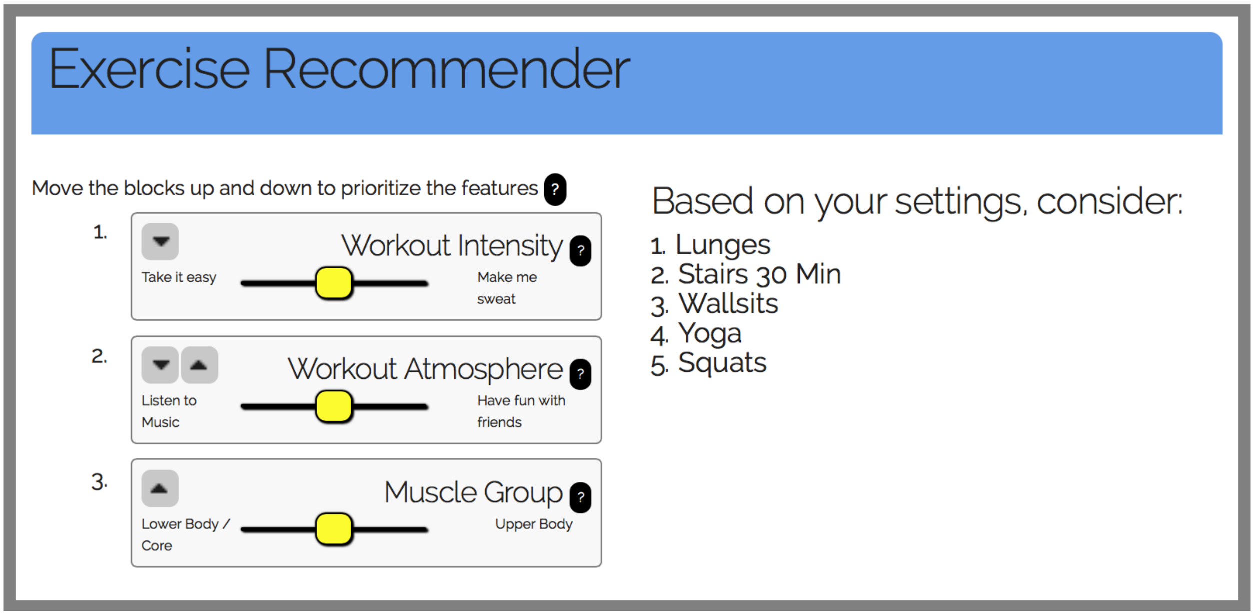
Custom only interface



Exercise Recommender

Move the blocks up and down to prioritize the features ?

You have no recommendations yet



Exercise Recommender

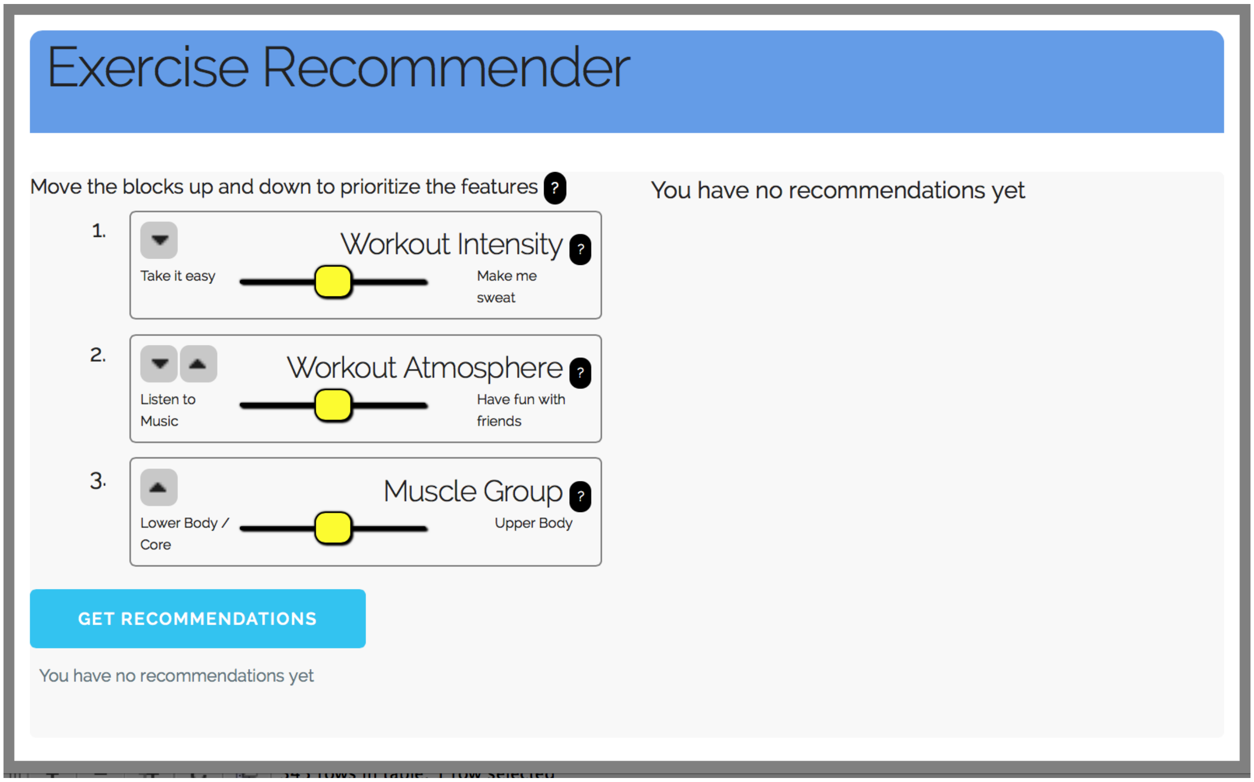
Move the blocks up and down to prioritize the features ?

Based on your settings, consider:

1. Lunges
2. Stairs 30 Min
3. Wallsits
4. Yoga
5. Squats

Figure G.1 (cont'd).

Both algorithms interface

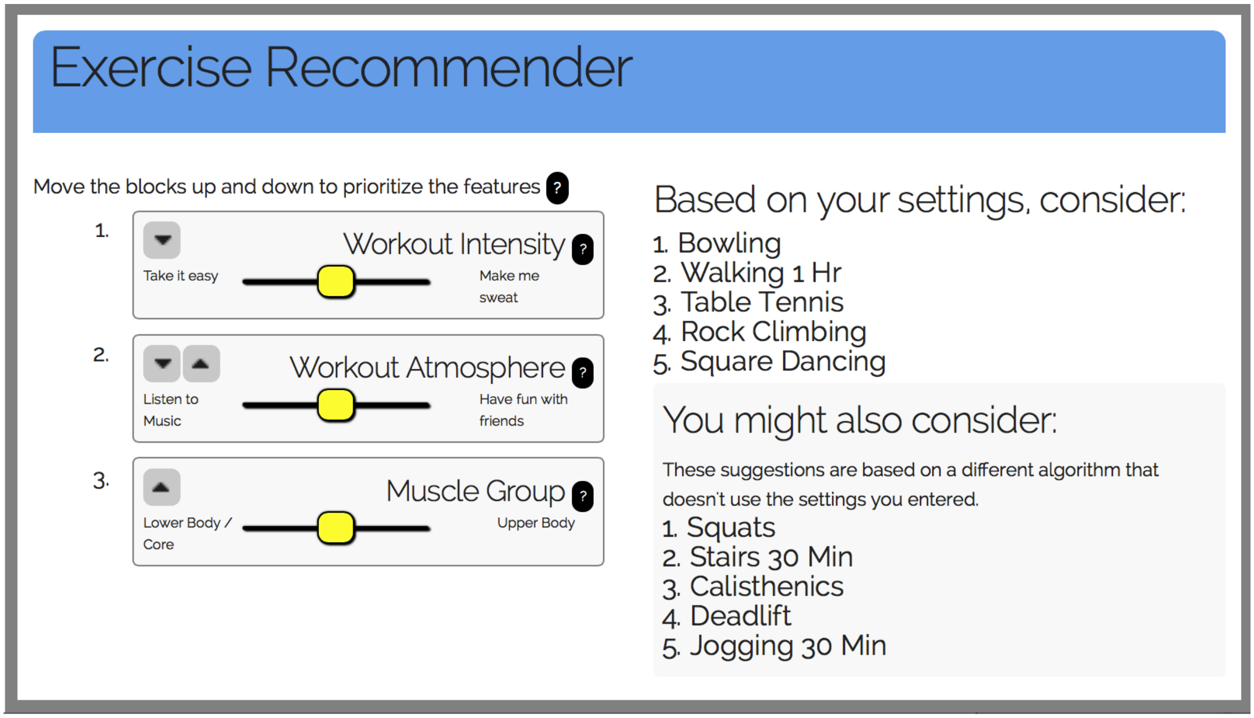


Exercise Recommender

Move the blocks up and down to prioritize the features ?

