University of California
Santa Barbara

A Quantitative Investigation into the Design Trade-offs in Decision Support Systems

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Computer Science

by

James Austin Schaffer

Committee in charge:

Professor Tobias Höllerer, Chair
Professor Xifeng Yan
Dr. John O'Donovan
Dr. Jonathan Bakdash

March 2017

The Dissertation of James Austin Schaffer is approved.

Professor Xifeng Yan

Dr. John O'Donovan

Dr. Jonathan Bakdash

Professor Tobias Höllerer, Committee Chair

December 2016

A Quantitative Investigation into the Design Trade-offs in Decision Support Systems

by

James Austin Schaffer

Dedicated to my barbell. I may not have operated you perfectly, but whenever I picked you up, you were always 20kg.

Acknowledgements

First and foremost I would like to thank my supervising professor, Tobias Höllerer, for opening up the opportunities to pursue this difficult, interdisciplinary research. Next, I would like to express extreme gratitude to John O'Donovan and his bottomless patience, which allowed him to endure the many debates and arguments that strengthened this work. Third, I would like to express gratitude towards Jon Bakdash, Cleotilde Gonzalez, Xifeng Yan, George LeGrady, Laura Marusich, Emrah Onal, and Michael Yu for their willingness to share their expertise. Finally, I would like to thank my family, especially my wife, for their loving support throughout the pursuit of a PhD.

Curriculum Vitæ

James Austin Schaffer

Education

2016	Ph.D. in Computer Science, University of California, Santa Barbara.
2010	B.E. in Computer Science, University of California, Santa Barbara.

Publications

1. James Schaffer, John O'Donovan, Tobias Höllerer, Human Cognition and Decision Support in the Diner's Dilemma To be submitted to International Journal of Human-Computer Studies.
2. James Schaffer, John O'Donovan, Tobias Höllerer, Recommender Systems: Human Factors and Decision-making. To be submitted to International Journal of Intelligent Systems.
3. James Schaffer, John O'Donovan, Tobias Höllerer, Yinglong Xia, Sabrina Lin, An Analysis of Student Behavior in Two Massive Open Online Courses ASONAM 2016.
4. James Schaffer, John O'Donovan, Laura Marusich, S. Yu Michael, Cleotilde Gonzalez, Tobias Hllerer. A Study of Dynamic Information Display and Decision-making in Abstract Trust Games. Under review, International Journal of Human-Computer Studies.
5. Laura Marusich, Jonathan Bakdash, Emrah Onal, Michael S. Yu, James Schaffer, John O'Donovan, Tobias Höllerer, Norbou Buchler, and Cleotilde Gonzalez. Effects of Information Availability on Command and Control Decision-Making: Performance, Trust, and Situation Awareness.
6. James Schaffer, Tobias Höllerer, John O'Donovan. "Hypothetical Recommendation". FLAIRS 2015.
7. James Schaffer, Prasanna Giridhar, Debra Jones, Tobias Höllerer, Tarek Abdelzaher, John O'Donovan. "Getting the Message? A Study of Explanation Interfaces for Microblog Data Analysis". IUI 2015.
8. James Schaffer, Tarek Abdelzaher, Debra Jones, Tobias Höllerer, Cleotilde Gonzalez, John O'Donovan. Truth, Lies, and Data: Credibility Representation in Data Analysis". IEEE COGSIMA, 2014.
9. Onal, Emrah, John O'Donovan, Laura Marusich, S. Yu Michael, James Schaffer, Cleotilde Gonzalez, and Tobias Höllerer. "Trust and Consequences: A Visual Perspective." Virtual, Augmented and Mixed Reality. Designing and Developing Virtual and Augmented Environments. Springer International Publishing, 2014. 203-214.

10. Emrah Onal, James Schaffer, John O'Donovan, Laura Marusich, Michael S. Yu, Cleotilde Gonzalez, Tobias Höllerer. "Decision-making in Abstract Trust Games: A User Interface Perspective". IEEE COGSIMA, 2014. (Best Paper Award)
11. Domagoj Barievid, James Schaffer, Theodore Kim. "PhysPix: Instantaneous Rigid Body Simulations of Rasters". ACM SIGGRAPH Posters, 2013.
12. James Schaffer, Byungku Kang, Tobias Höllerer, Hengchang Liu, Chenji Pan, Siyu Giyu, John O'Donovan. "Interactive Interfaces for Complex Network Analysis: An Information Credibility Perspective (Invited Paper)." IEEE International Conference on Pervasive Computing and Communications, 2013.
13. James Schaffer, Brian E. Ruttenberg and Ambuj K. Singh. "RAKE: A Visual System for Mining Spatial Images." Center for Bioimage Informatics (Demos), 2010.

Abstract

A Quantitative Investigation into the Design Trade-offs in Decision Support Systems

by

James Austin Schaffer

Users frequently make decisions about which information systems they incorporate into their information analysis and they abandon tools that they perceive as untrustworthy or ineffective. Decision support systems - automated agents that provide complex algorithms - are often effective but simultaneously opaque; meanwhile, simple tools are transparent and predictable but limited in their usefulness. Tool creators have responded by increasing transparency (via explanation) and customizability (via control parameters) of complex algorithms or by improving the effectiveness of simple algorithms (such as adding personalization to keyword search). Unfortunately, requiring user input or attention requires cognitive bandwidth, which could hurt performance in time-sensitive operations. Simultaneously, improving the performance of algorithms typically makes the underlying computations more complex, reducing predictability, increasing potential mistrust, and sometimes resulting in user performance degradation. Ideally, software engineers could create systems that accommodate human cognition, however, not all of the factors that affect decision making in human-agent interaction (HAI) are known.

In this work, we conduct a quantitative investigation into the role of human insight, awareness of system operations, cognitive load, and trust in the context of decision support systems. We conduct several experiments with different task parameters that shed light on the relationship between human cognition and the availability of system explanation/control under varying degrees of algorithm error. Human decision making behavior is quantified in terms of which information tools are used, which information is

incorporated, and domain decision success. The measurement of intermediate cognitive variables allows for the testing of mediation effects, which facilitates the explanation of effects related to system explanation, control, and error. Key findings are 1) a simple, reliable, domain independent profiling test can predict human decision behavior in the HAI context, 2) correct user beliefs about information systems mediate the effects of system explanations to predict adherence to advice, and 3) explanations from and control over complex algorithms increase trust, satisfaction, interaction, and adherence, but they also cause humans to form incorrect beliefs about data.

Contents

Chapter 1

Introduction

As cognitive tasks are increasingly automated, users may lose their opportunity to exercise tacit and procedural knowledge [1]. Additionally, computer algorithms are not yet flexible or intelligent enough to robustly handle unforeseen situations. Thus, in many domains, human decision making will likely remain an important component for the foreseeable future. For example, scientists must scrutinize the provenance of all data used in publications to verify no mistakes were made. Military intelligence analysts make decisions that can impact human lives. In these contexts, users must therefore remain "in the loop" [2]. Designers of decision supports systems (systems that automatically retrieve and summarize information from databases) for these domain applications can help users understand their systems through explanations and allow more control over operations, but designing transparent systems is costly on several fronts: additional usability testing, additional time for design, potentially increased cognitive overhead of the user, and increases in the amount of screen space. If the benefits, costs, and cognitive impacts of these usability features were better understood, they could be employed at the right time and to the right users.

Information retrieval systems have evolved to decrease human mental effort and im-

prove the amount of data that can be incorporated into the human decision making process. When users leverage the algorithms in these systems they are accessing stored procedural knowledge and benefiting from expertise that might not be known at the time of use. In some cases, algorithms have remained simple and useful, such as filtering and keyword matching. However, more complex algorithms have repeatedly demonstrated their usefulness despite pushing the user out of the loop, such as path-finding algorithms for automobile navigation [3] and collaborative filtering for movie recommendations [4]. Ideally, the complexity of these algorithms could be reduced to the level of matching and filtering, but this is not always possible. The conundrum of usefulness vs. simplicity was identified by Norman as early as 1986 and he writes "simple tools have problems because they can require too much skill from the user, intelligent tools can have problems if they fail to give any indication of how they operate and of what they are doing" [5].

In information analysis, humans extract insights from data to build up knowledge [6][7]. As knowledge is increased, the capacity to make better decisions also increases (complete information is a requisite of, but not sufficient for, optimal decision making). Information analysis is not always done during the collection of data, for example, in scientific analysis the process can be more exploratory and hypotheses are not necessarily formed at the outset. This contrasts with "online" decision making, where operators are simultaneously analyzing incoming data and making decisions. In this former case, we do not expect, for instance, that extra time spent accessing and ingesting explanations from intelligent agents would negatively impact outcomes. In the online case, it might be that time spent ingesting explanations of system operations would detract from time and attention spent on more immediate concerns.

Algorithms as provided by intelligent agents are manually invoked or automatically provided through a user interface, typically on a two-dimensional display, by providing the system with a set of input data. The algorithm will then provide an output, which

is a transformation of the original data, typically reduced in size, more informative, and more manageable for humans than the original data. As Norman has pointed out, trouble arises when users can only see the input and output of these algorithms, despite that the calculations may contain hundreds of intermediate steps. In these cases, the user cannot fully understand the limitations of the algorithm, the precise points where it might break down, or how it may potentially bias perception or affect awareness, and therefore decision making. Many systems may also have many tune-able parameters (which we refer to as *control* features). Without a good understanding of how the algorithm operates, users can fail to give feedback to the algorithm, which could improve the effectiveness.

The applicability of this work is limited to situations where a human needs to fill an information requirement to reach a decision. One or more information retrieval tools (i.e. collections of algorithms) are used to fulfill the information requirement, which we refer to as *decision support systems* (DSS). Some definitions of DSS have been extended to include any system that supports a decision [8], however, we limit the use of DSS only to "complex" systems (complexity is defined more concretely later in this chapter). A DSS is also referred to as an "agent" and we use the term "human-agent interaction (HAI) system" to refer to the human/DSS decision making system. We assume the final decision is made by the human based on a change in knowledge that occurred during the evolution of the HAI system [9][10].

We propose profiling complex, automated algorithms using what we refer to as the explanation, control, and error (ECR) profile. We profile human users, and use the human and machine profile to investigate the human cognitive and behavioral reactions to variations in these profiles. Key to the investigation is understanding human beliefs about the data domain and the complex algorithms that are being used. The factors investigated are then used to explain three types of human decision-making behaviors: interaction behaviors, adherence behaviors, and domain decision making behaviors.

This dissertation is organized into nine chapters. This chapter introduces the terminology and concepts used throughout the paper as well as defining the research questions and outlining the thesis contribution. Chapter 2 introduces related work to this thesis. Chapter 3 describes a user study that investigates global situation awareness and decision support in the Prisoner's Dilemma. Chapter 4 describes a user study that investigates the role of explanation and control over recommender systems and how that affects user interaction decisions. Chapter 5 describes a user study that investigates the impact of system explanation on task performance and data beliefs. Next, Chapter 6 describes the measurement framework that was developed in response to the results of the first three user studies. Chapters 7 and 8 apply this framework in two new user studies that experimentally manipulate system explanation, control, and error. Finally, Chapter 9 describes a meta-analysis of results as well as presenting an evaluation of the framework presented in Chapter 6 with future research directions.

1.1 Research Questions

This research attempts to investigate the most general factors that affect human-decision making and system adoption in several HAI task contexts. Specifically, we investigate the relationship between human cognitive variables and system explanation, control, and error. Of additional interest to this research is how the use of a DSS affects changes in human beliefs about data. This leads us to the research questions:

1. Which factors explain variability in decision making (interaction, adherence, success) in the HAI system?

2. How do personal user characteristics and ECR determine decision-making behavior?

3. What is the relationship between correct beliefs about algorithms, their use, and

trust?

4. What is the relationship between user beliefs about algorithms and insight in data analysis?

1.2 Introduction to Terminology and Cognitive Factors

Many of the terms in this research have been given different definitions in other contexts. This section attempts to clearly define the ontological and semantic meaning of each factor. This includes:

- Information Tools and their Complexity
- Explanation, Control, and Error
- Situation Awareness
- Trust, User Experience, System Perception
- Cognitive Load
- Insight (Domain Knowledge)

1.2.1 Information Tools and their Complexity

Here, the term **information tool** is used to denote any function or algorithm that takes data (along with any other parameters) as an input, and produces output in the form of a new set of data (whether autonomously or otherwise). We assume the tool is not created by the user. Both the input and the output data could be arbitrarily

long or short, and of any form. This includes simple algorithms that sort lists as well as the more complex **decision support systems**. For example, information tools could take multiple data sets as input, relate them, and produce a single vector of items as output. Furthermore, when multiple tools are involved, data might be transformed and then passed on for further calculation, or might even serve as control parameters.

Here, complexity refers to the number operations performed by an information tool and its non-linearity. Therefore, algorithms that perform linear math (such as the + operator) are simple, while sorting algorithms that exhibit non-linear behavior or hard-to-predict behavior are more complex. Algorithms whose intermediate steps consist of many sorting algorithms are even more complex. It might be expected that humans can develop accurate beliefs about simple tools faster than complex tools.

1.2.2 Explanation, Control, and Error

In this work, information tools are profiled by their levels of explanation, control, and error (ECR). Explanation level is the amount of output (and thus visual) bandwidth that is allocated to indicating operation. For instance, showing intermediate sorting steps would be an explanation of a sorting algorithm. Control level is the amount of input bandwidth that an information tool provides to the user in addition to the data being sensed. For instance, selection of a kernel in a support vector machine would be considered a control parameter (but not the training data set). Explanation features are sometimes intentionally designed to accommodate control features, such as the selection of an alternate route in a GPS system. Finally, all computational functions and algorithms solve some well defined problem but due to limitations in information or processing power, errors can occur. For instance, recommender algorithms attempt to predict user preferences in sets of items - but complete knowledge of a user's preferences can only be

estimated from the user's item profile, which only partially defines their tastes.

1.2.3 Situation Awareness

The theory of situation awareness (SA) can answer some questions about human decision making in contexts where intelligent agents are present [11][12]. Maximal SA is a requirement for optimal decision making. If an analyst cannot understand what an intelligent agent is doing and an error is made, it could potentially result in loss of life when a critical decision is involved (e.g. drone or aircraft operation). For example, the Air France 447 crash[1] was caused by a combination of system error and lack of transparency. In this crash, measurements from speed sensors became invalid due to ice and due to a lack of system understanding, the pilots could not compensate for the slowing of the plane and disengagement of the autopilot. The pilots pulled up, resulting in the crash. Another example is that powerful tools like R and Matlab can be misused by learners and students when key assumptions made by the system about the domain are not shared by the user. For instance, it has become very easy to "go fishing" for statistically significant results using SPSS or R, where it might be possible to blast data with every statistical test and pick the one that yields "significant" results. Real significance would actually be dependent on a handful of assumptions about independence of variables, order logistics, and so on. Problems caused by these mismatched assumptions are easily solved when the user's understanding of the algorithm is complete, but effecting a user to that state through some facility of the computational environment remains a challenge.

Measurement methodologies are well established [13], those these rely on tailoring to specific domains. SA is defined is the perception of environmental elements with respect to time or space, the comprehension of their meaning, and the projection of their status after some variable has changed. Mica also defines three distinct levels of SA,

[1] http://www.vanityfair.com/news/business/2014/10/air-france-flight-447-crash

corresponding to perception (level 1), comprehension (level 2), and projection (level 3). Higher SA leads to better decisions, both in the data analysis process and also to the application of insights to real world problems after the analysis is completed.

In this work, SA-based agent transparency is used. This is defined as: perception of an information tool, the comprehension of its meaning, and the projection of its status in the future or after some control parameter has changed. This definition is discussed further in chapter 2 and made explicit for our work here in chapter 6.

1.2.4 Trust, User Experience, and System Perception

The word "trust" has been used to describe a number of phenomena in many different domains and therefore it is carefully defined in this section. In this work, the word trust refers to the user's *perception* that he or she can rely on the system. We would therefore expect that users who trust a system would respond strongly to Likert-scale questionnaire items such as "I can rely on the system" or "I trust the system." Trust can affect the way that information tools are used and whether they are used at all. Additionally, trust has been studied extensively in recommender systems [14][15] and human factors [16][17]. Moreover, trust has been shown to be strongly correlated with other types of system perceptions [18][19]. Because automated information tools are not 100% reliable, human over-trusting can cause problems in decision-making situations.