You have no recommendations yet

You have no recommendations yet



Exercise Recommender

Move the blocks up and down to prioritize the features ?

Based on your settings, consider:

1. Bowling
2. Walking 1 Hr
3. Table Tennis
4. Rock Climbing
5. Square Dancing

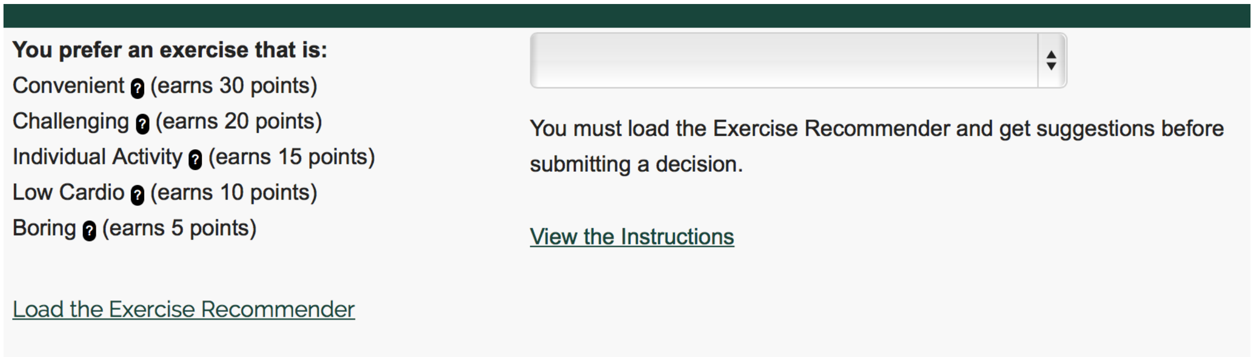
You might also consider:

These suggestions are based on a different algorithm that doesn't use the settings you entered.

1. Squats
2. Stairs 30 Min
3. Calisthenics
4. Deadlift
5. Jogging 30 Min

Figure G.1 (cont'd).

Experiment Interface



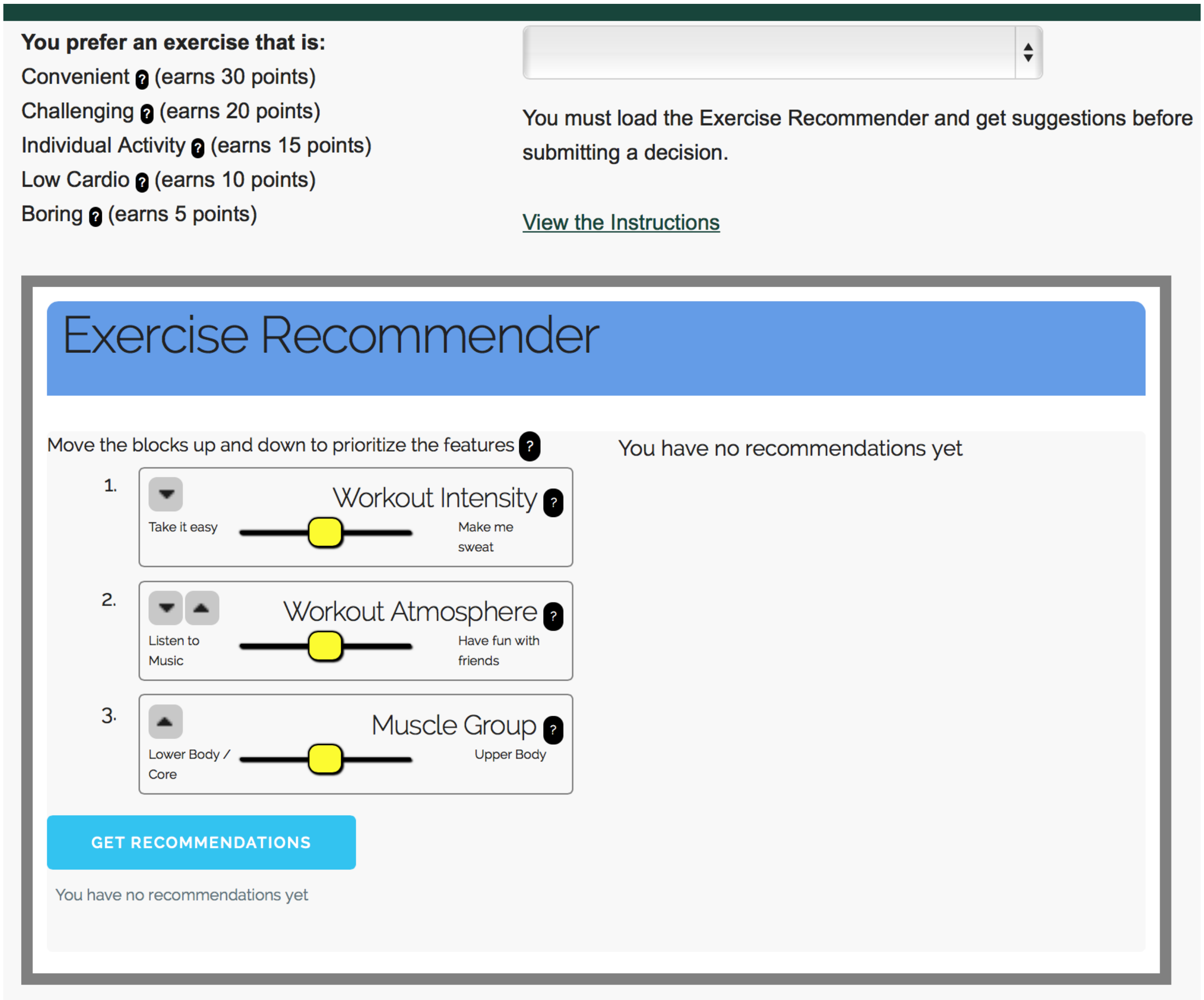
You prefer an exercise that is:

Convenient ? (earns 30 points)
Challenging ? (earns 20 points)
Individual Activity ? (earns 15 points)
Low Cardio ? (earns 10 points)
Boring ? (earns 5 points)

Load the Exercise Recommender

You must load the Exercise Recommender and get suggestions before submitting a decision.

View the Instructions



You prefer an exercise that is:

Convenient ? (earns 30 points)
Challenging ? (earns 20 points)
Individual Activity ? (earns 15 points)
Low Cardio ? (earns 10 points)
Boring ? (earns 5 points)

You must load the Exercise Recommender and get suggestions before submitting a decision.

View the Instructions

Exercise Recommender

Move the blocks up and down to prioritize the features ?

1. Workout Intensity ? — Take it easy / Make me sweat

2. Workout Atmosphere ? — Listen to Music / Have fun with friends

3. Muscle Group ? — Lower Body / Core / Upper Body

GET RECOMMENDATIONS

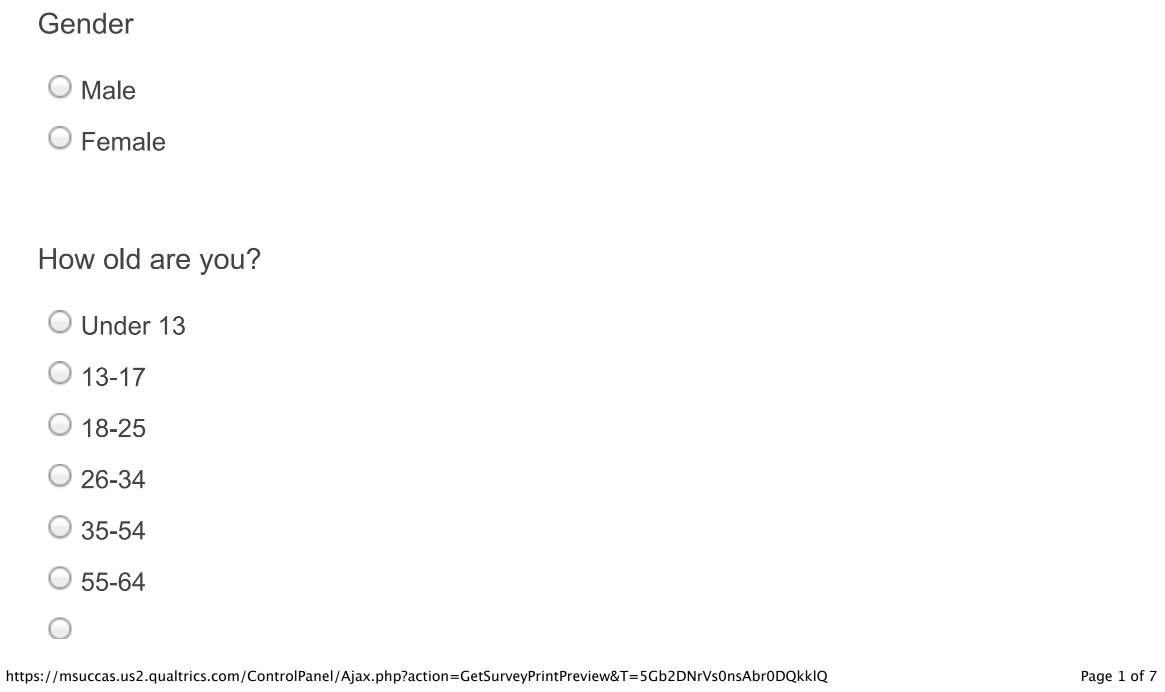
You have no recommendations yet

You have no recommendations yet

APPENDIX H

POST-TEST QUESTIONNAIRE FOR EXERCISE RECOMMENDER STUDY

Figure H.1: Exercise recommender post-test survey.



Gender

○ Male

○ Female

How old are you?

○ Under 13

○ 13-17

○ 18-25

○ 26-34

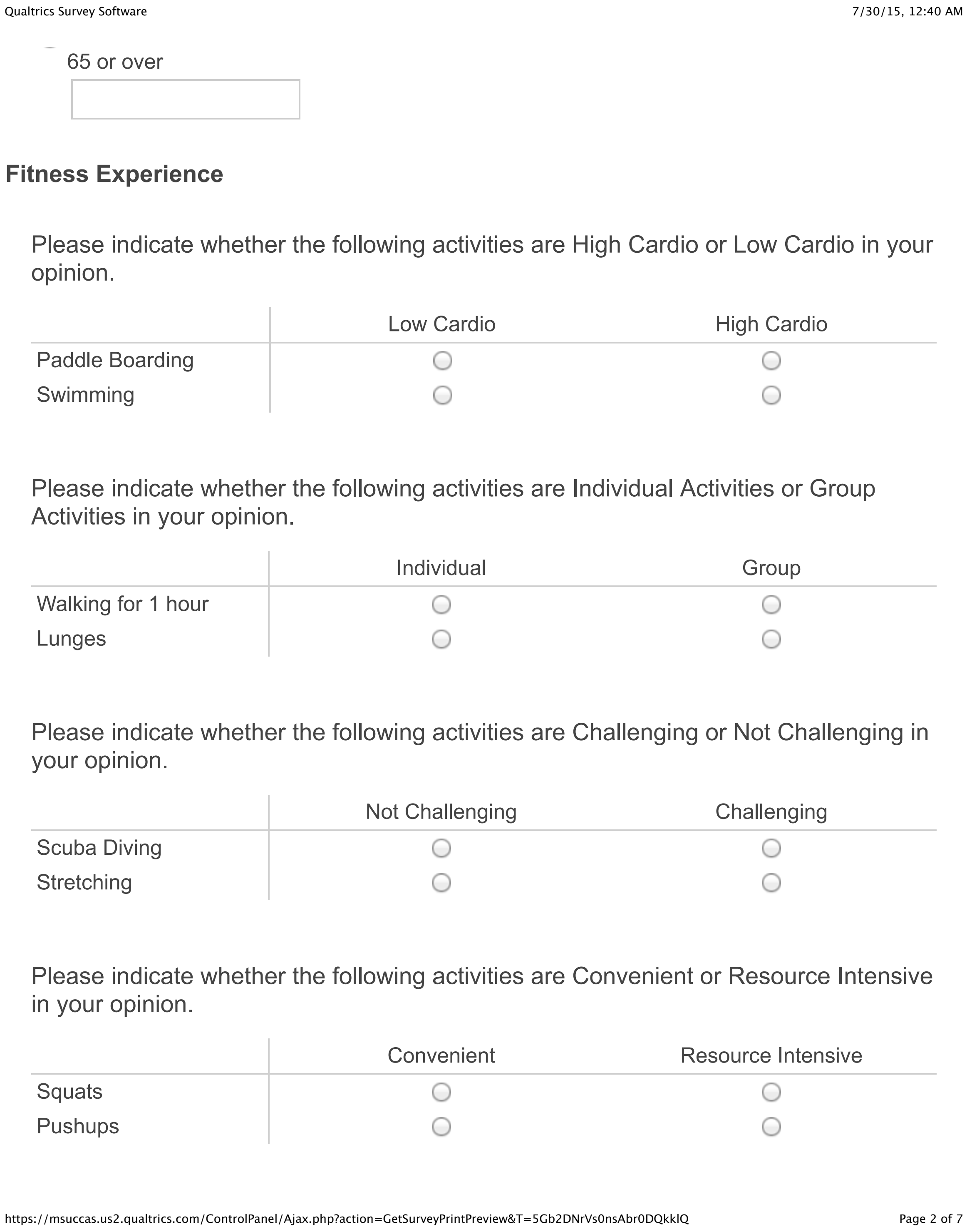
○ 35-54

○ 55-64

○

https://msuccas.us2.qualtrics.com/ControlPanel/Ajax.php?action=GetSurveyPrintPreview&T=5Gb2DNrVs0nsAbr0DQkklQ Page 1 of 7

Figure H.1 (cont'd).



Qualtrics Survey Software 7/30/15, 12:40 AM

65 or over

Fitness Experience

Please indicate whether the following activities are High Cardio or Low Cardio in your opinion.

	Low Cardio	High Cardio
Paddle Boarding	○	○
Swimming	○	○

Please indicate whether the following activities are Individual Activities or Group Activities in your opinion.

	Individual	Group
Walking for 1 hour	○	○
Lunges	○	○

Please indicate whether the following activities are Challenging or Not Challenging in your opinion.

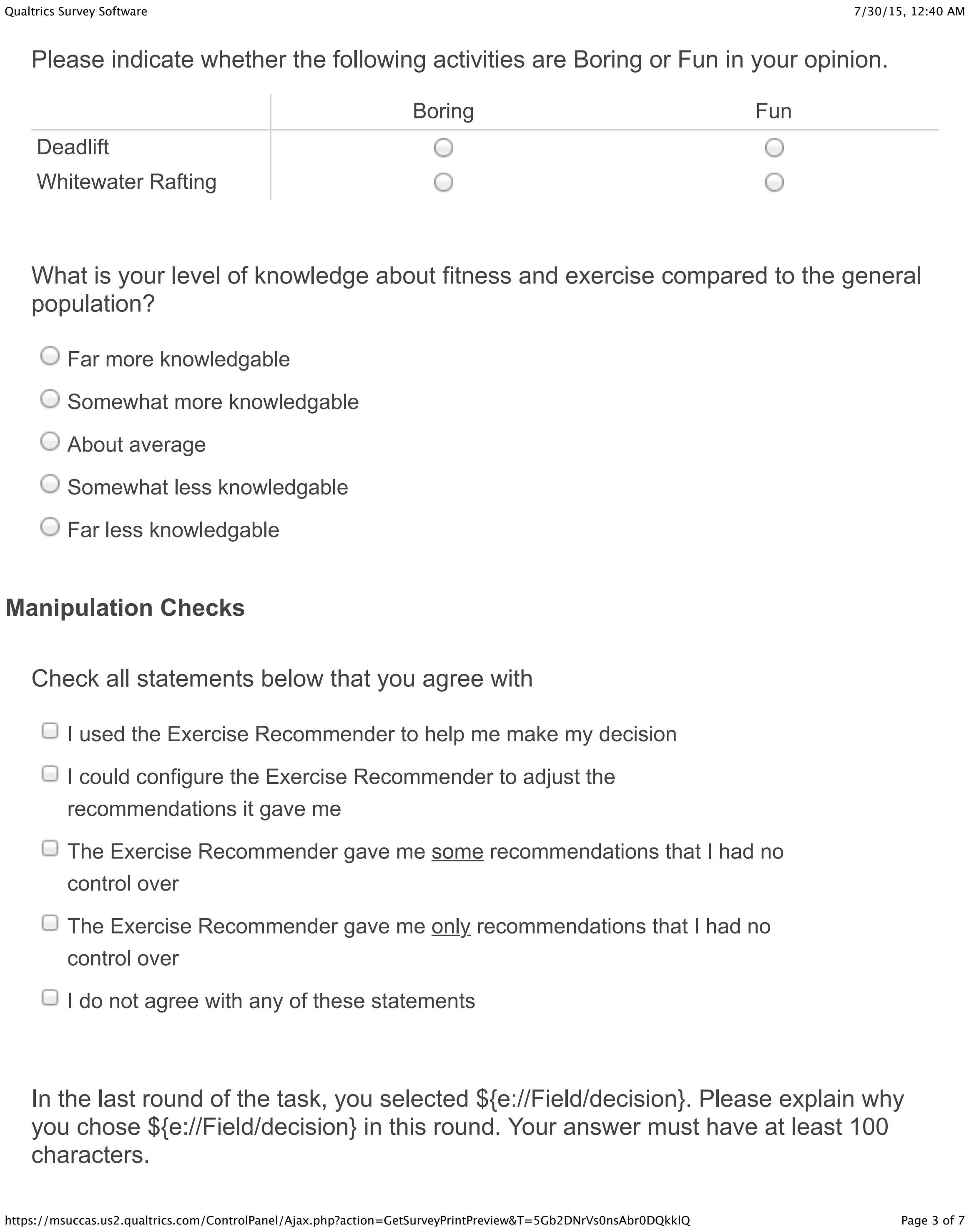
	Not Challenging	Challenging
Scuba Diving	○	○
Stretching	○	○

Please indicate whether the following activities are Convenient or Resource Intensive in your opinion.