1.2.5 Cognitive Load

Human attention is a limited resource [11], and users of a tool that contains many integrated information tools will have to spend quite some time to ingest the information from all of them, even if the tools are extremely effective in summarizing their target data. Blindly incorporating high levels of transparency and provenance into a tool may

unnecessarily inundate a user with computational details, triggering information overload [20].

Expert users of a particular system may not be as affected by large amounts of information, as they will develop the ability to look in the right place at the right time. Experts in a particular domain can also easily understand domain-oriented explanations and such types of provenance can then result in better decision making [21]. Furthermore, when the goal is learning, novices greatly benefit from explanations, and will actually attempt to access them more often. As users become more familiar with the actual algorithm and corresponding system, cognitive overhead and information overload decreases as users naturally memorize the visual layouts. Adaptive user interfaces [22][23] have gone some way when reducing information overload, though understanding exactly when to show explanations or expose advanced features is still not well understood.

1.2.6 Insight/Domain Knowledge

Insight, which has been called "the atomic unit of knowledge," has been characterized well enough in the past [6] so that it can be measured indirectly by observing other variables. Users form beliefs about data during analysis and correct beliefs correspond to insight. In our work, insight is assessed using testing methodologies similar to visual analytics literature [24][25].

Here, the term "domain knowledge" refers to what a user knows about the real processes in a particular domain that, when sensed, generate data. Thus, users analyze *specific* instances of data from a particular domain, which then builds up into *general* knowledge relating to that domain. Insight is specific to a particular data set (e.g., the distribution of values of a particular dataset column) and domain knowledge is general to all data sets that are generated by sensing that domain (e.g., the distribution of values

of a particular column of a data set is likely to be a Gaussian distribution).

1.3 Contribution

We present a measurement model that is based on previous research and outcomes from three quantitative investigations (Chapters 3,4,5) which are described in this dissertation. These first three investigations uncovered several interesting relationships between the investigated factors and provided the groundwork for the measurement model presented in Chapter 6. The final two investigations in this dissertation evaluate this formalized measurement model. All of the investigations provided some quantitative evidence that was used to answer the four research questions above.

1.3.1 A measurement model for HAI

Previous research in expert systems and recommender systems have shown many benefits of explanation and control for decision support systems. However, measurements of agent-based situation awareness, domain knowledge, and trust have not been conducted simultaneously, so trade-offs between these variables cannot yet be understood. The measurement model presented here is evaluated in terms of its ability to explain adherence and decision quality in the conducted experiments.

In chapter 6, provide recommendations for adapting our general, domain independent factors (e.g. domain knowledge, SA) into task-specific items and terms. Additionally, two applications/specifications of the framework are provided with quantitative evaluation (Chapters 7 and 8). An evaluation of the measured factors is presented in Chapter 9.

1.3.2 New Understanding of Human Decision-making behavior in the presence of a DSS

The primary contribution form this research is new knowledge about how human cognition responds to the presence and configuration of decision support systems. We attempt to identify general system, user, and cognitive factors that predict decision behaviors related to interaction with systems, incorporate of system predictions (adherence), and domain decision success. Better understanding of how humans react to decision support systems informs future system design. In this work, several domains are studied, which allows for evaluation of the generality of effects which are measured. There are three primary, novel findings in this work:

1. the user profiling metrics: trust propensity, cognitive reflection, reported expertise, and insight increase the ability to predict decision making behaviors in the presence of a DSS

2. correct user beliefs (SAT) about DSS mediate the effect of system explanation when predicting adherence to recommendations

3. while explanations and control increase trust, user perception, interaction, and adherence with DSS, they also have the potential to cause human analysts to form incorrect beliefs, which can lead to incorrect decisions or affect future decision-making behavior

A full list of discoveries and a meta-analysis of findings is presented in Chapter 9.

1.4 Limitations and Applications

The results from this work should be applicable to any domain where a user interface is used to control complex tools that summarize data, for example, recommendation algorithms, intelligence analysis, scientific analysis, information visualization, or dynamic systems like aircraft piloting or factory automation control. Robust data from every domain possible, with varying user goals, motivations, and personal investment is needed to create a general model of HAI, but the collection of such data is beyond the scope of this work. Once this data is collected, perfect prediction may still be out of the reach of researchers due to the complexity and non-linearity of the HAI system.

Furthermore, this research predicts how the HAI system will respond to changes in *facilities of information tools*. For all of the experiments, we use similar user interface designs for each treatment and we do not experiment with varying visual presentations or interface control mechanisms. For that reason, this theory cannot assist and guide the user interface design for a particular tool. It could be true that especially fantastic (or terrible) user interface design could have an impact on the parameters we measure in this work. This is an area for future research.

Finally, in this work we do not answer the question "how should explanation and control be designed?" In our work, pilot studies are done to indirectly measure the quality of explanation/control for a given tool based on user situation awareness and feedback, and then the level of explanation/control is varied experimentally for data collection. In practice each system will have to be targeted to its domain and designed on a case-by-case basis, with extensive usability testing and feedback.

Chapter 2

Background

In this chapter, relevant background work from human factors, human-computer interaction, visual analytics, recommender systems, expert systems, and scientific computing are surveyed to create a clear image of how increasing complexity of decision support systems has affected these fields. In the next section we'll give an overview of how complexity has created issues in different applications, and how usability theory has characterized "interacting with automation." Next, we'll discuss previous research in understanding and characterizing interactive data analysis and how decision support systems can assist in this process. Next, cognitive concerns in human decision making are discussed. This chapter concludes with a survey of different approaches to explanation and control in recommender systems, expert systems, and scientific workflow systems.

2.1 Decision Support Systems

Complex computer programs, sometimes called intelligent agents, have grown increasingly complex due to increased processing resources, larger data stores, and more sophisticated computational techniques. Additionally, human-agent interaction (HAI)

has been a field of study since the early work on the MYCIN system (see Shortliffe et al [26]). One kind of intelligent agent, sometimes called a decision support system, are now used ubiquitously, assisting road navigation, recommending movies and music [27], assisting in military operations [28][29], and computing in science (especially biology [30]). While not all users may be capable of programming these tools, conceptual understanding is often within the grasp of users. For instance, a user can understand what a GPS navigation algorithm is presenting at a conceptual level (shortest path in a network) while still being ignorant of hardware operation, physics in wireless communication, and so on. Understanding, then, occurs at the level of human procedural knowledge, rather than at the level of bits and circuits.

When decision support systems are employed, *users can benefit from the procedural knowledge that is latent in the system.* That is, making decision support systems *more usable* means that procedural knowledge known by the model designer becomes more reusable. Attempting to interact with the model becomes an attempt to leverage and understand the procedural knowledge known by the designer - in that sense, using such models could be considered a *collaboration* problem. We assert that even if the barriers to programming are lowered significantly and computational literacy becomes common, provenance features will still be needed since they will increase the re-usability of procedural knowledge by increasing the efficacy at which foreign decision support systems can be understood and integrated into new applications.

2.1.1 Human and Machine Tasks

Hancock and Scallen began addressing problems with human-machine "function allocation," or the theory of how tasks should be broken up between humans and machines [31]. In this article, they presented the famous "Fitts list," or a breakdown of how humans

surpass machines in some capacities and vice-versa. Humans are noted to have greater ability to detect small amounts of visual/acoustic energy, the ability to quickly perceive patterns, to improvise and use flexible procedures, store large amounts of information for long periods and recall this information quickly, and to reason inductively and exercise judgment. Machines are noted to respond quickly to control signals, perform repetitive tasks, store information briefly and erase it completely, apply deductive reasoning, and handle complex operations (do many things at once). Hancock notes, however, that these lists have failed to guide how functions should actually be allocated between humans and machines and that the effort to make machines surpass humans has not been matched by an effort by humans to push their performance to the machine's level. Thus, the allocation of functions really depends more on measures of actual task efficiency rather than satisfying the needs of people at work [32]. The fundamental problem is time: humans are an "open system" in that they are indeterminate in time, while machines are "closed" and only a fixed number of situations are encoded in hardware/software. The period during which this article was written represented a change from a machine-centered design to a human-centered one, where tasks are assigned dynamically and take contextual information into account. Evaluating the design of a computational system is thus based on the effectiveness of the human-machine system tested as a whole.

Controversy seems to exist over whether to create "cognitive agents" that carry out the user's tasks or to create more "tool-like" interfaces where the user is in control and invokes the execution of any computational action [33] (think of *Clippy* from earlier versions of Microsoft Office). Work in expert systems has attempted to automate types of deductive reasoning and provide advice and recommended courses of action to analysts [34], while modern scientific workflows [35] have taken a tool-based approach. This implies that the ideal system design is going to somewhat be dependent on the goals of the system as a whole, and is likely to fall in the middle of the two extremes of cognitive

agent and invoke-able tool. System designers have also taken to designing "adaptive" user interfaces, that record data about the user and adjust tools accordingly, or to allow users to provide certain types of feedback to cognitive agents [36]. A converse approach to user modeling and agent design is to design systems that are consistent, predictable, and controllable - this "tool" approach is described in the next section.

2.1.2 Execution and Evaluation

"Rising levels of automation bring benefits but [they] can also increase dangers" [37]. Shneiderman goes on to write that controlling dangers posed by automation will increase trust, ensuring broader use and safety. The first threats come from errors in the code or design, but systems are often used beyond their original intentions, as users are statistically bound to use the system in ways that were not anticipated, and the unpredictability of real-world conditions can cause malfunction. Effective design, in Shneiderman's opinion, recognizes human responsibility and provides advanced levels of control parameters and user interface design. However, creating systems that match a "tool" approach require much more from the user and increase cognitive overhead. Here, we encounter Norman's "gulf of execution" and "gulf of evaluation" [5].

When a user develops an intent to act while using a system, he or she will run into the "gulf of execution," or, "how can I get the system to do what I want?" The user must map their intent, a collection of psychological variables, to some action sequence in the system (as in programming). The user must determine how to manipulate system inputs to execute the actions which execute the sequence. Once execution has been bridged, the user needs to understand "did the system do what I wanted it to do?" The output from the system is likely to have a complex relationship with the "psychological variables" that caused the user to form their original intent, making evaluation difficult. What's

more, if the change in system state occurs too long after the action was executed, the delay can impede the process of evaluation, as the user may have forgotten the action sequence that corresponded with the output.

Norman [5] gives us two options for bridging the gap between user and system: move the user closer to the system, or move the system closer to the user. Increasing control over and transparency from systems moves the system closer to the user. Norman has already laid out several properties and challenges of this approach. First, the attempt to aid evaluation by presenting extra information can impair the formation of user intentions, but failing to provide such information can make it harder for the user to understand if a job was completed. Second, evaluation can be aided by using the best visual structures available - in some cases graphs and pictures will be ideal, in others, moving pictures or words. And finally, the main challenge for the user is the mapping of system variables to psychological variables - the user must translate their conceived goals into actions suitable for the system. This effort is only diminished when the user finally becomes an expert of the system.

> *The lesson applies to almost any aspect of design. Add extra help for the unskilled user and you run the risk of frustrating the experienced user. Make the display screen larger and some tasks get better, but others get more confused. Display more information, and the time to paint the display goes up, the memory requirement goes up, programs become larger, bulkier, slower. It is well known that different tasks and classes of users have different needs and requirements.*

Design is thus a series of trade-offs. The prototypical trade-off is information vs. time: factors that increase information decreases the amount of visual space available and the ability of the system to respond quickly to new input, due to computational overhead.

2.1.3 Programming and Usability

Finally, we will make a mention here of the considerable effort that has gone into the usability of programming languages, as they have been especially enlightening for the broader field of human-computer interaction. In the past few decades, considerable effort has gone into making programming languages more usable and empowering all end users to be able to create their own programs [38]. Myers examined learning barriers in more detail in [39], which reflect the gulf of execution and gulf of evaluation well. Myers describes six programming barriers, four of which will be discussed here. First *Selection* barriers relate to a system's facilities for finding what programming interfaces are available and which can be used for a particular behavior. Second, *use* barriers are properties of a programming interface that obscure in what ways it can be used, how to use it, and what effects such use will have. Third *understanding* barriers are properties of a program's external behavior that obscure what a program did or did not do at compile/run-time, and *information* barriers are properties of a system that make it difficult to acquire information about a program's internal behavior at runtime.

2.2 Computer-Driven Data Analysis

This section work on studying the interactive data analysis process. Here, theories, definitions, and requirements specifications from visual analytics are presented.

2.2.1 Data, Information, Knowledge, and Wisdom

The abstract idea of data and its representation/analysis has been extensively studied in information theory, database management, scientific computing, perception, and among many other fields which are not strictly concerned with computation, most no-

tably mathematics and statistics. Perhaps the most useful modern definition of data falls into Russell Ackoff's Data/Information/Knowledge/Understanding or Wisdom (DIKW) framework [9] which has been widely adapted in many fields, including information visualization [40] and artificial intelligence [41]. Despite this, the definition of the knowledge and understanding/wisdom portions of the DIKW framework remain nebulous, and definitions may differ based on individual field of study. The more straightforward concepts, data and information, have more agreement in definition - most sources will at least agree that data is not information. More specifically, *data* is often defined as simply a raw number or symbol with no significance attached (e.g. the binary string 01101110 or the tuple [32,5]), and data which has been given meaning by way of a model or human interpretation has become *information*. For instance, the familiar relational model specifies data as a series of rows where each column and row has some significance that gives the data structure and meaning. Definitions beyond these two simple concepts diverge.

Bellinger has defined knowledge as the process whereby information is amassed or accumulated, synonymous with the idea of memorization, and understanding has been defined as the knowledge of rules that can explain the 'why' of data and information (for instance, the function that generates a series of random values). Knowledge is distinct from understanding in the sense that knowledge has no appreciation of 'why' or 'how', and Bellinger asserts that understanding is the essential catalyst where an analyst can up through the hierarchy from data to knowledge. Finally, Bellinger defines wisdom to be on a higher order than knowledge, and that we must move beyond understanding patterns to understanding principles to achieve this level of data comprehension. Bellinger asserts that it is not possible for a machine to obtain wisdom, implying that the human an indispensable part of the data analysis process.

Chen also created a useful dichotomy of the DIKW framework for analysis, specifically visualization [40]. Chen shares the perceptual definitions of data and information,

but is more specific about these definitions in the computational space. Chen defines computational information as data that represents the results of a computation, such as a statistical analysis, that assigns meaning to the data. Knowledge is thus data that represent the result of a computer-simulated cognitive process, such as the rules formulated by a decision tree or the deductive reasoning applied by intelligent systems and case-based reasoning. Chen does not explicitly bring wisdom into computational space, stating that knowledge is sufficient to capture other high levels of understanding as far as computation is concerned. In Chen's framework, data, information, and knowledge can serve as both input AND output to a computational system, and the analysis ends when a sufficient amount of knowledge has been amassed in the user. In other words, the computational tool assists in the process of transferring information or knowledge in the computational space to the user's perceptual space.

2.2.2 Insight and Exploratory Data Analysis

Insight-based evaluations of visualization and data analytics systems have recently appeared in visualization research. While not all authors agree on an exact definition, Saraiya et al writes: 'insight is an individual observation about the data by the participant, or "a unit of discovery" [42]. This definition was leveraged in [7]. Unfortunately, this definition is not really useful in practice. North et al offers a compelling characterization of insight [6]: insight is complex, deep, qualitative, unexpected, and relevant. To elaborate:

- **Complex**: Insight involves all or most of the data and is not concerned with individual values. This means that insights that involve more of the observed data are therefore more meaningful.

- **Deep**: Insight builds up over time, meaning that some insights are logically depen-

dent on others. This means that multiple passes over the data might be necessary to generate a complete understanding.

- **Qualitative**: Insight is not exact, and can be subjective or uncertain. This means that some insights might only be able to be captured by text descriptions or probabilistic models.

- **Unexpected**: Insight is unpredictable, serendipitous, and creative. This implies that analysis systems need to be designed to support exploratory analysis, rather than fixed pipelines. It also implies that automated algorithms that ignore domain semantics to find patterns can significantly contribute to this process, since their data search is not biased by prior theory.

- **Relevant**: Data is deeply rooted in the data domain, meaning that generalized analysis of the raw variables is not enough to generate an insight. This implies that patterns discovered by automated approaches must be related back to the theory of the source domain before they can become useful.

Furthermore, insight has been contextualized in a three stage cyclical framework of hypothesis, exploration, and insight. This is one proposed model of exploratory data analysis (EDA). Supporting and developing for more exploratory systems that trigger insight has become an important end-goal of visualization tools, and the necessity of measuring such a quantity has been emphasized [43]. This has caused some shift in the design of visualization systems from rather straightforward tools that create visuals to exploratory toolboxes with a large suite of built-in analyses and programmability [44]. North's characterization has gained some traction among researchers [45][46]. The theory, while primarily for evaluating the effectiveness of visualization systems, can be reasonably applied to any human-machine analytic system as a metric for success. Thus

questionnaires that attempt to measure insight are dependent on their ability to represent North's five characteristics. Plaisant later used these guidelines for a visualization contest where open-ended search goals had participants submitting subjective evaluations of data [47].

In the visual analytics community, EDA has been defined as the extraction of meaningful insights from large, noisy sets of data [48], [49]. The primary approach is to use hybrid data mining and visualization systems and draw on the flexibility, creativity, and background knowledge of human analysts to improve the knowledge base from which decisions are made. Creating flexible and scalable systems that employ complex models yet remain usable to domain experts is still an open challenge.

The most recent model of exploratory data analysis is found in [7]. This work describes a high-level model for the visual analytics process. Sacha notes that computers miss the creativity of human analysis that allows them to create unexpected connections between data and the problem domain, but they are not able to deal efficiently and effectively with large amounts of information. For this reason, "models" need to be employed, which can be as simple as descriptive statistics or as complex as a data mining algorithm. Sacha also differentiates his definition of insight and knowledge, since weak evidence might lead to an insight that still needs to be validated to become knowledge. Sacha then provides a looping model of interactive data analysis, where users are choosing between various system actions (such as model usage, visualization interaction, etc.) based on their current internal state, which can be "exploration," "verification," or "knowledge generation." They compare their system to other models that have been developed previously (e.g. Green's human cognition model [50]), and note that the analysis of real world problems requires both expertise about the analysis and the domain, and thus domain experts and analytics experts will need to continue to collaborate.

2.2.3 Evaluation of Interactive Interfaces

The visual analytics community has begun to favor open-ended protocols over benchmark tasks for the evaluation of interactive interfaces [25][6][24]. Researchers recommend that participants be allowed to explore the data in any way they choose, creating as many insights as possible, and then measuring their insight with a think-aloud protocol or qualitative measures, such as quantity estimation or distribution characterization. This contrasts starkly with typically well-defined benchmark tasks, which usually have users do things such as find minimum or maximum values, find an item that meets a specific criterion, etc. North [6] cautions that most benchmark tasks may only evaluate an interface or visualization along a very narrow axis of functionality. North seems to be striving towards measuring a latent variable (insight) with a battery of questions that could be understood as indicator variables, but statistical theory such as structural equation modeling [51] has never been applied in their work.

2.2.4 Joining Information Visualization and DSS

At its core, visual analytics aims at employing more intelligent methods in the analysis process [49]. Keim writes that for informed decisions, it is indispensible to include humans in the data analysis process to combine flexibility, creativity, and domain knowledge with the computational power of modern computers. Complex computational capabilities should augment the discovery process, but the ultimate goal is to gain insight into the dataset from the human perspective. Keim goes on to identify several challenges in visual data analytics, one of which is the creation of visual analytics methods for the field of problem solving and decision science. Decision-support systems already exist to reproduce expert knowledge, results from experimentation with these kinds of systems will be discussed in the section on expert systems below. Another problem that was

identified was the issue of user acceptability, or the problem that users are very resistant to changing their working routines, and new automated methods for extracting information from complex data sets need to communicate their goals and abilities more clearly to users.

2.3 Cognitive Concerns

In this section, background work in human cognitive modeling is surveyed, specifically situation awareness, cognitive reflection, cognitive load, user experience, and trust.

2.3.1 Global Situation Awareness

First is Mica's early work on the theory of situation awareness [11] and its measurement [13]. From Mica, SA is defined as the perception of elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future. Since then, Mica's theory of SA has become well accepted [52], and is a useful tool to model human decision-making in complex, dynamic environments. Mica details the three states of SA, which are as follows:

- Level 1 SA (Perception) is simple awareness of multiple situation elements (objects, events, people, systems, environmental factors) and their present states (locations, conditions, modes, actions),
- Level 2 SA (Comprehension) is achieved by integrating Level 1 SA elements through time to understand their past states and how this will impact goals and objectives, and
- Level 3 SA (Projection) is achieved through integrating Level 1 and 2 SA information and extrapolating this information to project future actions and states of the

elements in the environment

SA is a state of knowledge that exists separately from the processes that went into effecting that state - in other words, it exists internally in a user. Furthermore, SA is context specific, pertains only to the state of a dynamic environment, and it exists separately from decision making and performance. It is to be expected that poor performance occurs when SA is incomplete, but it can also occur when the correct action to take in a given state is not known by the user or cannot be calculated in time. Not all system designs are not equal in their ability to convey the information that is most needed or in the way that they are the most compatible with human cognition, so the measurement of SA is useful when performing usability evaluations of interactive interfaces.

Mica recommended using what is known as a situation awareness global assessment test (SAGAT) freeze [13][53] to measure situation awareness, noting that this type of freeze was least likely to be disturbing to an operator's actions when compared with other measurement techniques. During a task, the SAGAT freeze requires a participant to answer several questions related to the current state of the environment, evaluating the participant's responses with the actual state of the environment, which is precisely known in simulation scenarios.

2.3.2 SA-based Agent Transparency

The theory of SA has already been applied to the problem of agent transparency [54]. Chen's theory is called SA-based Agent Transparency (SAT), which is based on Mica's theory of SA and others. Chen refers to Mica's SA as "global" SA, while SAT is relevant only to transparency requirements relevant to understanding the intelligent agent's task parameters, logic, and predicted outcomes. Briefly, SAT is defined as:

- Level 1 SAT (Perception) is the understanding of the agent's purpose, desire, in-

tentions, and performance,

- Level 2 SAT (Comprehension) is the understanding of why the agent behaves the way that it does, an understanding of the reasoning process, and an understanding of the environment and other constraints, and
- Level 3 SAT (Projection) is the understanding of what will happen in the future, including limitations and likelihood of error.

Incorporating all three levels should help a user gain understanding of an agent's reasoning and operation and help the user make informed decisions about "intervention," or what we call here as the manipulation of a "control" parameter. Chen notes that automation reliability strongly influences a user's attitude toward automation which can have significant impacts on trust, and thus has an impact on the degree to which that automation is leveraged. Overtrusting automation leads to improper automation use, and under-trusting results in disuse of the automation, which can impact performance. Chen notes that information visualization and the display of uncertainty are key factors in understanding automation, and discussed this in more detail in [55].

2.3.3 Information and Cognitive Load

Effective user interface design can overcome limitations in the user's attention and working memory [56][11]. Additionally, by increasing provenance of computational processes, user interface design can facilitate the correct perception of trust and data provenance [57]. Developers and interface designers must deal with the challenge of combining and representing large amounts of data at the right time in the right format. Unfortunately, there are finite limits on a human's ability to efficiently and effectively summarize large amounts of data, especially when tasks are time-sensitive.

Attention is a major limit on situation awareness [11]. Direct attention is required for perception and processing of cues, but also for the later stages of decision making. People typically employ a process of rapid information sampling from several cues, following a pattern dictated by their long term memory which concerns the relative priorities of information, and is proportional to the frequency at which information changes. Since the supply of attention is limited, more attention to some elements may increase the SA on those elements, but may decrease SA on other elements when attention limits are reached.

The term "cognitive load" originates from education and learner theory [58] and problem solving [59] and is loosely defined as a "multidimensional construct representing the load that performing a particular task imposes on the learners cognitive system." Greater cognitive effort by users of systems leads to increased error when performing tasks. Paas [60] surveys numerous methods of measuring cognitive load during participant tasks, noting that cognitive load can be assessed by measuring mental load (portion of cognitive load that originates from task to subject relationship characteristics), mental effort (the actual effort exerted as demanded by task requirements), and performance. Participant self-reported rating scale techniques have been apparently successful, as participants seem capable of accurately reporting their mental burden. Physiological techniques, such as the measurement of heart rate, brain activity, and pupil dilation have also been successful. Finally, other kinds of performance measures can be applied, such as measuring the participant's effectiveness at managing a secondary task periodically while performing the primary task.

The term information overload [20] is used to convey the notion of "receiving too much information," and has also been called "cognitive overload," "sensory overload," and "information fatigue syndrome." In information overload research, the primary metric is how performance changes in response to the amount of information that is exposed, and

most researchers have reported that more information leads to better performance up to a certain point, where it collapses. Ways to counteract information overload include increasing the quality of information and improving the organization of information.

2.3.4 Trust, Trusting Propensity, and System Perceptions

Trust propensity and its relationship to trust has been studied extensively in psychology, notably Colquitt et al [61] and Gill et al [62]. Behavioral outcomes are affected by trust propensity when partially mediated by trust and *trustworthiness*, which is information about a trustee. The effects of trust propensity on behavioral outcomes disappears when information about the trustee becomes more reliable. Other studies in e-commerce have also found similar mediating effects between trust and trust propensity [63]. This work hypothesizes that both trust and trust propensity need to be measured to understand adherence under different system explanation levels.

Trust and system perceptions are highly correlated, with many factors (enjoyment, perceived control, perceived usefulness, perceived transparency, perceived privacy, perceived security) being mediated or being highly correlated with trust and trust propensity [64][65][18][66]. Perceived system usefulness tends to correlate the most strongly with both overall system satisfaction and trust [19]. This work hypothesizes that system perceptions and trust play a strong role in predicting interaction and adherence.

2.3.5 Cognitive Reflection

Work on attention and cognitive reflection by Daniel Kahneman [67] has been successful in discriminating between "fast" and "slow" thinking using a variety of questions that effectively trick the human processing system. Since then, "cognitive reflection" tests have been frequently used due to its correlation with human intelligence and de-

cision making [68][69][70]. This work hypothesizes that cognitive reflection would be a strong predictor of user decision behavior when interacting with a DSS.

2.3.6 Reported Expertise

Collecting information from users related to *self-reported* domain expertise is fast and straightforward, however, there are concerns about the reliability of this type of metric [71] to indicate expertise. What these metrics can indicate, however, is the number of "unknown unknowns" relative to the user. Deficits in knowledge are a double burden for users, not only causing them to make mistakes but also preventing them from realizing they are making mistakes [72][73]. When people compare themselves to others, they typically inflate their own skills while ignoring the skills of others, additionally, this problem is exacerbated by cognitive load [74], which increases anchoring on a person's initial assessment of their own skills.

The work on the Dunning-Kruger effect thus leads us to hypothesize that self-reported experts are less likely to interact with or adhere to decision support systems, due to their inability to correctly assess the relative accuracy of the system (e.g., "GPS systems are not *that* accurate - I can find a better route myself!")

2.4 Explanation, Control, and Error

Explanation and control from automated algorithms has been studied as early as 1975 [75]. This section presents work on explanation and control features in a three research areas, recommender systems, expert systems and scientific computing. We will also survey research where the accuracy of decision support systems was experimentally manipulated.

2.4.1 Recommender Systems

Over the last 15 years, research has shown that explanation of a recommender system's reasoning can have a positive impact on trust and acceptance of recommendations. Recent keynote talks [76] and workshops [77] have helped to highlight the importance of usability. many recommender systems function as *black boxes*, providing no transparency into the working of the recommendation process, nor offering any additional information beyond the recommendations themselves [78]. This may negatively affect user perceptions of recommendation systems and the trust that users place in predictions. To address this issue, static or interactive/conversational explanations can be given to improve the transparency and control of recommender systems [79]. Bilgic et al. [80] furthered this work and explored explanation from the promotion vs. satisfaction perspective, finding that explanations can actually improve the user's impression of recommendation quality. Later work by Tintarev and Masthoff [81] surveyed literature on recommender explanations and noted several pitfalls to the explanation process, notably including the problem of confounding variables. This remains a difficult challenge for most interactive recommender systems [82], where factors such as user ability, mood and other propensities, experience with the interface, specific interaction pattern and generated recommendations can all impact on the user experience with the system. [83] note that users liked and felt more confident about recommendations they perceived as transparent. The importance of system transparency and explanation of recommendation algorithms has also been shown to increase the effectiveness of user adoption of recommendations by Knijnenburg in [18].

Of important note is the user experience framework created by Pu [19], which uses a number of subjective system aspects such as perceived transparency, perceived accuracy, perceived control, and overall satisfaction. They showed that SSA can be used to explain

how explanation and control relate to participant-reported use intention. Knijnenburg et al [18] used a similar evaluation framework to argue for the importance of inspectability (explanation) and control. In contrast to Pu's study, choice satisfaction was recorded, which gave an extra dimension to the analysis. Knijnenburg also used several user-modelling constructs, such as familiarity with recommenders, music expertise, trusting propensity, and effort to use the system.

2.4.2 Expert Systems

Work in knowledge-based or "expert systems" has illuminated the effects of exposing explanations from complex agents. Gregor et al. [34] provide an excellent summary of the theory of crafting explanations for intelligent systems. User studies which test the effects of explanation typically vary explanation level and quantify concepts such as like adherence or knowledge transfer. Key findings show that explanations will be more useful when the user has a goal of learning or when the user lacks knowledge to contribute to problem solving. Explanations also have been shown to improve learning overall and improve decision making. The impact of explanation on both novices and experts has also been extensively studied [21]: novices are much more likely to adhere to the recommender/expert system due to a lack of domain knowledge, and expert users require a strong 'domain-oriented' argument before adhering to advice. Experts are also much more likely to request an explanation if an anomaly or contradiction is perceived. Most of these studies focus on decision making domains (financial analysis, auditing problems) and were conducted before the explosion of data which now characterize typical web databases. When browsing or analyzing data that is too large to be analyzed by hand, decision makers have no choice but to utilize automated filtering techniques as part of their search strategy - this creates new questions about what might change in the

dynamics between humans and automated algorithms.

2.4.3 Scientific Computing

What's known as "data provenance" has been studied extensively in scientific computing [84]. The drive for rich metadata attached to the output of computational analyses originates from the need to make analyses reliable and reproducible [85]. Rapidly changing needs and a requirement for flexibility has driven scientists in many fields toward "scientific workflow systems" such as Taverna [86] and Kepler [35]. These systems serve as a high level substitute for scripting languages, where scientists invoke and pass their data through numerous complex computational processes, essentially "stitching together" many different entities. McPhillips [87] has identified numerous usability problems in this domain and formed multiple desiderata for such systems, including clarity, predictability, and reportability. Clearly, individual decision support systems in such frameworks will have to put significant effort into provenance design to effectively satisfy these desiderata.

Chapter 3

Improving Situation Awareness through Information Support in the Diner's Dilemma

In this chapter, we'll describe our work on how varying levels of support from a user interface affect situation awareness and decision making in Diner's Dilemma. Here, the "levels of support from the user interface" directly refers to the amount of decision support that users were given, with the low level model simply showing the most recent state of the data, the second model showing all states of the data (with a control for which states are shown at a given time), and the most supportive model allowing the user to calculate what the best decision would be based on two parameters that the participant had to infer from the second model, or come up with on his own. Results support the following claims:

- Decision support interfaces, when designed and introduced to the user properly, increase global situation awareness

- Increased global situation awareness positively impacts decision making (in this case, performance a binary choice task under various conditions)

3.1 Introduction to Trust Games

Today, many decisions are made online through user interfaces (UI) with multiple collaborators and complex computer support systems. In such settings, mutual cooperation is important for effective and efficient completion of work and analyses, however, self-interested actions frequently threaten to undermine cooperation [88]. When users make decisions through a user interface, unique challenges and opportunities arise for the designers of those interfaces. For instance, showing the right information at the right time to a decision maker may improve the quality of decision making, but hiding information from a self-interested actor at a critical moment may improve cooperation, thus benefiting the group as a whole.