	Convenient	Resource Intensive
Squats	○	○
Pushups	○	○

https://msuccas.us2.qualtrics.com/ControlPanel/Ajax.php?action=GetSurveyPrintPreview&T=5Gb2DNrVs0nsAbr0DQkklQ Page 2 of 7

Figure H.1 (cont'd).



Qualtrics Survey Software 7/30/15, 12:40 AM

Please indicate whether the following activities are Boring or Fun in your opinion.

	Boring	Fun
Deadlift	○	○
Whitewater Rafting	○	○

What is your level of knowledge about fitness and exercise compared to the general population?

○ Far more knowledgable

○ Somewhat more knowledgable

○ About average

○ Somewhat less knowledgable

○ Far less knowledgable

Manipulation Checks

Check all statements below that you agree with

□ I used the Exercise Recommender to help me make my decision

□ I could configure the Exercise Recommender to adjust the recommendations it gave me

□ The Exercise Recommender gave me <u>some</u> recommendations that I had no control over

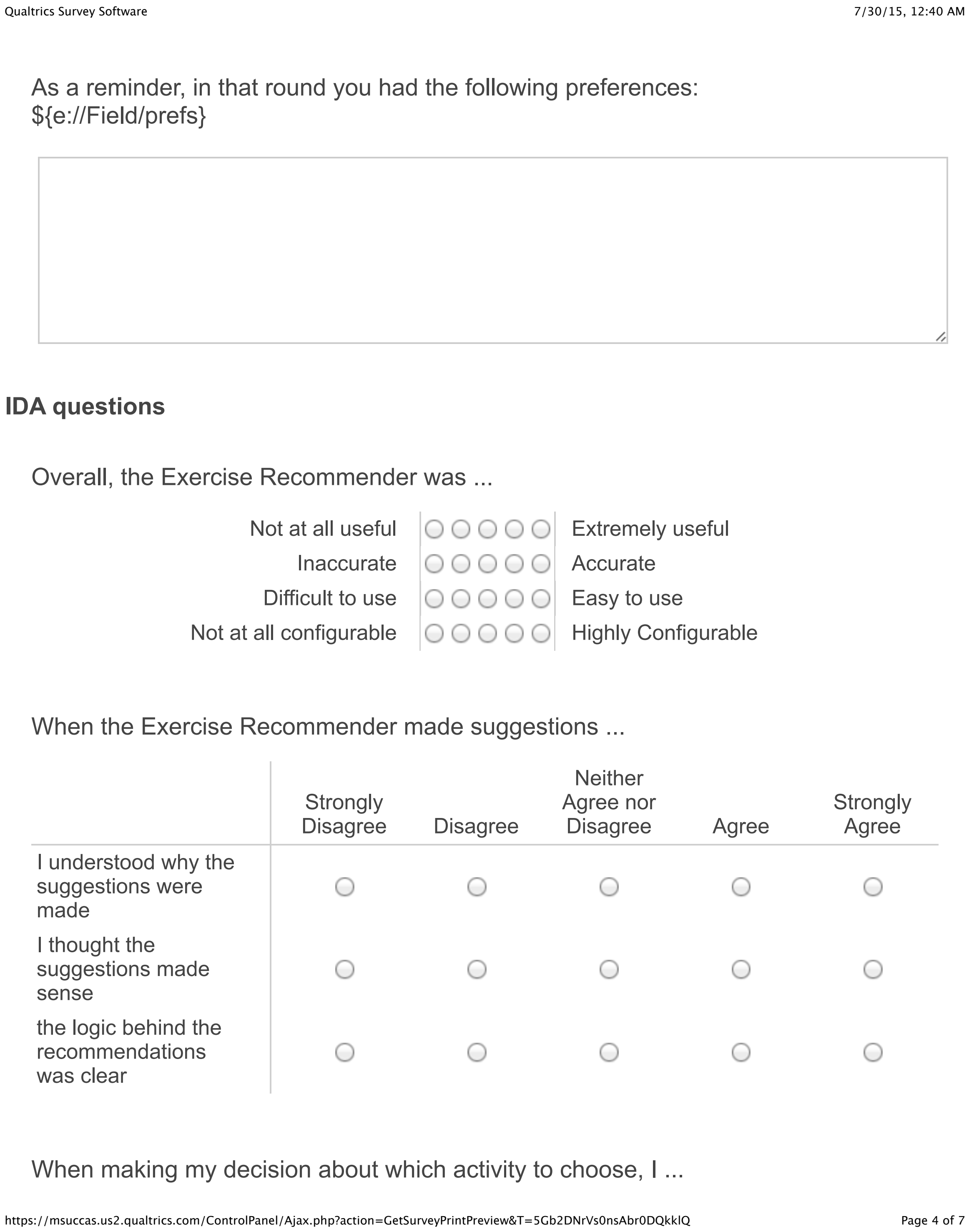
□ The Exercise Recommender gave me <u>only</u> recommendations that I had no control over

□ I do not agree with any of these statements

In the last round of the task, you selected ${e://Field/decision}. Please explain why you chose ${e://Field/decision} in this round. Your answer must have at least 100 characters.

https://msuccas.us2.qualtrics.com/ControlPanel/Ajax.php?action=GetSurveyPrintPreview&T=5Gb2DNrVs0nsAbr0DQkklQ Page 3 of 7

Figure H.1 (cont'd).



Qualtrics Survey Software 7/30/15, 12:40 AM

As a reminder, in that round you had the following preferences:
${e://Field/prefs}

IDA questions

Overall, the Exercise Recommender was ...

Not at all useful	○ ○ ○ ○ ○	Extremely useful
Inaccurate	○ ○ ○ ○ ○	Accurate
Difficult to use	○ ○ ○ ○ ○	Easy to use
Not at all configurable	○ ○ ○ ○ ○	Highly Configurable

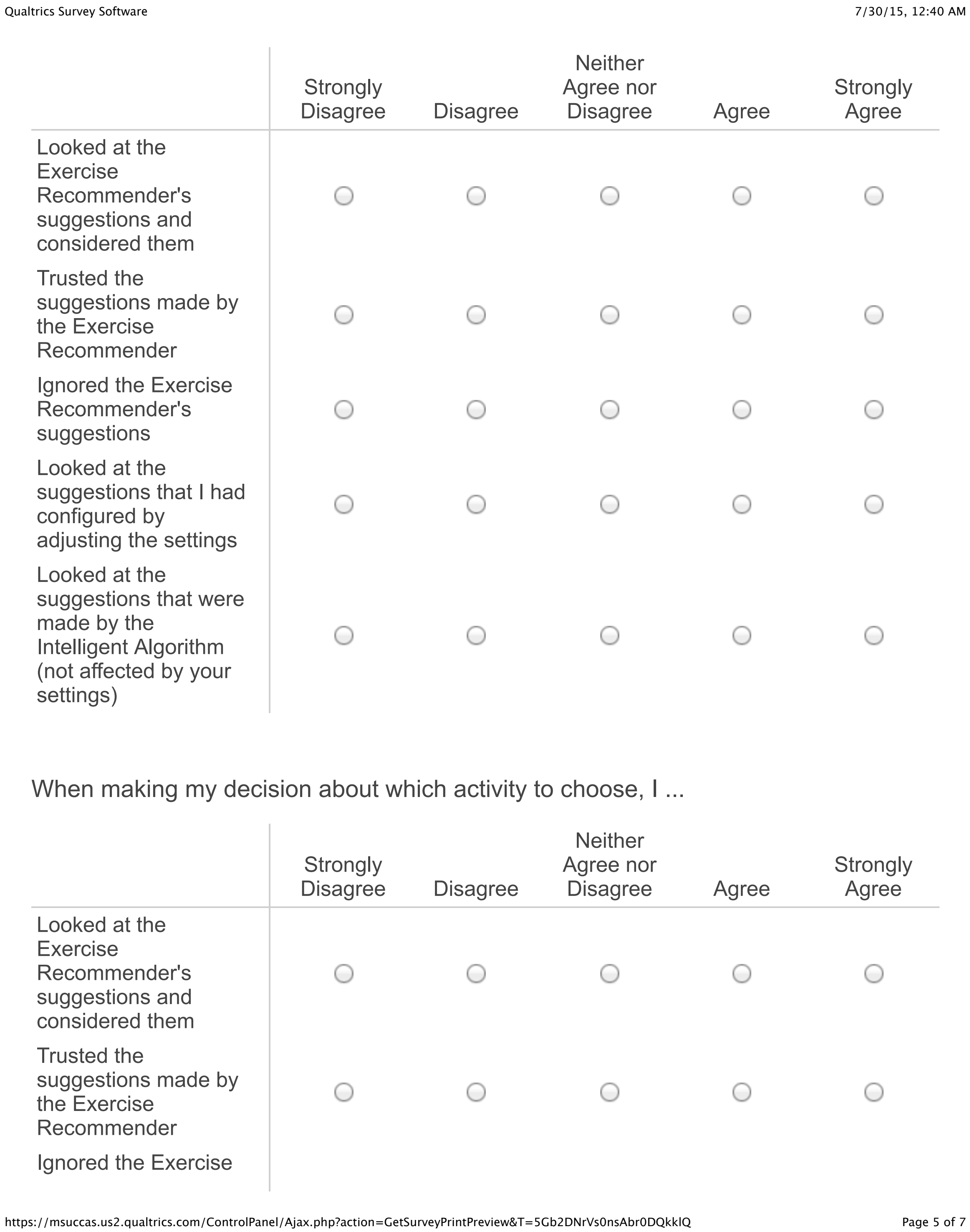
When the Exercise Recommender made suggestions ...

	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
I understood why the suggestions were made	○	○	○	○	○
I thought the suggestions made sense	○	○	○	○	○
the logic behind the recommendations was clear	○	○	○	○	○

When making my decision about which activity to choose, I ...

https://msuccas.us2.qualtrics.com/ControlPanel/Ajax.php?action=GetSurveyPrintPreview&T=5Gb2DNrVs0nsAbr0DQkklQ Page 4 of 7

Figure H.1 (cont'd).



Qualtrics Survey Software 7/30/15, 12:40 AM

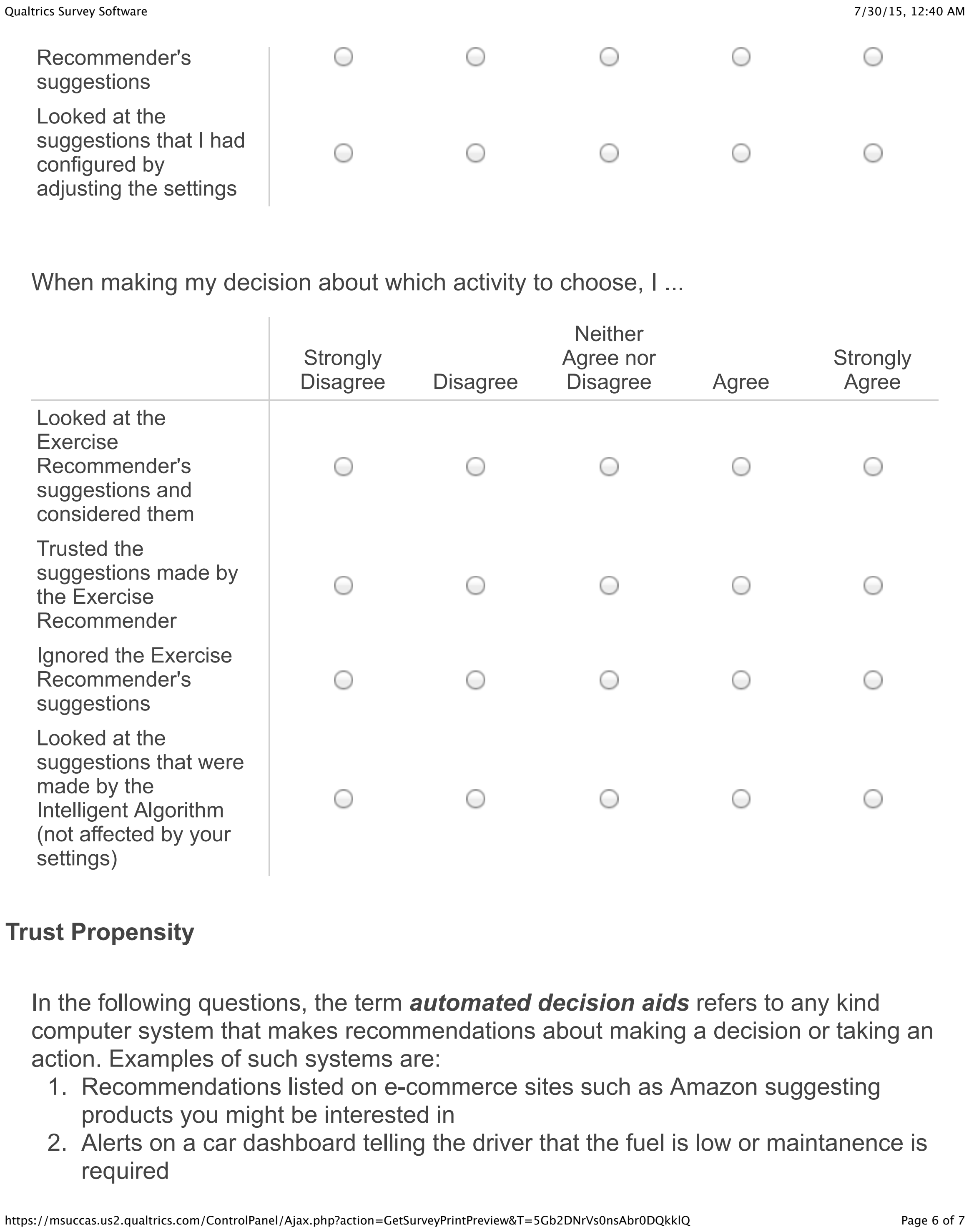
	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
Looked at the Exercise Recommender's suggestions and considered them	○	○	○	○	○
Trusted the suggestions made by the Exercise Recommender	○	○	○	○	○
Ignored the Exercise Recommender's suggestions	○	○	○	○	○
Looked at the suggestions that I had configured by adjusting the settings	○	○	○	○	○
Looked at the suggestions that were made by the Intelligent Algorithm (not affected by your settings)	○	○	○	○	○

When making my decision about which activity to choose, I ...

	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
Looked at the Exercise Recommender's suggestions and considered them	○	○	○	○	○
Trusted the suggestions made by the Exercise Recommender	○	○	○	○	○
Ignored the Exercise					

https://msuccas.us2.qualtrics.com/ControlPanel/Ajax.php?action=GetSurveyPrintPreview&T=5Gb2DNrVs0nsAbr0DQkklQ Page 5 of 7

Figure H.1 (cont'd).



Recommender's suggestions	○	○	○	○	○
Looked at the suggestions that I had configured by adjusting the settings	○	○	○	○	○

When making my decision about which activity to choose, I ...