Past studies have often leveraged abstract trust games, such as the Prisoner's Dilemma, to study human cooperative behavior [89, 90, 91, 92]. An iterated version of the Prisoner's Dilemma, called the Diner's Dilemma (DD), is applied here to study the relative magnitude of co-actor behavior and user interface design on a decision-maker. Here, we will use term "diner" when referring to an individual decision maker and "co-diner" when referring to this player's co-actors. In the Diner's Dilemma, several diners eat out at a restaurant over an unspecified number of days with the agreement to split the bill equally each time. Each diner has the choice to order the inexpensive dish (cooperation) or the expensive dish (defection). Diners receive a better dining experience (here, quantified as *dining points*) when everyone chooses the inexpensive dish compared to when everyone chooses the expensive dish. However, individual diners are better off choosing the expensive dish regardless of what the others choose to do. Nonetheless, when the same

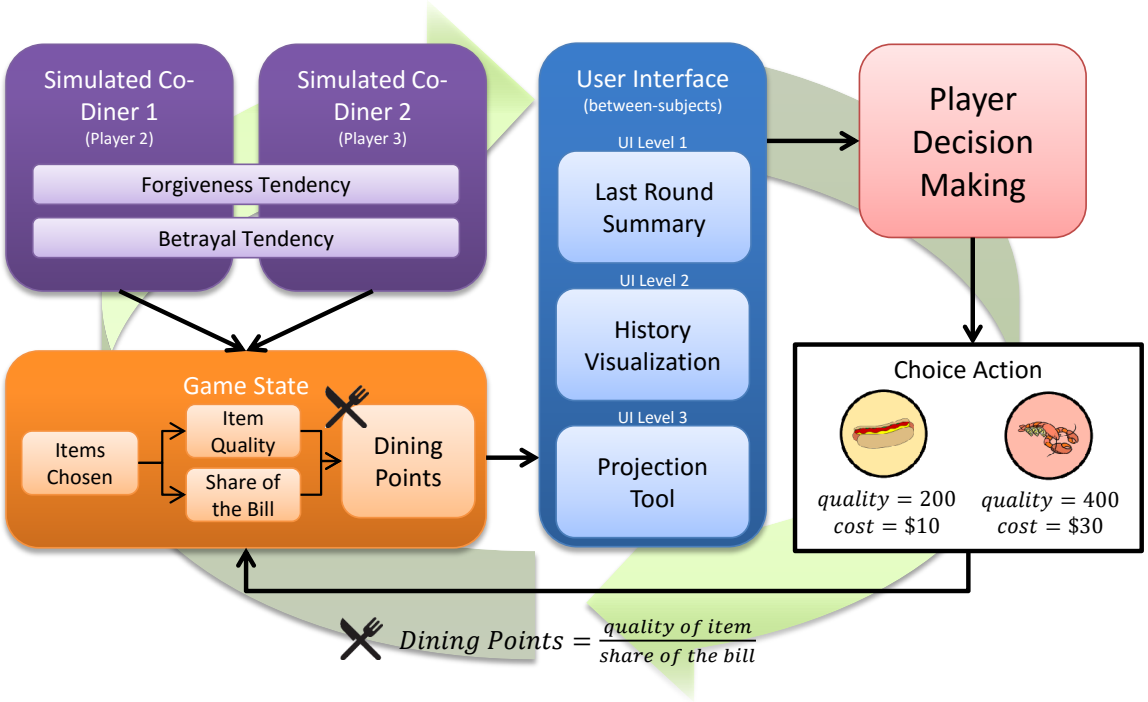


Figure 3.1: Overview of the Diner's Dilemma game and experiment. Player participants interacted with a user interface (Figure 4.1) to play the Diner's Dilemma game. One of three different user interfaces and one of eleven different co-diner strategies were chosen randomly for each participant. Forgiveness and betrayal refer to the degree to which the co-diners would forgive defection and punish cooperation, respectively. Players tried to maximize their Dining Points, which are determined each round by the quality of the item selected divided by the player's share of the bill. The "History Panel" which showed a summary of past rounds or the "Projection Panel" were available depending on the treatment assigned.

group of diners meets repeatedly under the same bill-sharing agreement, cooperation may develop, leading to a better overall group dining experience.

In [93] and [94], it was suggested that the amount of information available to an agent has an effect on their decision or next course of action by affecting trust and cooperation. [93] showed a positive correlation between trust and situation awareness, suggesting that user interfaces might be an effective way to encourage trust in social settings. Specifically, these previous experiments suggest that showing increased amounts

of information through a user interface about past, present, and future decisions made by co-actors in social settings tends to increase cooperation (and therefore, trust). Not yet understood are the full range of parameters that define the interplay between the design of the user interface, the propensity of co-diners to exploit, individual user cooperation, situation awareness (SA), and the resulting performance of the individual user. In this work, participants played the Diner's Dilemma game with different variations of simulated co-diners through one of three variations of a user interface (Figure 3.1). We hypothesized that the user interface may have a different effect on cooperation under certain co-diner forgiveness and betrayal rates. This leads to the following research questions:

1. How does co-diner cooperation and defection affect human cooperative behavior under different UI support conditions?

2. Can a user interface be used to encourage or discourage human cooperative behavior?

3. To what extent can user interfaces improve situation awareness?

4. How do (2) and (3) affect overall performance of the individual?

3.2 Related Work

In addition to summarizing experiments directly related to this work, this section surveys theory in abstract trust games, information displays, and decision making.

3.2.1 Information Displays and Human Decision Making in Trust Games

[95] and [96] conducted experiments that demonstrate how information displays can affect decision-making in trust games. A key finding of these works was that an increase in information in the user-interface led to an increase in cooperation behavior, joint-performance, and satisfaction. Pairs of participants in the experiment were given different levels of interdependence information across four different levels of information exposure. The increase in cooperation seen in these experiments might be explained by a participant's feeling of obligation to reciprocate when historical data was laid out before them via the UI.

3.2.2 Trust and The Cooperation Problem

Mutual cooperation can create a situation where the whole is greater than the sum of the parts. Unfortunately, the short-term payoffs from defection are always tempting, making collaboration in social situations inherently difficult [97]. In economics, the Prisoner's Dilemma is an abstract game used to study cooperation. In the Prisoner's Dilemma, two players decide to take an action (cooperate or defect) without communication beforehand, where defection leads to a higher outcome for an individual regardless of the other players actions, but with mutual cooperation leading to outcome higher outcome than mutual defection. Iterated forms of this game (Diner's Dilemma or the Public Goods Game) are useful for studying dynamically changing situations, such as military arms races. The proportion of mutual cooperation in an iterated game of Prisoner's Dilemma can be considered a measure of trust between the two participants, as choosing to cooperate makes an individual vulnerable to the worst possible outcome.

The notion of trust has diverse meanings across a number of fields such as sociology,

psychology, economics, and computer science [98]. Here, we take a general definition of trust to be "a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions and behaviors of another" [99]. Trust dictates how people interact with each other, their tools, and their environment; thus "trustability" impacts the effectiveness of interactive interfaces. In this experiment, participants in the Diner's Dilemma game develop a trust-based relationship with the simulated co-diners and the information display that was provided. Moreover, the UI is positioned as a mediator of interpersonal trust by exposing and cataloging the actions of each individual in real-time, which has been shown to increase cooperation and interdependence. In the Diner's Dilemma, trust develops as the co-diners develop the knowledge to predict each other's actions over time.

Psychological motives to cooperate in the Prisoner's Dilemma can be complex. Early empirical evidence suggests that human behavior was not dictated only by selfish considerations but also by other regarding preferences (see for example, [100]; [101]). Research models suggest a direct benefit from the well-being of others (often described as altruism, for example, [102]), a preference for equality in outcomes [103], and a preference for "kind" behaviors that allow the other player access to better outcomes [104]. In addition to the reputation benefits of cooperating in an iterated Prisoner's Dilemma, these models help to promote cooperation in the face of more selfish motives. The latter two preferences for equity and kindness also allow for mutual defection and call for it when/if the other players defect. If we translate these theories to our current design, all models call for cooperation with a baseline Tit-for-Tat (TFT) strategy, and there are at least two potential arguments for defection in TFT strategies with random defection. Moreover, evidence suggests that people are willing to pay a cost to retaliate to defection [105].

In spite of the findings of pro-social preferences, selfish motives continue to be observed in these experimental designs, though less strongly than would be predicted by traditional

economic rationality. These motives do not seem to be inherent individual differences but could potentially be acquired. Findings indicate that those exposed to a traditional economic education are generally less inclined to cooperate in a public goods game than others [106]. People will also often engage in selfish motives when they can "get away with it." In experiments with the dictator game [107], people were willing to accept a smaller payout that couldn't be shared than to take a larger payout which could have been shared, i.e., people are willing to pay to be (secretly) selfish. As random forgiveness increase in the TFT strategies in our design, the consequences of being selfish are reduced.

In social dilemmas, participants are not always aware about how their actions influence other people and vice-versa. This can quickly create a situation where less-than-optimal results are achieved for all participants, as higher levels of information about the game and the strategies taken by other players have been shown to greatly improve the outcome for all [100]. More recently, a study by [97] have shown that consistent contributors, actors that consistently contribute to the public good regardless of the actions that their co-actors take, can have a significant positive impact on the behavior of the group as a whole. Consistent contributors occurred naturally in four previously-collected data sets, and were shown to improve overall cooperation. Clearly, awareness of interdependence encourages pro-social behavior and trust in these interactions.

In this work, we examine the boundaries where pro-social behavior breaks down and also the impact of consistent contributors with a varying level of UI support. This paints a comprehensive picture of the interplay between the user interface and interpersonal cooperation.

3.2.3 Related Experiments

The work by Teng et al [93] was, to our knowledge, the first experiment that examined the relationship between situation awareness and user interfaces for a version of the iterated Prisoner's Dilemma. This work focused on the relationship between behavior and situation awareness in Diner's Dilemma, wherein a human played repeated rounds of the game with two computer opponents. Based on SA theory and design principles, the authors developed three different UIs that were expected to represent the information needed to support a specific SA Level. Several Trust-related metrics were also assessed, including percentage of cooperation over time and subjective level of self-reported trust toward the opponents. They found that participants in the most simple UI treatment cooperated more frequently when simulated co-diners encouraged cooperation, but participant defection increased when the user interface displayed more information. It was also concluded that cooperation level is a good indicator of the trust that participants place in their co-diners.

The work by Onal et al [94] significantly built on Teng's study, expanding the sample size, varying the opponent strategies, and revamping all UI levels to induce the desired SA and using a SAGAT style questionnaire to assess participant understanding of the game and the interface. To study the effects of UI components on awareness and decision-making behavior, an online study of 95 users was conducted using Amazon's Mechanical Turk (AMT). Participants played repeated trials of the DD game, and answered evaluative questionnaires at multiple stages in the game. The experiment highlighted two key results: First, there is a strong correlation between SA and performance in the game, and second, UI composition and information presentation have an impact on human trust and cooperation behavior.

3.3 The Diner's Dilemma Web Game

In this experiment, participants interacted with a web-based implementation of the Diner's Dilemma game and were recruited online through AMT. During the game, the user's goal was to maximize his or her "dining points," defined as the ratio of the food quality of the chosen meal divided by the diner's share of the bill. In each round, the participant must weigh the pros and cons of selecting either hot dog or lobster by assessing the cost/value trade-offs involved, the opponent behavior, and the long-term gain of a chosen strategy.

The simulated co-diners played variants of Tit-for-Tat (TFT), a simple strategy in which the opponent makes the same choice that the participant did on the previous round. Opponent strategies varied from pure TFT along two parameters: forgiveness and betrayal. The higher the forgiveness parameter, the more likely the simulated co-diner will respond to a lobster order with a hotdog order. The higher the betrayal parameter, the more likely the opponent will respond to a hotdog order with a lobster order. To make the game more understandable for the human participant and to simplify result analysis, simulated co-diners reacted only to the human decision and not to each other.

When beginning the game, participants completed a pre-study questionnaire that collected some basic demographic and expertise information, and were required to answer three screening questions to test their attention. They were then directed to an interactive training session that explained the game rules in detail. After the training, if the participants were ready to continue, participants played a 100-round game of Diner's Dilemma. When all rounds were completed, the users were directed to a post-study questionnaire where they provided feedback on the game and the simulated opponents.

3.4 Experiment Setup

In related work, results suggested that exposing the participant to more information increased situation awareness and increased situation awareness led to more cooperation in the situations where cooperating was an advantageous strategy. This experiment is also interested in how the opponent defection affects participant behavior in the presence of different support interfaces, however, the full range of co-diner behavior is tested under all user interface conditions.

3.4.1 Prestudy and Poststudy

A pre-study and post-study were given to participants before and after playing the game, respectively (refer to Figure 3.3). Important to our analysis was the measurement of trust propensity (via participant response to the question "I am a trusting person") and altruism. Trust in co-diners was taken post-study with the question "How much do you trust this pair of co-diners?" Similar to the dictator game [108] and the trust/investment game [109], altruism was measured with the following scenario:

> *"You have $50. You can keep this money and do with it whatever you wish or you can send some or all of it to another person in another room (whom you will never see or meet). They are also given $50 and the same instructions. Any money sent will be tripled on the way to the other person. Thus, if you send them $10, they will receive $30; if they send you $30, you will receive $90, and so on. You can send them any amount that you wish. You can send them nothing if you wish. This decision is completely up to you."*
>
> *"How much of your $50 would you send?"*

3.4.2 Study Design

An 11x3 between-subjects experiment was designed to investigate above hypotheses. Two simulated co-diners played Tit-for-Tat with eleven variations of forgiveness and betrayal parameters. Three different UIs were designed that exposed varying degrees and complexity of information. The UIs and a training module were iteratively improved through pilot testing before experimentation. The pilots revealed that the game was most easily explained to new players through the concept of reciprocity, which was then used to explain game rules to the participants in the primary experiment.

A key challenge in the experimental design was the measurement of SA. Although there are several approaches for the direct measurement of SA, the SAGAT is a widely tested and validated technique [53] for objectively measuring SA across all of its elements (levels 1, 2, and 3) with numerous studies supporting its validity and reliability [110, 111]. The SAGAT questions were based on an analysis of SA requirements of the game. The SAGAT questionnaire required participants to have a knowledge of basic game parameters, observed co-diner behavior, and optimal game strategy in the future. There were a total of 8 questions in each of two questionnaires (one after 50 rounds and after 90 rounds), all with multiple-choice answers. Participants were informed about the questionnaires, but not about their timing. The game's UI was not visible during the questionnaire phase. The range of possible scores was 0-8 on each questionnaire.

For each participant, detailed round data was taken, allowing each session to be reconstructed in its entirety. Our main research questions were related to overall dining points, cooperation percentage, and score on the two SAGAT questionnaires. Independent, dependent, and game variables are listed in Tables 3.1,3.2, and 3.3. More details on the game and the reasoning behind the selection of values for the independent variables are given in the following two sections.

Term Name	Description
UI Level	1, 2, or 3 selected randomly when the participant starts the study. The Level 1 UI only provides the central panel and the 'Last Round' panel, Level 2 provides the Level 1 components plus the "History panel", and the Level 3 UI contains the Level 2 components plus the "Estimate my Score" panel. See Figure 4.1 for a visual of each component. We also measured the quantity of interaction with the UI Level 3 component (integer). Any slider movement counted as an interaction.
Simulated Co-Diner Strategy	A variant of Tit-for-Tat, with one noise parameter for the co-diner's reaction when the player cooperates, and one noise parameter for the reaction when the player defects.

Table 3.1: Independent variables in the study

Term Name	Description
Cooperation	Proportion of rounds that participant chose Hotdog.
Performance	To measure performance, closeness to optimal strategy was used, which is the difference between observed cooperation percentage and optimal cooperation percentage (dependent on simulated co-diner strategies).
SA	Situation awareness: performance on the questionnaires given during the game which test both game understanding, and the past/future state of the game. Two identical SAGAT tests were given - SAGAT 1 and SAGAT 2, at fixed times during the middle of the game (timing was not told to participants).
Trust Propensity	Participant response to the statement "I am a trusting person."
Altruiusm	A noisy measure of altruism, measured by participant response to a question in the pre-study similar to the dictator and investment/trust games.
Trust	Participant reported trust on a Likert scale, the participant's response to the question "How much do you trust this pair of co-diners?"

Table 3.2: Dependent variables in the study.