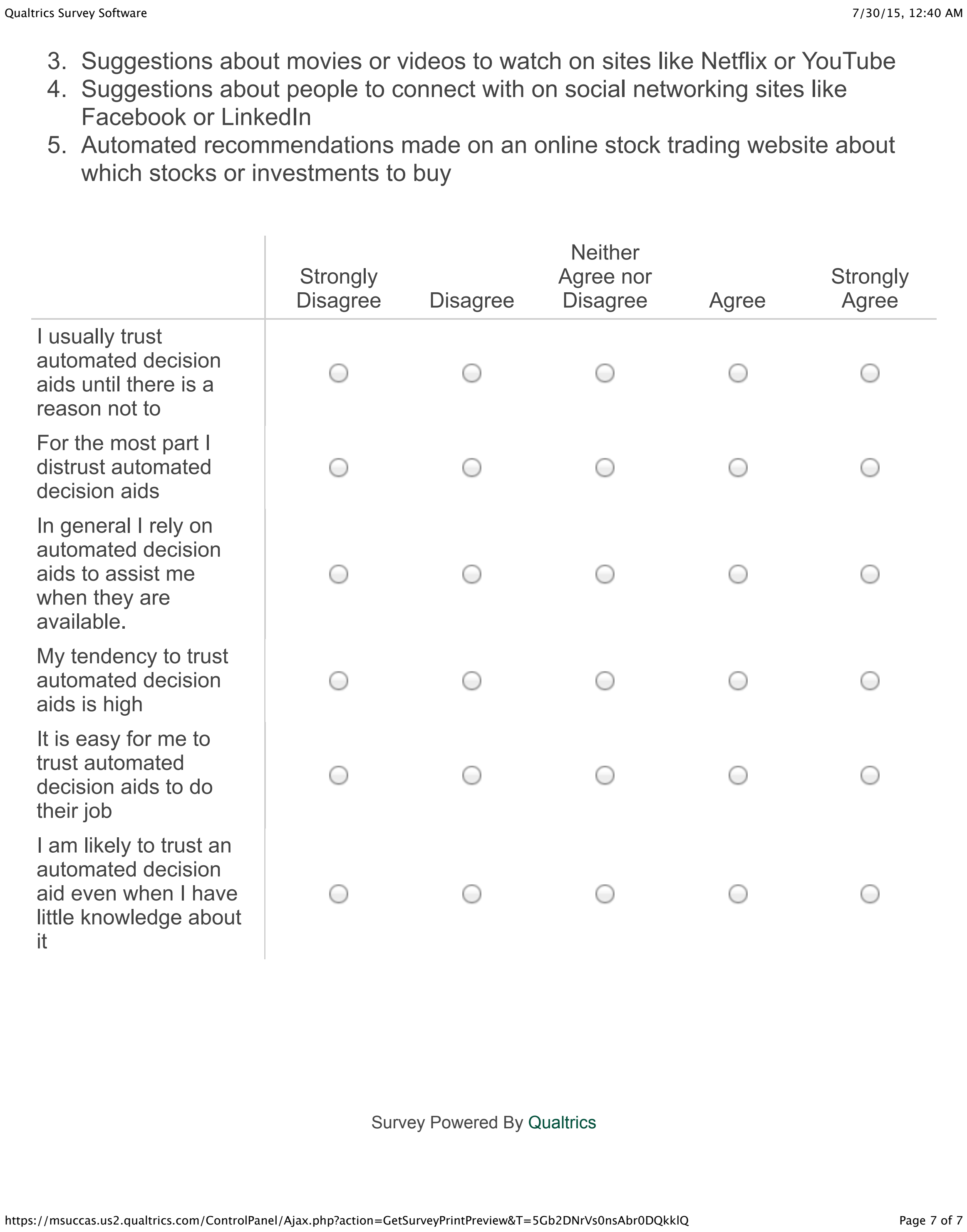
	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
Looked at the Exercise Recommender's suggestions and considered them	○	○	○	○	○
Trusted the suggestions made by the Exercise Recommender	○	○	○	○	○
Ignored the Exercise Recommender's suggestions	○	○	○	○	○
Looked at the suggestions that were made by the Intelligent Algorithm (not affected by your settings)	○	○	○	○	○

Trust Propensity

In the following questions, the term ***automated decision aids*** refers to any kind computer system that makes recommendations about making a decision or taking an action. Examples of such systems are:

1. Recommendations listed on e-commerce sites such as Amazon suggesting products you might be interested in
2. Alerts on a car dashboard telling the driver that the fuel is low or maintanence is required

Figure H.1 (cont'd).



Qualtrics Survey Software 7/30/15, 12:40 AM

3. Suggestions about movies or videos to watch on sites like Netflix or YouTube
4. Suggestions about people to connect with on social networking sites like Facebook or LinkedIn
5. Automated recommendations made on an online stock trading website about which stocks or investments to buy

	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
I usually trust automated decision aids until there is a reason not to	○	○	○	○	○
For the most part I distrust automated decision aids	○	○	○	○	○
In general I rely on automated decision aids to assist me when they are available.	○	○	○	○	○
My tendency to trust automated decision aids is high	○	○	○	○	○
It is easy for me to trust automated decision aids to do their job	○	○	○	○	○
I am likely to trust an automated decision aid even when I have little knowledge about it	○	○	○	○	○

Survey Powered By Qualtrics

https://msuccas.us2.qualtrics.com/ControlPanel/Ajax.php?action=GetSurveyPrintPreview&T=5Gb2DNrVs0nsAbr0DQkklQ Page 7 of 7

REFERENCES

REFERENCES

Adar, E., Tan, D. S., & Teevan, J. (2013). Benevolent deception in human computer interaction. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1863–1872).

Aksoy, L., Bloom, P. N., Lurie, N. H., & Cooil, B. (2006). Should recommendation agents think like people? *Journal of Service Research*, *8*(4), 297–315.

Alberdi, E., Povyakalo, A., Strigini, L., & Ayton, P. (2004). Effects of incorrect computer-aided detection (CAD) output on human decision-making in mammography. *Academic Radiology*, *11*(8), 909–918. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/15354301`

Al-Natour, S., Benbasat, I., & Cenfetelli, R. T. (2008). The effects of process and outcome similarity on users' evaluations of decision aids*. *Decision Sciences*, *39*(2), 175–211.

Amento, B., Terveen, L., Hill, W., Hix, D., & Schulman, R. (2003). Experiments in social data mining: The topicshop system. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *10*(1), 54–85.

Andreoni, J. (1988). Why free ride?: Strategies and learning in public goods experiments. *Journal of Public Economics*, *37*(3), 291 - 304. Retrieved from `http://www.sciencedirect.com/science/article/pii/0047272788900436` doi: 10.1016/0047-2727(88)90043-6

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion.* Princeton university press.

Barnett, G. O., Cimino, J. J., Hupp, J. A., & Hoffer, E. P. (1987). Dxplain: an evolving diagnostic decision-support system. *Jama*, *258*(1), 67–74.

Berner, E. S. (2007). *Clinical decision support systems: theory and practice.* Springer Science & Business Media.

Berner, E. S., Maisiak, R. S., Heudebert, G. R., & Young Jr, K. R. (2003). Clinician performance and prominence of diagnoses displayed by a clinical diagnostic decision support system. In *Amia annual symposium proceedings* (Vol. 2003, p. 76).

Blom, J. (2000). Personalization: a taxonomy. In *Chi'00 extended abstracts on human factors in computing systems* (pp. 313–314).

Bostandjiev, S., O'Donovan, J., & Höllerer, T. (2012). Tasteweights: a visual interactive hybrid recommender system. In *Proceedings of the sixth acm conference on recommender systems* (pp. 35–42).

Bostandjiev, S., O'Donovan, J., & Höllerer, T. (2013). Linkedvis: exploring social and semantic career recommendations. In *Proceedings of the 2013 international conference on intelligent user interfaces* (pp. 107–116).

Bright, T. J., Wong, A., Dhurjati, R., Bristow, E., Bastian, L., Coeytaux, R. R., ... Lobach, D. (2012). Effect of clinical decision-support systems: A systematic review. *Annals of Internal Medicine*. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/22529043` doi: 10.1059/0003-4819-157-1-201207030-00450

Burkolter, D., Weyers, B., Kluge, A., & Luther, W. (2014). Customization of user interfaces to reduce errors and enhance user acceptance. *Applied ergonomics*, *45*(2), 346–353.

Carr, N. (2014). *The glass cage: Automation and us.* New York: W.W. Norton and Company.

Chen, L., & Pu, P. (2009). Interaction design guidelines on critiquing-based recommender systems. *User Modeling and User-Adapted Interaction*, *19*(3), 167–206.

Chen, L., & Pu, P. (2012). Critiquing-based recommenders: survey and emerging trends. *User Modeling and User-Adapted Interaction*, *22*(1-2), 125–150.

Choe, E. K., Lee, N. B., Lee, B., Pratt, W., & Kientz, J. A. (2014). Understanding quantified-selfers' practices in collecting and exploring personal data. In *Proceedings of the 32nd annual acm conference on human factors in computing systems* (pp. 1143–1152).

Clare, A. S., Cummings, M. L., How, J. P., Whitten, A. K., & Toupet, O. (2012). Operator object function guidance for a real-time unmanned vehicle scheduling algorithm. *Journal of Aerospace Computing, Information, and Communication*, *9*(4), 161–173.

Coiera, E., Westbrook, J., & Wyatt, J. (2006). The safety and quality of decision support systems. *Methods of Information in Medicine*, *45*(1), S20–5. doi: 10.1.1.111.9326

Cramer, H., Evers, V., Ramlal, S., Van Someren, M., Rutledge, L., Stash, N., ... Wielinga, B. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, *18*(5), 455–496.

Cremer, H., & Thisse, J.-F. (1991). Location models of horizontal differentiation: a special case of vertical differentiation models. *The Journal of Industrial Economics*, 383–390.

Cummings, M. L. (2004). Automation bias in intelligent time critical decision support systems. In *Aiaa 1st intelligent systems technical conference* (Vol. 2, pp. 557–562).

Cummings, M. L., & Bruni, S. (2009). Collaborative human–automation decision making. In *Springer handbook of automation* (pp. 437–447). Springer.

de Winter, J., & Dodou, D. (2014). Why the fitts list has persisted throughout the history of function allocation. *Cognition, Technology & Work*, 1–11.

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation*, *14*(20), 1–11.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697–718.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *44*(1), 79–94.

Ehrlich, K., Kirk, S. E., Patterson, J., Rasmussen, J. C., Ross, S. I., & Gruen, D. M. (2011, February). Taking advice from intelligent systems: The double edged sword of explanations. In *Proceedings of the 15th international conference on intelligent user interfaces - iui '11* (p. 125). New York, New York, USA: ACM Press. Retrieved from `http://dl.acm.org/citation.cfm?id=1943403.1943424` doi: 10.1145/1943403.1943424

Ekstrand, M. D., Riedl, J. T., & Konstan, J. A. (2011). Collaborative filtering recommender systems. *Foundations and Trendstextregistered in Human-Computer Interaction*, *4*, 175-243.

Fitts, P. M. (1951). Human engineering for an effective air-navigation and traffic-control system.

Gino, F., Sharek, Z., & Moore, D. A. (2011). Keeping the illusion of control under control: Ceilings, floors, and imperfect calibration. *Organizational Behavior and Human Decision Processes*, *114*(2), 104–114.

Glaser, W. T., Westergren, T. B., Stearns, J. P., & Kraft, J. M. (2006, February 21). *Consumer item matching method and system.* Google Patents. (US Patent 7,003,515)

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical*

Informatics Association, *19*(1), 121–127.

Golfarelli, M., Rizzi, S., & Proli, A. (2006). Designing what-if analysis: towards a methodology. In *Proceedings of the 9th acm international workshop on data warehousing and olap* (pp. 51–58).

Guerlain, S., Brown, D. E., & Mastrangelo, C. (2000). Intelligent decision support systems. In *Systems, man, and cybernetics, 2000 ieee international conference on* (Vol. 3, pp. 1934–1938).

Han, S., He, D., Jiang, J., & Yue, Z. (2013). Supporting exploratory people search: a study of factor transparency and user control. In *Proceedings of the 22nd acm international conference on conference on information & knowledge management* (pp. 449–458).

Häubl, G., & Murray, K. B. (2003). Preference construction and persistence in digital marketplaces: The role of electronic recommendation agents. *Journal of Consumer Psychology*, *13*(1-2), 75–91.

Häubl, G., & Trifts, V. (2000). Consumer decision making in online shopping environments: The effects of interactive decision aids. *Marketing science*, *19*(1), 4–21.

Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000). Explaining collaborative filtering recommendations. In *Proceedings of the 2000 acm conference on computer supported cooperative work* (pp. 241–250).

Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, *22*(1), 5–53. doi: 10.1145/963770.963772

Hijikata, Y., Kai, Y., & Nishida, S. (2012). The relation between user intervention and user satisfaction for information recommendation. In *SAC '12 Proceedings of the 27th Annual ACM Symposium on Applied Computing* (pp. 2002–2007). doi: 10.1145/2231936.2232109

Hill, T., Marquez, L., O'Connor, M., & Remus, W. (1994). Artificial neural network models for forecasting and decision making. *International Journal of Forecasting*, *10*(1), 5–15.

Hoffman, R. R., Johnson, M., Bradshaw, J. M., & Underbrink, A. (2013). Trust in automation. *Intelligent Systems, IEEE*, *28*(1), 84–88.

Hostler, R. E., Yoon, V. Y., & Guimaraes, T. (2005). Assessing the impact of internet agent on end users' performance. *Decision Support Systems*, *41*(1), 313–323.

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis.

Psychological methods, *15*(4), 309.

Introne, J., & Iandoli, L. (2014). Improving decision-making performance through argumentation: An argument-based decision support system to compute with evidence. *Decision Support Systems*, *64*, 79–89.

Kahai, S. S., Solieri, S. A., & Felo, A. J. (1998). Active involvement, familiarity, framing, and the illusion of control during decision support system use. *Decision Support Systems*, *23*(2), 133–148.

Kay, J., & Kummerfeld, B. (2012). Creating personalized systems that people can scrutinize and control: Drivers, principles and experience. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *2*(4), 24.

Klein, G. (2008). Naturalistic decision making. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(3), 456–460. doi: 10.1518/001872008X288385

Knijnenburg, B. P., Bostandjiev, S., O'Donovan, J., & Kobsa, A. (2012). Inspectability and control in social recommenders. In *Proceedings of the sixth acm conference on recommender systems* (pp. 43–50).

Knijnenburg, B. P., Willemsen, M. C., & Kobsa, A. (2011). A pragmatic procedure to support the user-centric evaluation of recommender systems. In *Proceedings of the fifth acm conference on recommender systems* (pp. 321–324). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/2043932.2043993` doi: 10.1145/2043932.2043993

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*(8), 30–37.

Kottemann, J. E., Boyer-Wright, K. M., Kincaid, J. F., & Davis, F. D. (2009). Understanding decision-support effectiveness: A computer simulation approach. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, *39*(1), 57–65.

Kottemann, J. E., Davis, F. D., & Remus, W. E. (1994). Computer-assisted decision making: Performance, beliefs, and the illusion of control. *Organizational Behavior and Human Decision Processes*, *57*(1), 26–37.

Kulesza, T., Stumpf, S., Burnett, M., & Kwan, I. (2012). Tell me more?: the effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1–10).

Lam, X. N., Vu, T., Le, T. D., & Duong, A. D. (2008). Addressing cold-start problem in recommendation systems. In *Proceedings of the 2nd international conference on* *ubiquitous information management and communication* (pp. 208–211).

Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology*, *32*(2), 311–328. doi: 10.1037/0022-3514.32.2.311

Lee, G., & Lee, W. J. (2009). Psychological reactance to online recommendation services. *Information and Management*, *46*(8), 448 - 452. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0378720609000962` doi: http://dx.doi.org/10.1016/j.im.2009.07.005

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243–1270.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 50–80.

Leino, J. (2014). *User factors in recommender systems: Case studies in e-commerce, news recommending, and e-learning* (Unpublished doctoral dissertation). University of Tampere.

Li, A. C., Kannry, J. L., Kushniruk, A., Chrimes, D., McGinn, T. G., Edonyabo, D., & Mann, D. M. (2012). Integrating usability testing and think-aloud protocol analysis with "near-live" clinical simulations in evaluating clinical decision support. *International journal of medical informatics*, *81*(11), 761–772.

Lim, B. Y., Dey, A. K., & Avrahami, D. (2009). Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 2119–2128).

Linden, G., Smith, B., & York, J. (2003). Amazon. com recommendations: Item-to-item collaborative filtering. *Internet Computing, IEEE*, *7*(1), 76–80.

Liu, J., Dolan, P., & Pedersen, E. R. (2010). Personalized news recommendation based on click behavior. In *Proceedings of the 15th international conference on intelligent user interfaces* (pp. 31–40).

Madhavan, P., & Phillips, R. R. (2010). Effects of computer self-efficacy and system reliability on user interaction with decision support systems. *Computers in Human Behavior*, *26*(2), 199 - 204. doi: 10.1016/j.chb.2009.10.005

Madhavan, P., & Wiegmann, D. A. (2007, July). Similarities and differences between human–human and human–automation trust: an integrative review. *Theoretical Issues in Ergonomics Science*, *8*(4), 277–301. doi: 10.1080/14639220500337708

Manzey, D., Reichenbach, J., & Onnasch, L. (2008). Performance consequences of automated aids in supervisory control: The impact of function allocation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 52, pp. 297–301). doi: 10.1177/154193120805200421

Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human performance consequences of automated decision aids: The impact of degree of automation and system experience. *Journal of Cognitive Engineering and Decision Making*, 1555343411433844.

Marathe, S., & Sundar, S. S. (2011, May). What drives customization? In *Proceedings of the 2011 annual conference on human factors in computing systems - chi '11* (p. 781). New York, New York, USA: ACM Press. Retrieved from `http://dl.acm.org.proxy2.cl.msu.edu/citation.cfm?id=1978942.1979056` doi: 10.1145/1978942.1979056

Massa, P., & Avesani, P. (2007). Trust-aware recommender systems. In *Proceedings of the 2007 acm conference on recommender systems* (pp. 17–24).

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of management review*, *20*(3), 709–734.

McDonald, D. W., & Ackerman, M. S. (2000). Expertise recommender: a flexible recommendation system and architecture. In *Proceedings of the 2000 acm conference on computer supported cooperative work* (pp. 231–240).

McNee, S. M., Lam, S. K., Konstan, J. A., & Riedl, J. (2003). Interfaces for eliciting new user preferences in recommender systems. In *User modeling 2003* (pp. 178–187). Springer.

Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2012). I trust it, but i don't know why effects of implicit attitudes toward automation on trust in an automated system. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 0018720812465081.

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(2), 194–210.

Messier Jr, W. F., Kachelmeier, S. J., & Jensen, K. L. (2001). An experimental assessment of recent professional developments in nonstatistical audit sampling guidance. *Auditing: A Journal of Practice & Theory*, *20*(1), 81–96.

Miller, B. N., Albert, I., Lam, S. K., Konstan, J. A., & Riedl, J. (2003). Movielens unplugged: experiences with an occasionally connected recommender system. In *Proceedings of the 8th international conference on intelligent user interfaces* (pp. 263–266).