Term Name	**Description**
Dining Points	For each round, the ratio of item quality (either 200 for hotdog or 400 for lobster) to share of the bill. Dining points summed up over the entire game give an indication of absolute performance.
Cooperation	In this version of Diner's Dilemma, the action of ordering lobster.
Defection	In this version of Diner's Dilemma, the action of ordering lobster.
Reciprocation	A game situation: following a hotdog order with a hotdog order, or a lobster order with a lobster order. For example, the Tit-for-Tat strategy is to reciprocate 100% of the time.
Forgiveness	The tendency of a co-diner to choose not to reciprocate a lobster order, i.e., that co-diner "forgives" a lobster order by choosing hotdog in the next round instead.
Betrayal	The tendency of a co-diner to choose not to reciprocate a lobster order, i.e., that co-diner "betrays" a hotdog order by choosing lobster in the next round instead.
Strategy Regime	One of three "regions" (refer to Figure 3.6) of co-diner behavior which define the optimal strategy for that region: either "cooperation," (center) "exploitation," (left-side) or "avoidance" (right side). The points between each region (where it doesn't matter if the participant cooperates or defects) are termed the "pivot points."

Table 3.3: Diner's Dilemma game terms.

Dining Points Gained for Round Outcomes			
	Both Cooperate	**One Cooperates**	**Neither Cooperates**
Player Chooses Hotdog	20.00	12.00	8.57
Player Chooses Lobster	24.00	17.14	13.33

Table 3.4: Dining points gained based on the actions of co-diners. The cost of a hotdog is $10 with a quality of 200, and the cost of a lobster is $30 with a quality of 400. Dining points are calculated as quality divided by share of the bill.

3.4.3 Payoff Matrix and Co-diner Strategies

The quality-cost ratio of items available in a valid Diner's Dilemma game must be chosen so that the selection is a real dilemma for the player. First, if the player were dining alone, ordering hotdog should maximize dining points. Second, players must earn more points when they are the sole defector than when all players cooperate. Third, the player should earn more points when the player and the two co-diners all defect than when the player is the only one to cooperate. The payoff matrix in Table 3.4 was used in the game and adheres to this ordering. The exact values (Hotdog $10 w/ 200 quality, Lobster $30 w/ 400 quality) were refined based on the pilot study, which examined how well the participants understood the mechanics of the game. Participants in the pilot reported that they were able to quickly divide these numbers in their head to see the trade-offs.

The eleven co-diner strategies inflicted upon participants were created by varying two parameters. The first parameter was the probability of opponent "forgiveness," or the probability that the co-diner will cooperate (order hot dog) given that the player previously defected (ordered lobster). The second parameter was the probability of opponent "betrayal," or the probability that the co-diner will defect given that the player previously cooperated. The extent of forgiveness or betrayal can be thought of in terms

of distance from completely reciprocal (Tit-for-Tat) behavior. The eleven configurations and the number of participants who completed each condition are shown in Figure 3.4.

The choice of the eleven co-diner strategies creates three regimes of optimal decision making (for the participant player) based on the chosen payoff matrix (see Table 3.4). The first regime is the "cooperation" regime: these are cases where the human player's optimal strategy is to always cooperate. This occurs in the Tit-for-Tat opponent strategy and close to it (low rates of either forgiveness or betrayal). The second regime is the "exploitation" regime: cases where the human player's optimal strategy is to always defect in order to exploit co-diners. This occurs with high co-diner forgiveness rates. The third regime is the "avoidance" regime: cases where the human player's optimal strategy is to always defect in order to avoid being exploited by co-diners. This occurs with high co-diner betrayal rates. Two points do not fall into any regime, which are called the "pivot points," these points occur at the boundary between the three regimes, where cooperation being optimal switches to defection being optimal and vice versa. At these two points, there was no optimal strategy for the human player, since the long term performance average does not differ between hot dog and lobster.

3.4.4 User Interface

For the purpose of our study, we avoided showing the user information that might be considered an opaque recommendation or expert opinion, potentially biasing them towards cooperation or defection, in line with literature on system transparency and explanatory interfaces [112, 18, 113]. Instead, participants were shown one of three configurations of the UI with varying amounts of information. The Level 1 UI only displays the bare minimum amount of information necessary for a participant to perceive the environment, although clever users would still be able to achieve SA levels 2 or 3

by paying close attention or taking extra time to perform an analysis. The Level 2 UI aids comprehension of opponent behavior by displaying an enumerated game history that participants can examine to get a quick synopsis of opponent behavior from the outset to the current round. Finally, the Level 3 UI included a tool that allows a participant to create "what-if" scenarios for the long term gains of their choices.

- Level 1 UI (no support, see green box of Figure 4.1): All participants were shown, at a minimum, their current dining points, the food quality and cost of each menu item, the current round, and the results from the previous round in terms of dining points. This view explicitly reports on only the most current and recent game states, leading us to hypothesize that the participants would not be able to keep track of opponent behavior as easily as subjects using the more advanced interfaces, although it is possible that participants could resort to pen and paper to achieve the functionality available in the more complex interfaces.

- Level 2 UI (history, see blue box of Figure 4.1): This UI level includes all UI features from Level 1 UI, and adds a "History" panel to provide historical game information to the participant. In our first experiment, [93] presented both the participant and opponent score in a game history panel. Their results showed a drop in participant cooperation when the history panel was presented. Based on their observation that presenting opponent score can promote retaliatory behavior, we omit the score display feature from our user interface design.

- Level 3 UI (history + projection, see red box of Figure 4.1): This UI level includes all UI features from Level 1 and 2 UIs, and adds a "Projection" panel to provide long-term projection information. In this panel, the participant can enter his or her assumptions about opponent reciprocation behavior and calculate the expected dining points. The designers intended these assumptions to be drawn from the

Level 2 UI, but other assumptions can be entered at any time to explore the payoff space. By default, nothing is selected, so as to avoid biasing the participant in either direction.

3.4.5 Participants

The Diner's Dilemma game was deployed on AMT and data was collected from 901 participants. Only complete participant records were included in our analysis (35% of the participants dropped out and these records were discarded). AMT is a web service that provides attractive tools for researchers who require large participant pools and inexpensive overhead for their experiments. Numerous experiments have been conducted, notably [114], assessing the validity of using the service to collect research data, and these studies have generally found that the quality of data collected from AMT is comparable to (and perhaps even better than) what would be collected from supervised laboratory experiments [115].

3.4.6 Training

An interactive training session was designed to insure that participants had a basic understanding of the game and the interface before they could proceed to the game trial. First, the following description of the game was given:

> *"In this game, you will be dining with two co-diners multiple times at the same restaurant. Every round, you must choose to order either Hotdog (cheap but low quality) or Lobster (expensive but high quality). You can assume that you have enough money to dine out each time (your cash will not run out), however, you still prefer to save money! You have agreed to split the bill every round regardless of what items are ordered, and your overall performance is measured in terms of*

> *dining points, which is the ratio of food quality to money spent. If everyone orders Hotdog, you can get 20 points per round, but if you order Lobster and both your co-diners order hotdog, you can get 24 points (but your co-diners will lose out).*"

Next, information about the interface was provided through tooltips and components of the game were iteratively added as the participant proceeded through the training (Figure 3.5). Participants could play as many practice rounds as they wanted against two opponents playing Tit-for-Tat with 10% noise (i.e. simulated co-diners deviated from Tit-for-Tat 10% of the time), but were eventually prompted to complete some comprehension questionnaires (up to 3, one for each UI level). The information required to answer each question was explicitly available at the time the questionnaire was presented, and participants were allowed to submit answers as many times as they needed to complete the questionnaire. At the end of the training session, participants were allowed to continue using the interface as long as they liked before advancing to the next portion of the study.

3.4.7 Game Trial

Once the training session ended, participants were prompted that the game was about to begin and that they would be paid according to their performance in terms of dining points. After rounds 50 and 90, the game was stopped and the participant was required to respond to a short questionnaire test of the current state of the game. This allowed understanding of how the user interface was affecting proficiency of the game and awareness of the opponent strategies at different points.

3.5 Results

We considered the effect of varying the UI Level and opponent strategy on the co-operation rate, participant performance in terms of the participant's closeness from the

optimal strategy, situation awareness, and trust. An analysis of each relationship is given and then we construct a pathway model over all of our dependent variables to better understand how our measurements relate to each other.

3.5.1 Demographics

In the Diner's Dilemma experiment, the AMT participant age ranged from 18 to 75 with an average of 32. 49% of participants were male while 51% were female.

3.5.2 Cooperation Rate

Figure 3.6 shows the average participant cooperation percentage for each opponent strategy, grouped by UI level. Recall that forgiveness rate indicates the rate that simulated co-diners will respond to a lobster order with a hotdog order, and betrayal rate indicates the rate that simulated co-diners will respond to a hotdog order with a lobster order. Take notice of the three regimes divided by the pivot points: from left to right, the "exploitation," "cooperation," and "avoidance" regimes respectively.

3.5.3 Performance

How is UI level and performance related? Figure 3.7 shows how the UI level affected participant performance in terms of the optimal player strategy for each opponent strategy condition (refer to Figure 3.4). Participant performance was calculated as the closeness of the participant's cooperation proportion and the "optimal" proportion of cooperation responses (either 0 or 1) for the given opponent strategy condition. For example, in the pure Tit-for-Tat condition, the optimal strategy is to cooperate 100% of the time. If a participant's cooperation rate is 70% in this condition, this would produce a closeness score of 0.7 (this score would be 0.3 in a condition where 100% defection

was the best course of action). In this analysis, pivot points are excluded, as by design, cooperation and defection produce the same expected score, so there is no single optimal strategy in these two conditions. Smaller deviations from optimality indicate better performance in this analysis. Figure 3.8 plots observed participant behavior against the theoretical optimal strategy. Figure 3.7 shows that closeness to optimal behavior tends to increase (indicating improved performance) with increasing UI Level. A separate regression analysis (not shown) showed a significant linear improvement in performance with increasing UI Level ($b = -0.03, t(743) = 2.47, p = 0.01$).

Next, how are UI level and SA related? Results from the first and second SAGAT freezes are shown in Figure 3.9, across the three UI levels. Recall that these results indicate the participants performance on the questionnaire that was given during the game trial, and related to both general game strategy, the current and past states of the game, and asked the participant to project into the future.

An ordered logistic regression showed the effect of UI Level on the first and second test scores ($p < 0.001$ for each test). We also considered if there was an increased learning effect between SAGAT 1 and SAGAT 2, for each of the UI levels, however, no significant trend was found. A two-way mixed ANOVA showed a significant main effect of UI level upon SAGAT test scores ($F(2, 905) = 3.69, p = 0.03$), with scores improving with increasing UI level. There was also a main effect of SAGAT test time on score ($F(1, 905) = 11.91, p < 0.001$), with scores improving from the first to the second SAGAT questionnaire.

3.5.4 Trust

In this section, we examine differences between trust that participants reported in their co-diners during the post-study and observed trust in the game. For the purpose

of this discussion, we are using observed cooperation in the game as a proxy for trust, assuming that participants are aware that if the optimal state for the overall group is for everyone to always cooperate. So, our definition of trust is that the participants trusts in the group as a whole to achieve that optimal group-wise goal. Figure 3.11 shows the mean trust reported by users in each strategy group. The number used to compute each mean is shown as n on the x-axis.

Figure 3.10 shows the level of participant cooperation (y-axis) grouped by the co-diner strategy that they played against (x-axis). Note that the x-axis in this figure contains two scales –a betrayal rate and a forgiveness rate. At the center point, the 0-value means that the co-diners algorithm plays an unbiased Tit-for-Tat strategy. We apply this visual approach to allow for quick comparison of participants' reactions to different rates of forgiveness and betrayal. Figure 3.11 follows the same setup, with self-reported trust on the y-axis. These plots highlight two useful results. First, in both self-reported (trust) and observed (cooperation) cases, any bias towards betrayal by co-diners is quickly responded to with betrayal by participants (i.e: the sharp drop-off to the right of the peaks in both plots). Second, reaction to bias in forgiveness rate of the co-diners is not as pronounced as reaction to betrayal, and is not the same pattern for reported trust and observed cooperation. Note that the slope to the left of the peak (forgiveness side) of Figure 3.11 is flat compared to that in Figure 3.10. This is most likely a result of capitalization behavior, wherein participants trust that co-diners will cooperate, and then perform a betray move to increase their own dining points at the expense of the group.

Regressand	**Regression (←) or Covariance (↔) Term**	**Estimate (β)**	**Std. Error**	**P-value**
Cooperation ($R^2 = 0.26$)	← Forgiveness	-0.29	0.037	***
	← Betrayal	-0.62	0.037	***
	← Situation Awareness	0.15	0.029	***
Situation Awareness ($R^2 = 0.05$)	← Forgiveness	-0.16	0.033	***
	← UI Level	0.11	0.040	**
	← Trust Propensity	-0.11	0.033	**
Trust ($R^2 = 0.26$)	← Betrayal	-0.46	0.032	***
	← Cooperation	0.07	0.031	*
	← Trust Propensity	0.14	0.029	***
Performance ($R^2 = 0.07$)	← Forgiveness	0.12	0.041	**
	← Betrayal	0.25	0.041	***
	← Situation Awareness	0.14	0.032	***
	← Altruism	-0.08	0.032	*
	↔ Trust (reported)	-0.12	0.028	***

Table 3.5: A pathway model built on all study variables. Regressands (left-hand side variables) are shown in the left column, with all regression terms (right-hand-side variables, or regressors) shown in the second column. The regressand can be expressed as a linear sum of the regression terms multiplied by their coefficients, which are shown in the "Estimate" column. Except for UI (which ranges from 0-2), all variables were standardized. Recall that "Performance" is measured as closeness to optimal rather than the raw dining score. Significance levels for this table: *** $p < .001$, ** $p < .01$, * $p < .05$. Model fit: $N = 901$ with 22 free parameters = 41 participants per free parameter, $RMSEA = 0.039$ ($CI : [0.021, 0.058]$), $CFI = 0.974$, $TLI = 0.944$ over null baseline model. $\chi^2(12) = 28.592$.

3.5.5 Path Analysis

The results of fitting the data to a path model [51] are shown in Table 3.5. This model was constructed by ordering all variables in the study into groups based on their causal relationships (e.g., observed participant cooperation cannot affect the UI treatment that was assigned) and then saturating all regressions with all available variables. This saturated model was then trimmed of insignificant effects to produce four candidate models. The model shown in Table 3.5 had the best (lowest) Bayesian Information Criterion (BIC)/Akaike Information Criterion (AIC) score of all tested models. Figure 3.12 shows a visualization [116] of the model with regression coefficients (β) and covariance terms.

From Table 3.5, we can see that both co-diner forgiveness and betrayal parameters had an effect on the participant's cooperation. We can also see that participants were sensitive to the different opponent strategies to different degrees. This decrease appears to be steeper with increased betrayal than it is with increased forgiveness, with the coefficient for the effect of betrayal ($\beta = -0.62$) being about twice as large as the coefficient for forgiveness ($\beta = -0.29$). This effect can also be visually observed in Figure 3.8. Additionally, participants who performed well on the SAGAT questionnaires (SAGAT 1 score was typically within 95% of SAGAT 2 score, refer to Figure 3.9) cooperated more ($\beta = 0.15$).

Situation awareness was taken as a linear sum of the scores on both SAGAT questionnaires. The total SA score was influenced by co-diner forgiveness ($\beta = -0.16$), UI level ($\beta = 0.09$), and trust propensity($\beta = -0.11$). Reported trust in co-diners (taken during the post-study) was profoundly impacted by the degree of betrayal the participants endured ($\beta = 0 - .46$). Participants that cooperated more also reported higher trust in co-diners ($\beta = 0.07$). Final performance, in terms of closeness to optimal strategy, was

beneficially affected by forgiveness ($\beta = 0.12$) and betrayal ($\beta = 0.25$). Participants with higher SA performance also performed better ($\beta = 0.14$), but altruistic participants performed worse ($\beta = -0.08$). Reported trust and performance were negatively correlated ($\beta = -0.12$).

3.6 Discussion

This section will contextualize related work with the new results that were gathered and draw implications about the relationship between the user interface, situation awareness, cooperation, and performance.

3.6.1 Situation Awareness and the User Interface

First, in [93], the inclusion of more UI support did not necessarily increase cooperation and in some cases actually resulted in a decrease. This could be due to less effective user interface support and information, but the relationship between situation awareness and cooperation was not thoroughly explored. However, it was still identified that participant reported trust in co-diners and cooperation proportion were correlated, as was also the case in this experiment. For strategies in this previous experiment that discouraged cooperation, we saw that the UI encouraged cooperation even less than in the present results, most likely for the similar reason that the desire to punish bad behavior overcomes the influence of information from the user interface, and possibly the desire for improved individual outcome.