Mosier, K. L., Skitka, L. J., Heers, S., & Burdick, M. (1998). Automation bias: Decision making and performance in high-tech cockpits. *The International Journal of Aviation Psychology*, *8*(1), 47-63. doi: 10.1207/s15327108ijap0801_3

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, *27*(5), 527–539.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica: journal of the Econometric Society*, 69–85.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, *56*, 81—-103. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.87.2456

Nelson, P. (1970). Information and consumer behavior. *The Journal of Political Economy*, 311–329.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220. doi: 10.1037/1089-2680.2.2.175

Norman, D. A. (1990a). *The design of everyday things*. New York: Doubleday.

Norman, D. A. (1990b). The'problem'with automation: inappropriate feedback and interaction, not'over-automation'. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *327*(1241), 585–593.

Norton, M., Mochon, D., & Ariely, D. (2011). The 'IKEA effect': When labor leads to love. *Harvard Business School Marketing Unit Working Paper*(11-091). Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1777100

Ochi, P., Rao, S., Takayama, L., & Nass, C. (2010). Predictors of user perceptions of web recommender systems: How the basis for generating experience and search product recommendations affects user responses. *International Journal of Human-Computer Studies*, *68*(8), 472–482.

O'Donovan, J., & Smyth, B. (2005). Trust in recommender systems. In *Proceedings of the 10th international conference on intelligent user interfaces* (pp. 167–174).

Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2013). Human performance consequences of stages and levels of automation an integrated meta-analysis. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 0018720813501549.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *Systems, Man and Cybernetics, Part A: Systems* *and Humans, IEEE Transactions on*, *30*(3), 286–297.

Pariser, E. (2011). *The filter bubble: What the internet is hiding from you.* Penguin UK.

Park, S.-T., & Chu, W. (2009). Pairwise preference regression for cold-start recommendation. In *Proceedings of the third acm conference on recommender systems* (pp. 21–28).

Parra, D. (2013). User controllability in a hybrid recommender system.

Pereira, R. E. (2001). Influence of query-based decision aids on consumer decision making in electronic commerce. *Information Resources Management Journal*, *14*(1), 31.

Qin, S., Menezes, R., & Silaghi, M. (2010). A recommender system for youtube based on its network of reviewers. In *Social computing (socialcom), 2010 ieee second international conference on* (pp. 323–328).

Rader, E., & Gray, R. (2015). Understanding user beliefs about algorithmic curation in the facebook news feed. In *Proceedings of the 33rd annual acm conference on human factors in computing systems* (pp. 173–182).

Ross, J., Irani, L., Silberman, M., Zaldivar, A., & Tomlinson, B. (2010). Who are the crowdworkers?: shifting demographics in mechanical turk. In *Chi'10 extended abstracts on human factors in computing systems* (pp. 2863–2872).

Roth, M., Ben-David, A., Deutscher, D., Flysher, G., Horn, I., Leichtberg, A., … Merom, R. (2010). Suggesting friends using the implicit social graph. In *Proceedings of the 16th acm sigkdd international conference on knowledge discovery and data mining* (pp. 233–242).

Sanchez, J., Fisk, A. D., & Rogers, W. A. (2004). Reliability and age-related effects on trust and reliance of a decision support aid. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 48, pp. 586–589).

Sargan, J. D. (1958, July). The estimation of economic relationships using instrumental variables. *Econometrica*, *26*(3), 393-415.

Schafer, J. B., Konstan, J. A., & Riedl, J. (2004). View through metalens: usage patterns for a meta-recommendation system. *IEE Proceedings-Software*, *151*(6), 267–279.

Schein, A. I., Popescul, A., Ungar, L. H., & Pennock, D. M. (2002). Methods and metrics for cold-start recommendations. In *Proceedings of the 25th annual international acm sigir conference on research and development in information retrieval* (pp. 253–260).

Schuster, D., Jentsch, F., Fincannon, T., & Ososky, S. (2013). The impact of type and level of automation on situation awareness and performance in human-robot interaction. In *Engineering psychology and cognitive ergonomics. understanding human cognition* (pp. 252–260). Springer.

Sinha, R., & Swearingen, K. (2002, April). The role of transparency in recommender systems. In *Chi '02 extended abstracts on human factors in computing systems - chi '02* (p. 830). New York, New York, USA: ACM Press. Retrieved from `http://dl.acm.org/citation.cfm?id=506443.506619` doi: 10.1145/506443.506619

Skeels, M. M., & Grudin, J. (2009). When social networks cross boundaries: a case study of workplace use of facebook and linkedin. In *Proceedings of the acm 2009 international conference on supporting group work* (pp. 95–104).

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, *51*(5), 991 - 1006. doi: 10.1006/ijhc.1999.0252

Smith, V. L. (1976, May). Experimental Economics: Induced Value Theory. *The American Economic Review*, *66*(2), 274–279.

Solomon, J. (2014). Customization bias in decision support systems. In *Proceedings of the 32nd annual acm conference on human factors in computing systems* (pp. 3065–3074).

Solomon, J., Ma, W., & Wash, R. (2015). Don't wait!: How timing affects coordination of crowdfunding donations. In *Proceedings of the 18th acm conference on computer supported cooperative work & social computing* (pp. 547–556).

Solomon, J., & Wash, R. (2014, October). Human-what interaction? understanding user source orientation. In *Human factors and ergonomics society annual meeting proceedings.*

Stock, J. H., & Yogo, M. (2005). Testing for weak instruments in linear iv regression. *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg.*

Su, X., & Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. *Advances in artificial intelligence*, *2009*, 4.

Sundar, S. S. (2008). Self-as-source. In E. Konijn, S. Utz, M. Tanis, & S. B. Barnes (Eds.), *Mediated interpersonal communication* (pp. 58–74). New York: Routledge.

Sundar, S. S., & Marathe, S. S. (2010). Personalization versus customization: The importance of agency, privacy, and power usage. *Human Communication Research*, *36*(3), 298–322.

Sundar, S. S., & Nass, C. (2000, December). Source Orientation in Human-Computer Interaction. *Communication Research*, *27*(6), 683 –703. Retrieved from http://crx.sagepub.com/content/27/6/683.abstract doi: 10.1177/009365000027006001

Sundar, S. S., Oh, J., Bellur, S., Jia, H., & Kim, H.-S. (2012, May). Interactivity as self-expression. In *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems - CHI '12* (p. 395). New York, New York, USA: ACM Press. doi: 10.1145/2207676.2207731

Tintarev, N., & Masthoff, J. (2007). Effective explanations of recommendations: user-centered design. In *Proceedings of the 2007 acm conference on recommender systems* (pp. 153–156).

Tintarev, N., & Masthoff, J. (2008). The effectiveness of personalized movie explanations: An experiment using commercial meta-data. In *Adaptive hypermedia and adaptive web-based systems* (pp. 204–213).

Tintarev, N., & Masthoff, J. (2011). Designing and evaluating explanations for recommender systems. In *Recommender systems handbook* (pp. 479–510). Springer.

Tintarev, N., & Masthoff, J. (2012, February). Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction*. Retrieved from http://www.springerlink.com/content/9087j6w55r654255/ doi: 10.1007/s11257-011-9117-5

Tjur, T. (2009). Coefficients of determination in logistic regression models—a new proposal: The coefficient of discrimination. *The American Statistician*, *63*(4), 366–372.

Tu, Q., & Dong, L. (2010). An intelligent personalized fashion recommendation system. In *Communications, circuits and systems (icccas), 2010 international conference on* (pp. 479–485).

Verbert, K., Parra, D., Brusilovsky, P., & Duval, E. (2013). Visualizing recommendations to support exploration, transparency and controllability. In *Proceedings of the 2013 international conference on intelligent user interfaces* (pp. 351–362).

Wang, W., & Benbasat, I. (2007). Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. *Journal of Management Information Systems*, *23*(4), 217–246.

Wang, W., & Benbasat, I. (2008). Attributions of trust in decision support technologies: A study of recommendation agents for e-commerce. *Journal of Management Information Systems*, *24*(4), 249–273.

Wash, R. (2013). *Expectations and critical mass in online communities.* (NSF Proposal)

Wash, R., & Solomon, J. (2014). Coordinating donors on crowdfunding websites. In *Proceedings of the 17th acm conference on computer supported cooperative work & social computing* (pp. 38–48).

Woolley, D. J. (2007). The influence of decision commitment and decision guidance on directing decision aid recommendations. *Academy of Information and Management Sciences Journal*, *10*(2), 39.

Xiao, B., & Benbasat, I. (2007). E-commerce product recommendation agents: use, characteristics, and impact. *Mis Quarterly*, *31*(1), 137–209.

Yang, R., & Newman, M. W. (2013). Learning from a learning thermostat: lessons for intelligent systems for the home. In *Proceedings of the 2013 acm international joint conference on pervasive and ubiquitous computing* (pp. 93–102).

Yang, T. Y., & Swartz, T. (2004). A two-stage bayesian model for predicting winners in major league baseball. *Journal of Data Science*, *2*(1), 61–73.