Second, [94] explored the relationship between situation awareness and cooperation rates. The level 2 and level 3 user interfaces in that experiment, which differed slightly from the current user interfaces, had also increased performance and cooperation. We found that situation awareness and dining points were highly correlated. However, the

range of opponent strategies in [94] was limited, but cooperation behavior observed in the treatments with the more defection prone strategies opened the question of whether or not the UI could continue to maintain high SA (and therefore cooperation) even when co-diner betrayal increases.

Measuring SA allowed for the decoupling of user interface design and dependent metrics of interest. In this case, we were interested in measuring cooperation or performance, but when SA is used as a mediator between UI and these other dependent variables we found that our model fit the data better. This makes sense from a practical standpoint - if information is shown but is not internalized by users, the user interface cannot have an effect on decision making. Effective UI support that leads to increased SA can then lead to better decisions by individual participants. Thus a goal of user interface design could be to increase SA for a given task and researchers could eschew approaches that treat the participant as a black box, which bypass SA measurements and only consider performance.

It should be noted that, according to our model (Table 3.5), participants in more forgiving strategies displayed lower SA. Also note that participants performed worse overall in terms of closeness to optimality in the "exploitation" regime. The lower SA could be due to a lower perceived need for strategizing in these conditions. Participants with higher trust propensity also displayed lower situation awareness. These participants may have strategized less overall and may have spent more rounds attempting to get co-diners to cooperate.

3.6.2 Cooperation

Due to the nature of the chosen co-diner strategies and payoff matrix, higher participant cooperation in the game always maximized group benefit. Thus, participant

cooperation can be seen as a measure of total group outcome.

In general (across all co-diner strategies), more UI support lead to more SA which meant more cooperation. However, when looking at Figure 3.6, it can be seen that the effect of the UI was also dependent on co-diner strategy. More SA also increased participant responsiveness to changes in opponent behavior, which is shown by the overall increased performance. Looking again at Table 3.5, it can be seen that high SA increased cooperation overall. However, if we refer back to Figure 3.6, we can see that the effect of the UI varied based on co-diners strategy. In the "avoidance" regime (right side) participants using the level 1 UI consistently displayed higher cooperation. This could be due to the fact that the History Panel made the patterns of co-diner defection much easier to see. In the "exploitation" regime (left-side), the effects of the UI appeared to have less impact. Still, in the "cooperation" regime, the UI seemed to have the highest impact on participant cooperation. To verify these apparent interaction effects between UI and co-diner strategy we tested the following regression model:

$$\begin{aligned} cooperation =& \beta_1 ui2 + \beta_2 ui3 + \beta_3 exploitation + \beta_4 avoidance + \\ & \beta_5 ui2 * exploitation + \beta_6 ui2 * avoidance + \\ & \beta_7 ui3 * exploitation + \beta_8 ui3 * avoidance \end{aligned} \tag{3.1}$$

This multivariable regression (adjusted $R^2 = 0.24$, $p = 0.00$) revealed the effects of UI level 2 ($\beta_6 = -0.12, p = 0.03$) and UI level 3 ($\beta_8 = -0.14, p = 0.01$) in the "avoidance" regime. In the "exploitation" regime, effects of UI level 2 ($\beta_5 = -0.02, p = 0.81$) and UI level 3 ($\beta_7 = -0.13, p = 0.06$) were non-significant and marginal, respectively. This analysis suggests that different levels of UI support would be appropriate in different situations to maximize group benefit. It is possible that showing less information would

encourage group members to avoid all-defect situations when one or more agents act selfishly. A follow-up study using three human co-diners would be necessary to investigate this further.

3.6.3 Performance

Our co-diner strategies were quite simple (drawn from an independent and identically distributed random variable), so participants in the game were not able to sway co-diners into cooperation by consistent contribution behavior. The user interface was designed to increase the situation awareness and performance of the individual using it, without regard for the group's well-being. As noted in the previous section, participants using the level 1 UI still attempted to punish forgiving opponents, but to a marginally lesser extent than the participants in UI level 3. If the performance of the entire group is of concern, it may be prudent to design a UI that attempts to maximize that objective, rather than the individual's performance.

The model described in Table 3.5 found that participants were far more likely to retaliate to defection than they were to exploit over-forgiving co-diners. Figure 3.13 shows the observed difference in cooperation between participants in the far-left and far-right regions of Figure 3.6. As you move from the left to the right, you can see how many points are lost by deviating from the optimal strategy under these conditions (100% betrayal). Initially, we hypothesized that participants in the far left and right regions would react to the co-diner strategies in the same way, that is, the lines for each region in Figure 3.13 should be more or less the same. The observed behavior was much different: participants simply did not exploit the simulated co-diners as much as they could despite the obvious personal benefit, and participants in the high-betrayal conditions were far more likely to behave closer to the optimal strategy.

3.7 Summary

We designed a decision support interface for the Diner's Dilemma and conducted an N=901 experiment to study human decision making with various levels of user interface support. We found that: 1) participants were more likely to retaliate against defection than they were to initiate defection, especially when more UI support was given; 2) participant SA and cooperation increased as more UI support was given but decreased in the presence of more forgiveness from co-diners; and 3) performance increased when SA was higher and co-diner strategy was more extreme, especially when co-actors betrayed more. Empirical results from the experiment conducted can help user interface designers to encourage cooperative behavior in social settings or maximize the benefits to an individual decision maker.

At the beginning of this chapter we wrote that 1) decision support systems can positively impact situation awareness which in turn 2) improves decision making. We saw that increased UI level increased performance significantly on our global SA freezes (Figure 3.9), and that participants using the higher level UIs made significantly better decisions that the users in UI level 1 (Figure 3.7). Furthermore, our pathway model showed that situation awareness positively impacted game performance (Figure 3.12). It is important to point out here that the measure of *global* SA that was conducted here measures *insight* in the framework in chapter 6, in addition to co-diner and interface SA.

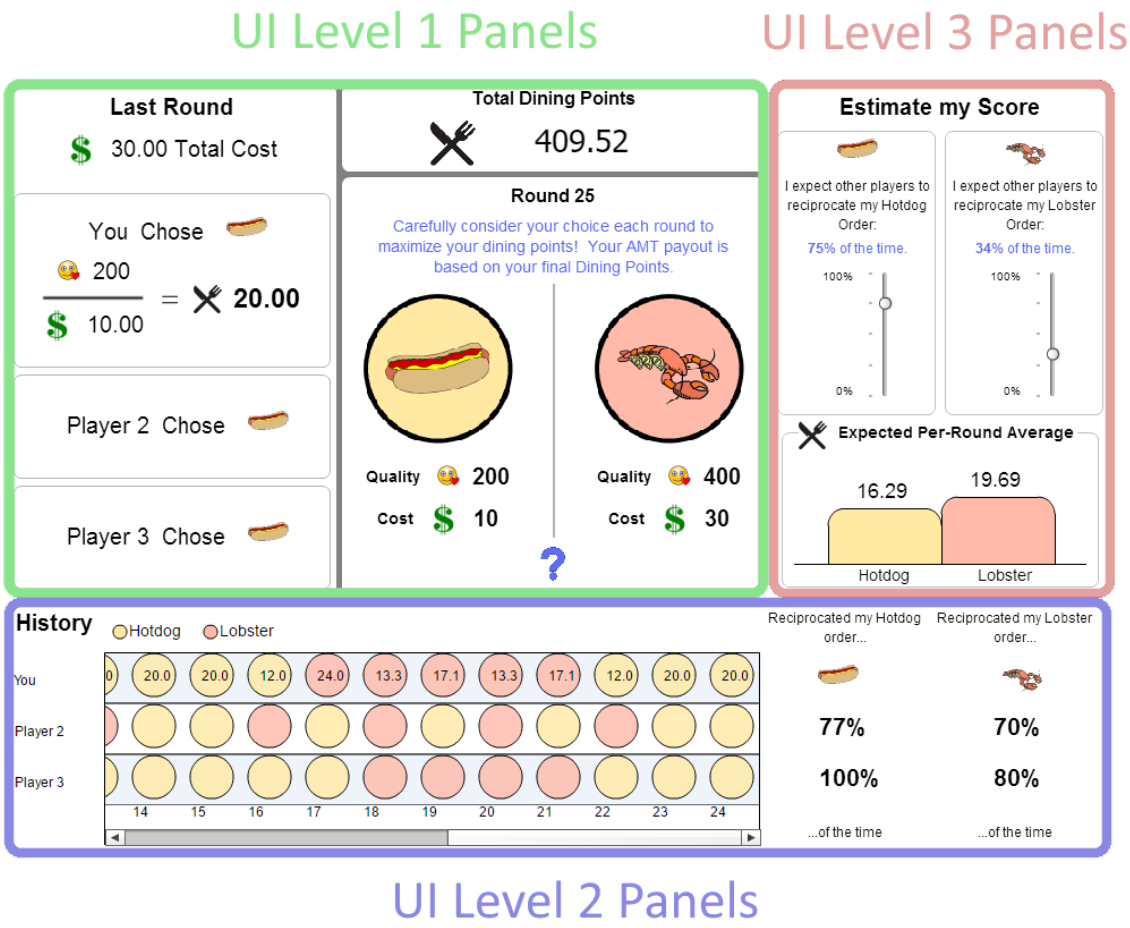


Figure 3.2: The user interface for the Diner's Dilemma game. Participants either interacted with UI Level 1, 2, or 3. Panels are indicated in green, blue, and red. Participants saw all panels below their level, for instance, participants in UI Level 2 saw the UI Level 1 and UI Level 2 Panels.

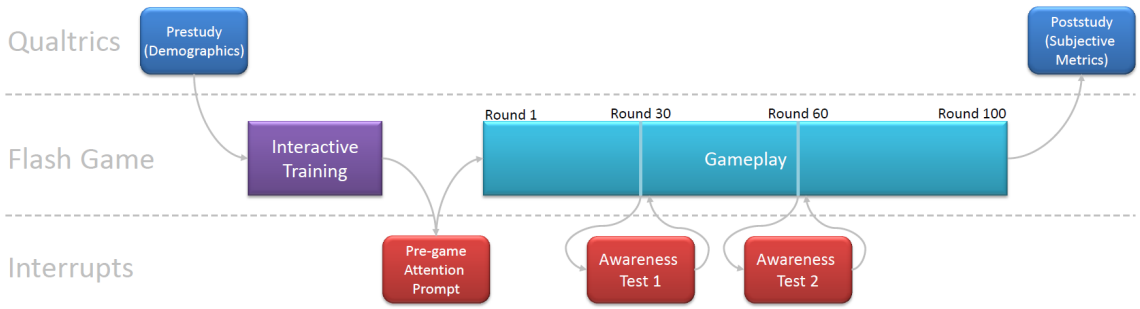
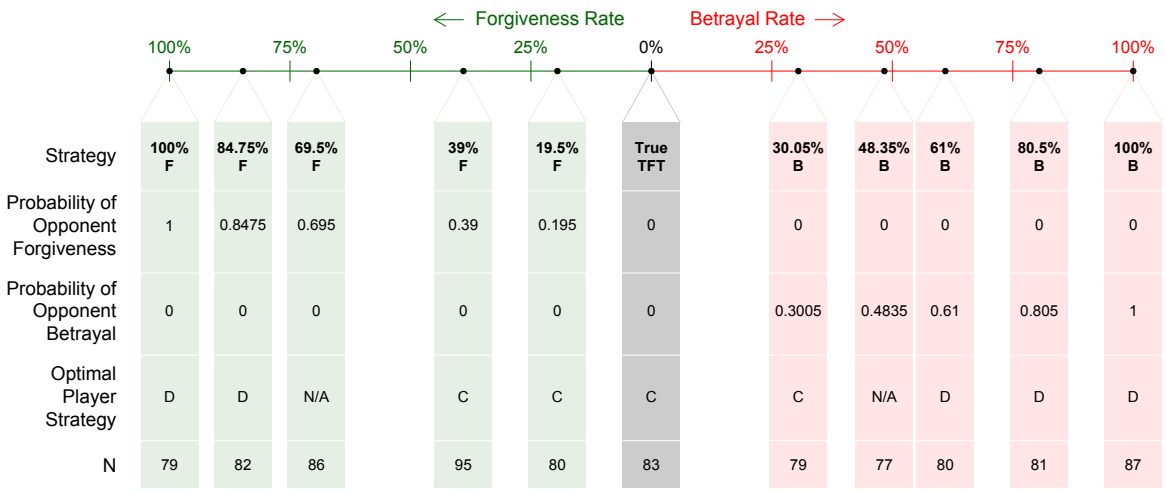


Figure 3.3: A participant's experience through the system. Participants were recruited from Amazon Mechanical Turk, and spent the majority of the time learning or playing the Diner's Dilemma game as seen in Figure 4.1. A prestudy and poststudy collected demographic metrics, while two popups tested participant knowledge at unexpected times during the game. Since participants were paid more for better scores but were not penalized during training, a prompt alerted them when the training ended and when they began to play for points.



Strategy	100% F	84.75% F	69.5% F	39% F	19.5% F	True TFT	30.05% B	48.35% B	61% B	80.5% B	100% B
Probability of Opponent Forgiveness	1	0.8475	0.695	0.39	0.195	0	0	0	0	0	0
Probability of Opponent Betrayal	0	0	0	0	0	0	0.3005	0.4835	0.61	0.805	1
Optimal Player Strategy	D	D	N/A	C	C	C	C	N/A	D	D	D
N	79	82	86	95	80	83	79	77	80	81	87

Figure 3.4: Quantities of participants in each opponent strategy condition. At 48.35% Betrayal and 69.5% Forgiveness, player choice does not matter statistically, as the expected dining points in each round is the same for the choice of hotdog or lobster.

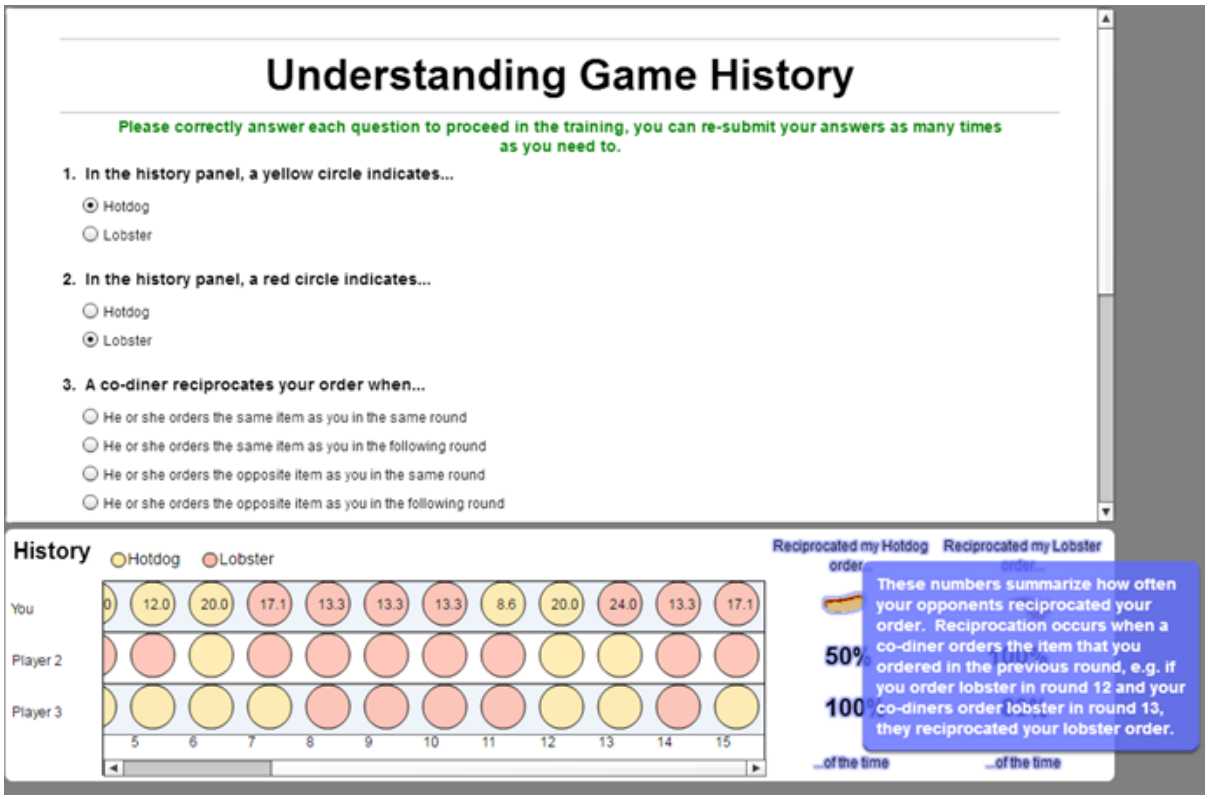


Figure 3.5: During the training session, participants were could spend as much time as needed collecting information about the interface and playing an indefinite number of practice rounds. They were intermittently re-prompted to proceed to the next part of the study, which required the correct answer of multiple choice questions about game concepts and information in the interface. Use of UI Level 2 and 3 required the understanding of the concept of reciprocity, which was tested. Participants in the pilot study had an easier time with this conceptualization than two other variants of the same information.

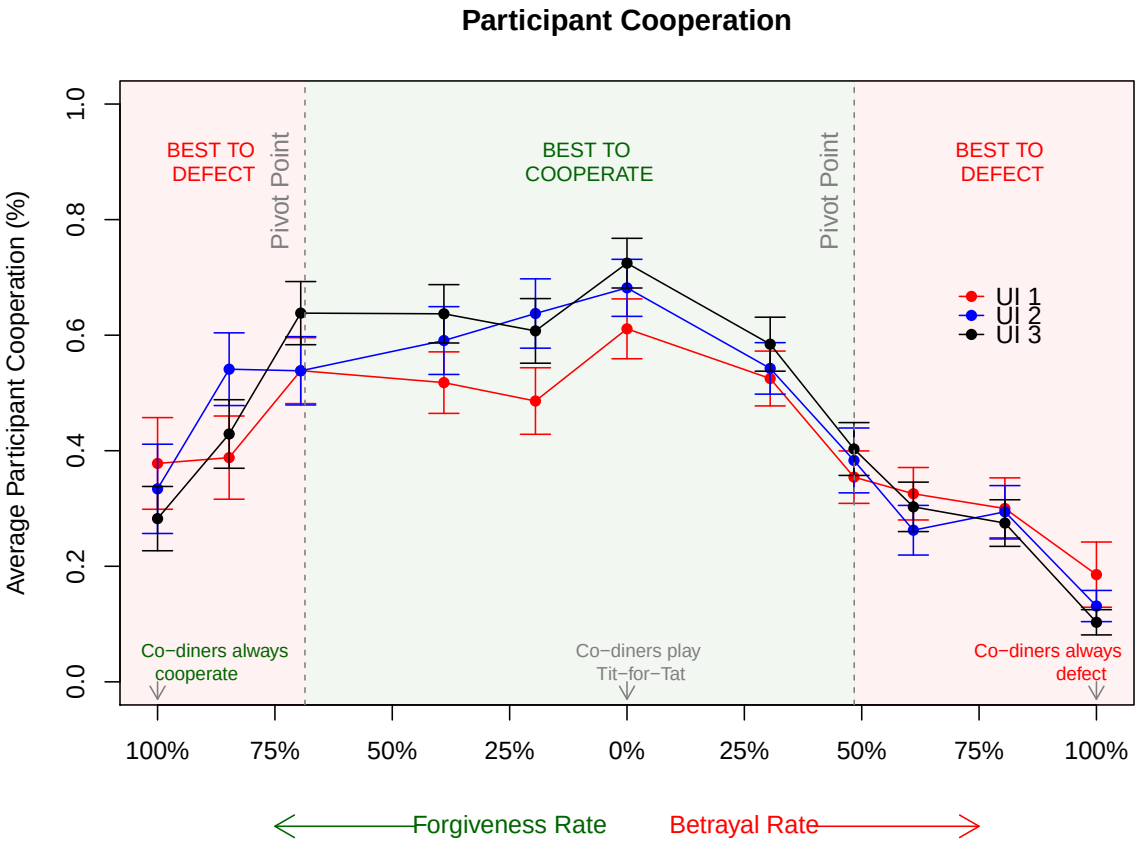


Figure 3.6: The cooperation percentage for each opponent strategy, grouped by UI Level, error bars are 95% confidence interval. Forgiveness rate indicates the rate that simulated co-diners will respond to a lobster order with a hotdog order, and betrayal rate indicates the rate that simulated co-diners will respond to a hotdog order with a lobster order. Note the three regimes divided by the pivot points; in the far left and far right regions the optimal strategy would be to defect 100% of the time, and in the central region 100% cooperation would yield the best performance. Observed behavior differed from this theoretical maximum.

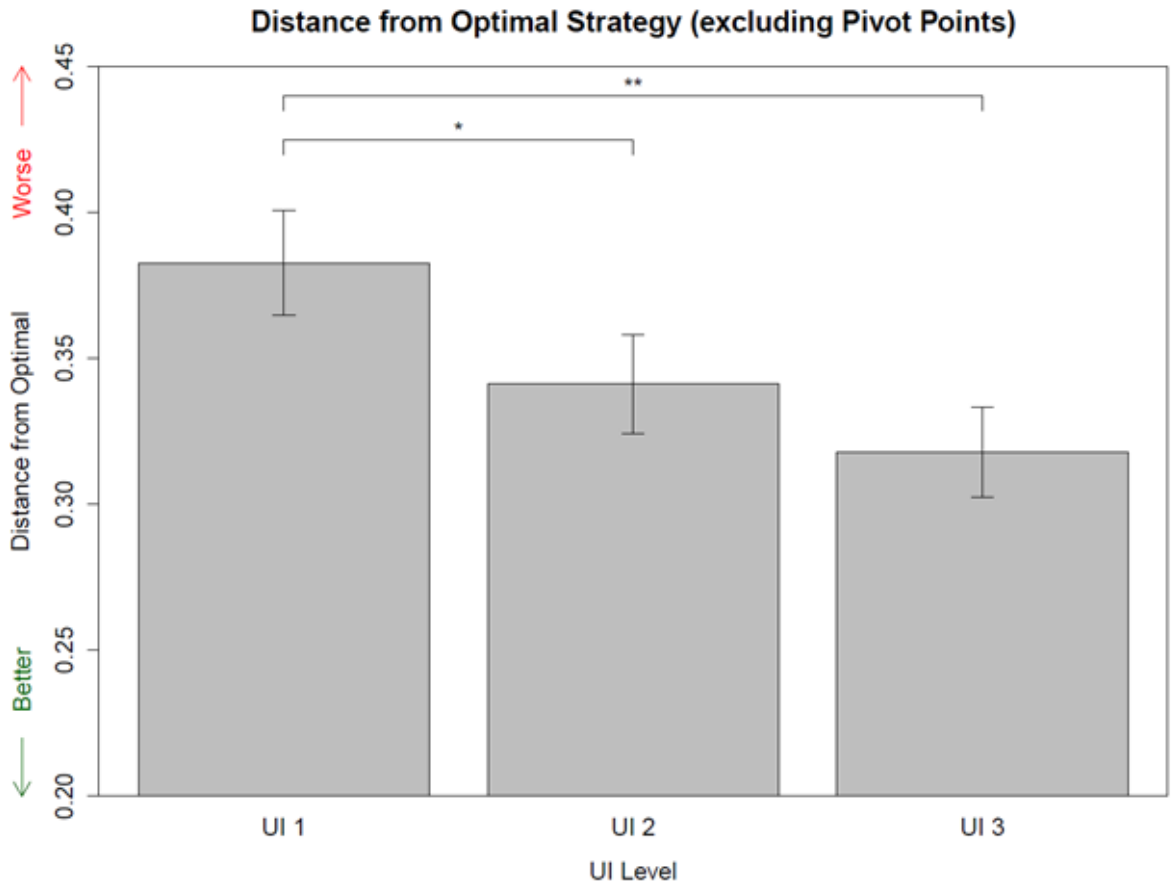


Figure 3.7: Closeness to optimal strategy, error bars are 95% confidence interval. Participants were much more likely to make an optimal decision when using UI level 3. Pivot points were excluded from this analysis, as player choice becomes arbitrary under those conditions. Significance levels: * = 0.05, ** = 0.005

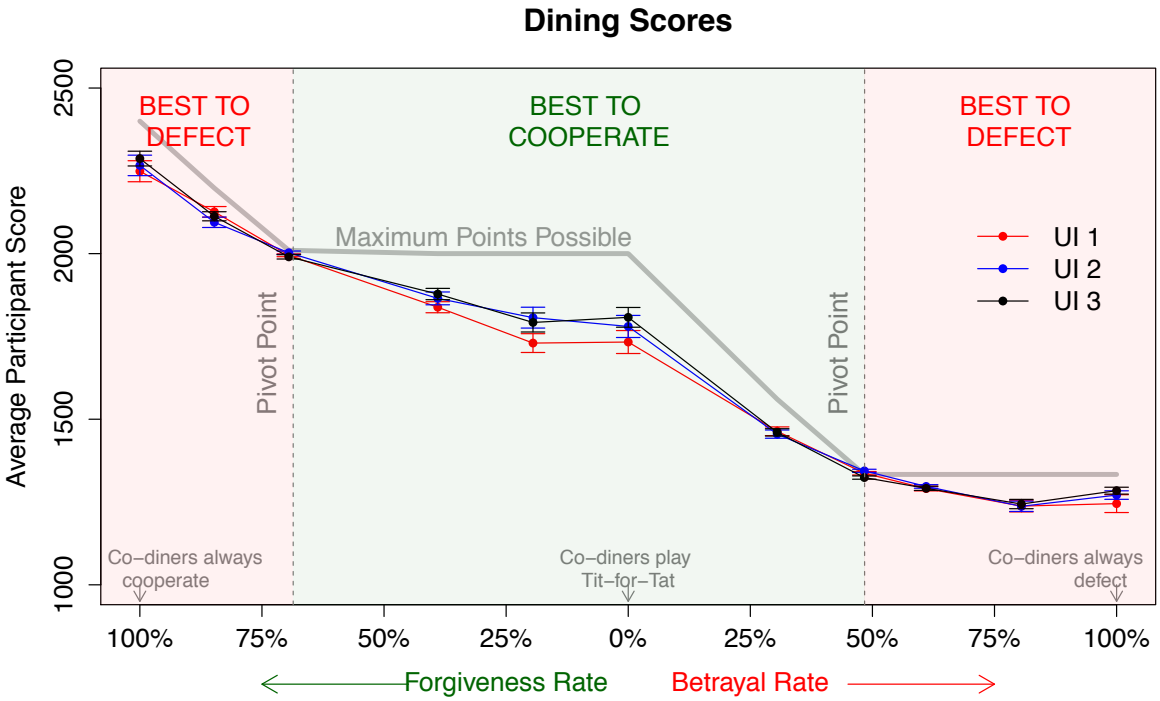


Figure 3.8: Dining points earned relative to the maximum available. Data is grouped by UI level. Players actually perform worse compared to optimal in the central "co-operation" regime, which is indicated by a larger gap between the gray line and the observed red/blue/black lines. At the pivot point, player choice no longer matters so we see the theoretical and observed behavior converge.

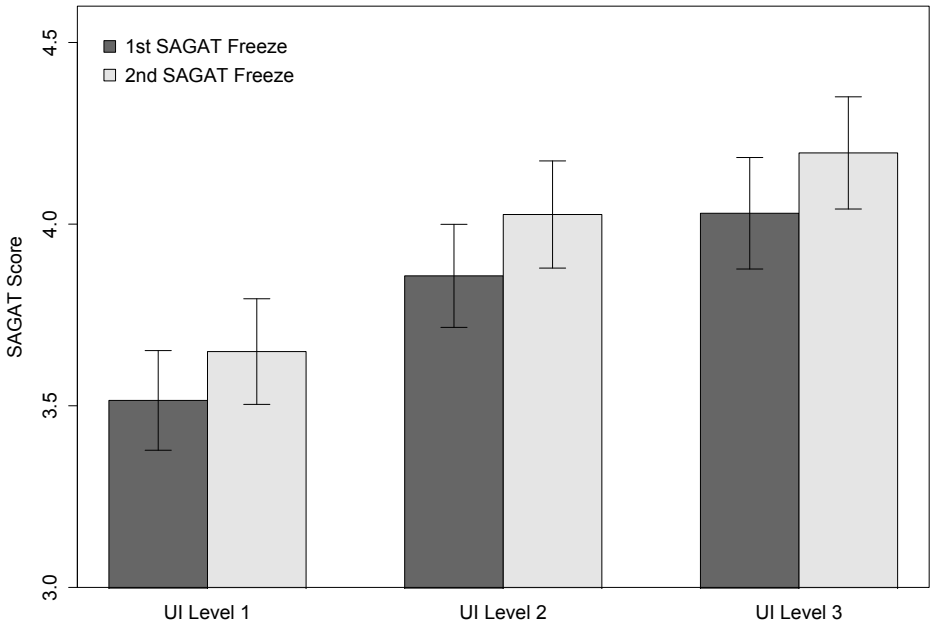


Figure 3.9: Average scores for UI Level 1 were (3.5,3.65), UI Level 2: (3.86, 4.02), UI Level 3: (4.02, 4.19). Error bars are one standard error.

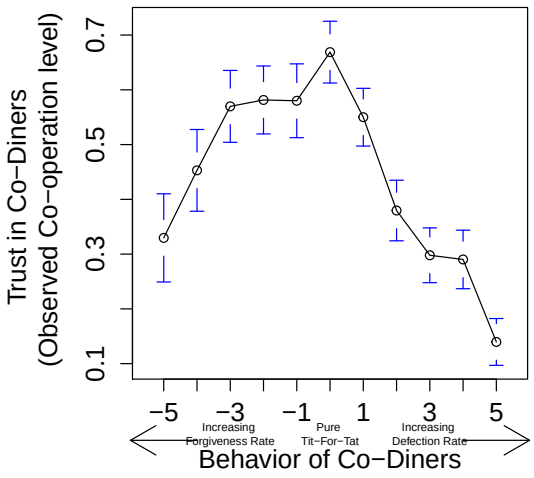


Figure 3.10: Cooperation observed from participants.

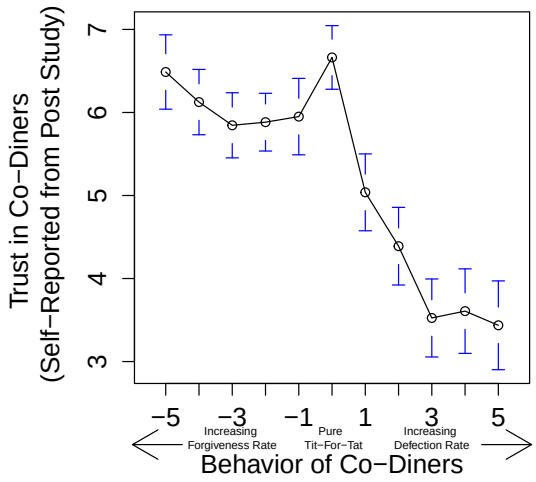


Figure 3.11: Self-reported trust in co-diners

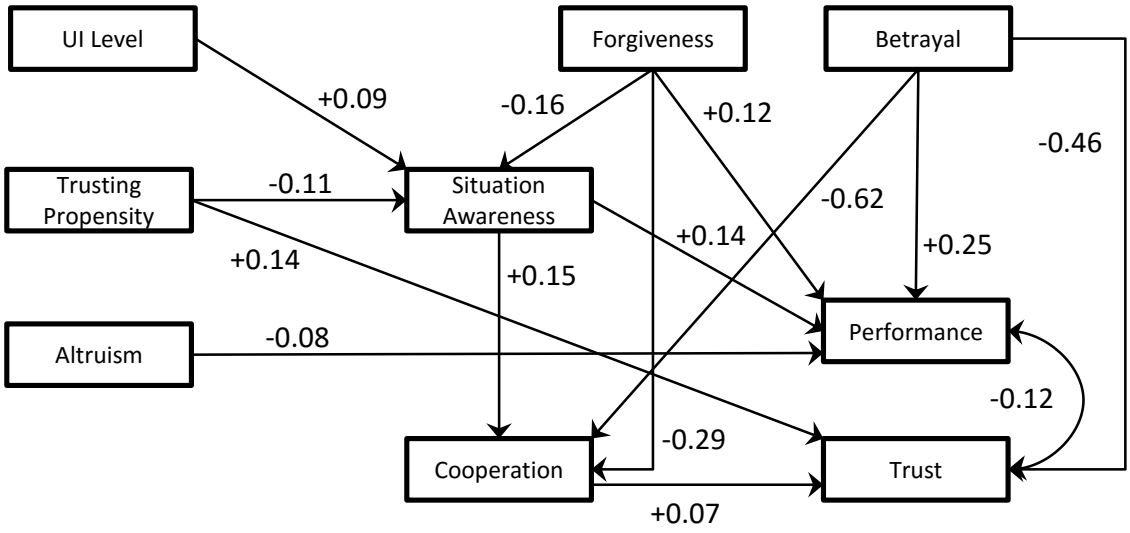


Figure 3.12: A network visualization of the pathway model. P-values and standard error values are shown in Table 3.5

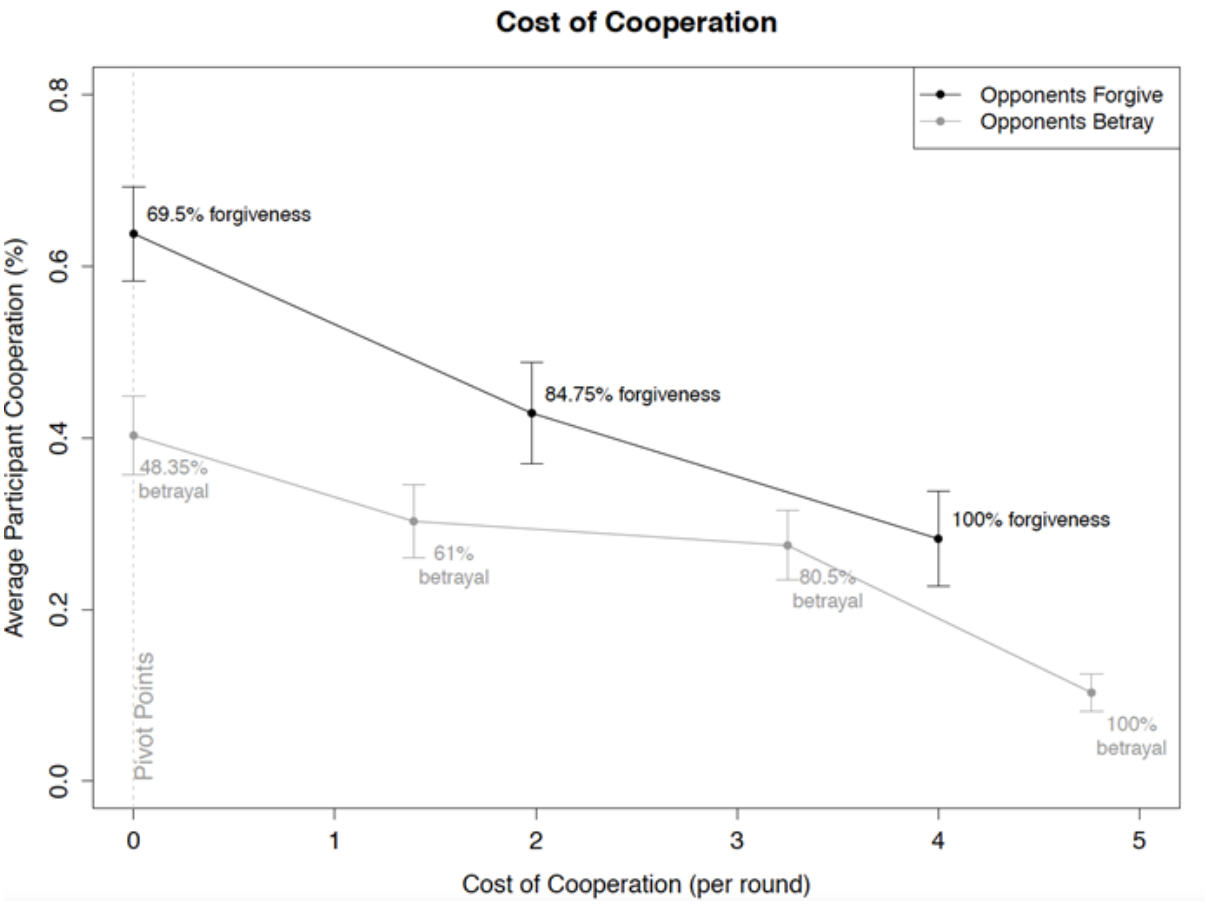


Figure 3.13: Observed difference in cooperation between participants who dealt with forgiving or exploitative co-diners. Error bars are 95% CI. Groups are broken down by opponent strategy (as in Figure 3.6), and the vertical axis represents observed cooperation percent for that group. The horizontal axis indicates the points lost when a participant cooperates 100% when the co-diner strategy is fixed, thus, the far left line indicates the pivot points, and the strategies between the pivot points in Figure 3.6 are accordingly not plotted. The dark black line represents the far left side of Figure 3.6, and the light grey line represents the far right side. The optimal strategy for all points graphed here should be to cooperate 0% of the time, but this was not observed.

Chapter 4

Dynamic Feedback from Collaborative Filtering: Hypothetical Recommendation

In this chapter, we'll show how improved explanatory feedback from a recommender can improve the quality of user actions when given equivalent control over the input to a recommendation algorithm. Here, users interacted with a collaborative filtering algorithm by manipulating their movie profile to get recommendations for new movies to watch. Unlike other studies in this dissertation, decision success is not measured due to lack of alternative options and participant choice - only subjective satisfaction with algorithm output is measured. Results support the following claims:

- Feedback explanation facilities can encourage users to interact with a recommendation algorithm
- Feedback explanation facilities can aid users when choosing control parameters
- Feedback explanation facilities increase perceived trust and accuracy in recommen-

dations after a profile update task

This work was previously published at the FLAIRS conference in 2015 [117].

4.1 Introduction to Explanation in Recommender Systems

Recommender systems have evolved to help users get to the right information at the right time [118, 119]. In recent years, a number of researchers and practitioners have argued that the user experience with recommendation systems is equally, if not more, important than accuracy of predictions made by the system [120]. Research has shown that providing "dynamic feedback" (live updates to recommendations with explanation) to users during the recommendation process can have a positive impact on the overall user experience in terms of user satisfaction and trust in the recommendations and to accuracy of predictions [121]. In many real-world recommender systems, however, user profiles are not always up-to date when recommendations are generated, and users could potentially benefit from adding, removing, and re-rating items to reflect current preferences. While researchers have explored the effects of conversational recommender systems [122], these studies focus on a granular refinement of requirement specifications for individual product search during an ad-hoc session. In this paper, we focus on evaluating how the experience of the recommendation consumer is affected by using low-cost, exploratory profile manipulations (an addition, deletion, or re-rate) on a pre-existing profile to generate what we call "hypothetical" recommendations. These are scenarios that allow a user to update a stale profile by asking questions of the form "what if I added product x?", "what if I rated these 10 songs?" Specifically, this paper describes a study involving 129 participants, designed to answer the following three research questions:

1. How does dynamic feedback affect which type of profile updates users perform?
2. What is the effect of different types of profile updates on recommendation error?
3. What is the effect of dynamic feedback on *perceived* accuracy, satisfaction, and trust?

Previous work on profile elicitation for collaborative filtering systems has focused on passive [123] and active [124] approaches. The experiments discussed in this paper can also be classed as a form of active profiling: the user is encouraged to add, delete, and re-rate items by assessing the feedback from the recommender while updating their preference profile. This research also considers the impact of interactive feedback for eliciting and encouraging profile manipulations from the user. Before we proceed with our discussion of the experiment itself, the following sections frame the experiment in the context of previous research on explanation and interaction aspects of recommender systems.

4.2 Related Work in Recommender Systems

Engaging Users

The majority of research in recommender systems is focused on improving recommendation algorithms (e.g. [125]), without specific focus on user experience. This research builds on a number of related research efforts that deal with visualization, interaction and control of recommender systems. Earlier work by Herlocker [78] demonstrated that explanation interfaces for recommender systems can improve the user experience, increasing the trust that users place in the system and its predictions. Cosley et al. [112] build on the explanation study to explore how explanations can change the opinions of a

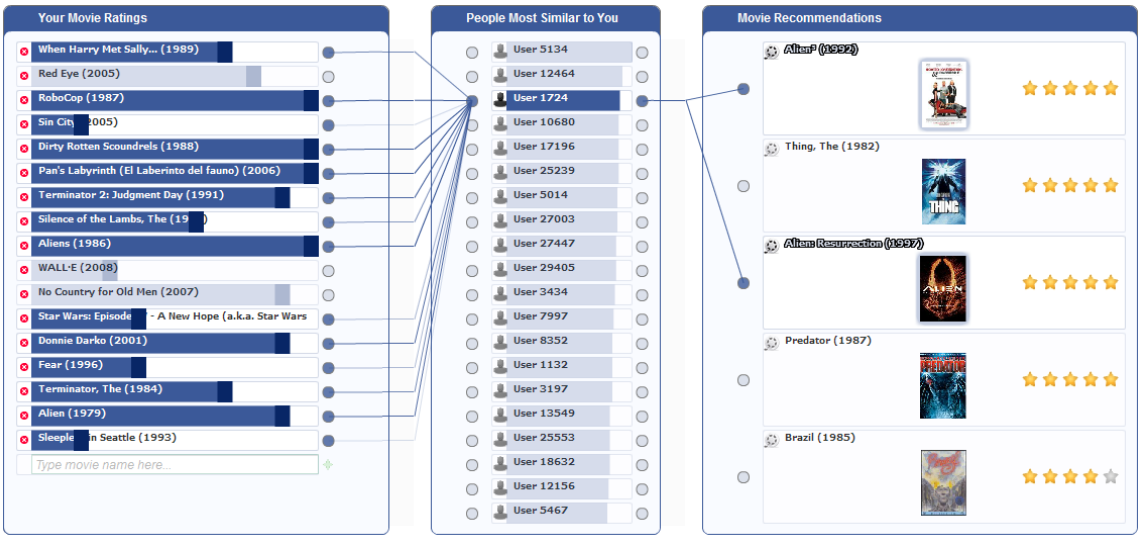


Figure 4.1: Screenshot of the interactive recommender system used in the experiment. From left to right the columns display: a user's profile items; top-k similar users; top-n recommendations. Adds, deletes or re-rates produce updated recommendations in real time. The dark blue lines appear when a user clicks a node and show provenance data for recommendations and the nearest neighbors who contributed to them. Clicking a movie recommendation on the right side of the page opens the movie information page on Rottentomatoes.com.

recommendation consumer, particularly in terms of rating behavior. They focus on consistency of re-rating behavior, impact of the rating scale and of dynamic feedback. Our experiment differs from Cosley's study [112] in that feedback is not explicitly controlled to be high or low quality, placing the focus on the true impact of hypothetical profile manipulations on the overall user experience. Work by Swearingen and Singha [126] finds that users tend to have higher trust in recommender systems that predict items that they already know and like. They posit two important considerations for interaction design: what user needs are satisfied by interacting and what specific features of the system lead to satisfaction of those needs? In the context of our experiment, we believe that the user "need" is a desire to explore and probe the information space, and that a low-cost "hypothetical recommendation" feature provided by an interactive visualization tool can fulfill this user requirement.

Interactive Recommendation Systems

Recent work in this area focuses on visual interactive explanation and control mechanisms for recommender system algorithms. O'Donovan et al. [127] describe an interactive visualization tool that supports genre-based manipulations of the k-nearest neighbors used in a collaborative filtering algorithm. They argue that "over-tweaking" can reduce the quality of recommendations if the interactive manipulations are not well balanced with the pre-existing user profile information. Bostandjiev et al. [121] describe a visual interface to a hybrid recommender system that supports user guided transitions between social and semantic recommendation sources, and this system is leveraged by [128] in an experiment to study the effect of inspectability and control in social recommender systems. In particular, Knijnenburg et al. finds that both inspectability and control have a positive impact on user satisfaction and trust in the recommender system, but they do not evaluate this effect as profiles are manipulated over time . Verbert et al. [129] further analyze the impact of information visualization techniques on user involvement in the recommendation process. Their evaluation of the Conference Navigator system [129] shows that the effectiveness of recommendations and the probability of item selection increases when users are able to explore and interrelate entities.

Our work differs from previous approaches in that we attempt to determine the individual impact of each *type* of profile manipulation (add, delete, re-rate) on recommendation error, and how dynamic feedback affects the frequency of each type being performed and any difference in magnitude when reducing recommendation error.

4.3 Experiment Setup

In this study, the interactive recommender system shown in Figure 4.1 was presented to participants, and they were asked to add, delete or re-rate items in their profile. The

Treatment	*First Phase*	*Second Phase*
1	Gathering	Manipulation (no dynamic feedback)
2	Gathering	Manipulation (w/ dynamic feedback)

Table 4.1: Breakdown of participant task and independent variables

Metric Name	*Explanation*
Manipulation	Participant's quantity of additions, deletions, and re-rates of profile items during the second phase of the task.
Rec. Error	Mean difference of ratings given by participants and ratings by the recommender.
Satisfaction	The participant's perceived satisfaction with the recommendations (1-100).
Trust	The participant's reported trust in the recommender (1-100).
Accuracy	The participant's perception of the accuracy of the recommender (1-100).

Table 4.2: Dependent variables in the study.

system recommended movies based on the MovieLens 10M dataset, through two different configurations of the user interface. The first group received dynamic feedback (on-the-fly recommendations after each profile update) while the second group did not receive feedback. Pre-existing profile information was retrieved by participants through a web service of their choice (Netflix, IMDb, etc.) and we asked users to rate recommendations from the system based on this initial profile as a benchmark. Ratings were given on a 1-5 star scale. Following this, users updated their profiles using the interactive interface and received iterative feedback from the recommender based on a treatment (feedback or no feedback), and were subsequently asked to rate the post-manipulation set of recommendations from the system.

Design and Metrics

Participants in the dynamic feedback condition received recommendations generated by the system on the fly as they manipulated their profile in the second phase, while the remaining did not (see Table 4.1). By comparing ratings from the first phase against the second phase, and between treatments, we were able to examine how manipulation of profiles affected recommendation error, satisfaction, trust, and perceived recommendation accuracy in the presence and absence of dynamic feedback (Table 4.2). To use our earlier analogy, profile manipulations can be used to establish "what-if" scenarios at low cost to the user. Our aim is to assess how users go about this process, and what the resulting outcome is for their final recommendations and overall user experience.

The recommender system was deployed on Amazon Mechanical Turk (AMT) and data was collected from 129 AMT workers. Previous studies have established that the quality of data collected from AMT is comparable to what would be collected from supervised laboratory experiments, if studies are carefully set up, explained, and controlled [130, 131]. Previous studies of recommender systems have also sucessfully leveraged AMT as a subject pool [121]. We carefully follow recommended best practices in our AMT experimental design and procedures.

Generating Recommendations

Since the focus of this paper is on examining the profile manipulation behavior of users, and not specifically on the underlying recommendation algorithm, we chose a standard dataset (10m Movielens) and a standard collaborative filtering algorithm [132].

Algorithm

A collaborative filtering algorithm was chosen for this experiment because it lends itself well to visualization, but other algorithms should be interchangeable in the context of this experiment. Note that users have reported being able to easily understand visual representations of the algorithm [121]. The variant of collaborative filtering in this study applies Herlocker damping to increase recommendation quality.

User Interface

Our experiment uses a three column representation of collaborative filtering, similar to [121]. From left to right, a user sees his or her movie profile, then similar users in the collaborative filtering database, and finally a list of top movie recommendations. The underlying algorithm represents results from collaborative filtering as a directed graph, connecting the user's profile items to database users with at least one overlapping item and specifying edge strength as a similarity score. This score is shown as a light-blue gauge on the node for simplicity. Thus, if a user clicks on a movie he has rated, they can see which other similar users have rated it, and which recommendations are a result of those ratings. The recommendation column uses a star notation rather than a bar, provides visuals from the movie in the form of a teaser poster, and, when clicked, takes the user to RottenTomatoes.com to get more information about the movie.

4.4 Experiment Results

More than 300 users started the study, but many users were unable to complete the task properly due to the scarcity of valid "stale" profiles. Participant age ranged from 18 to 65, with an average of 31 and a median of 29. 53% of participants were male while 47% were female. Since we are interested in profile manipulation behavior, our

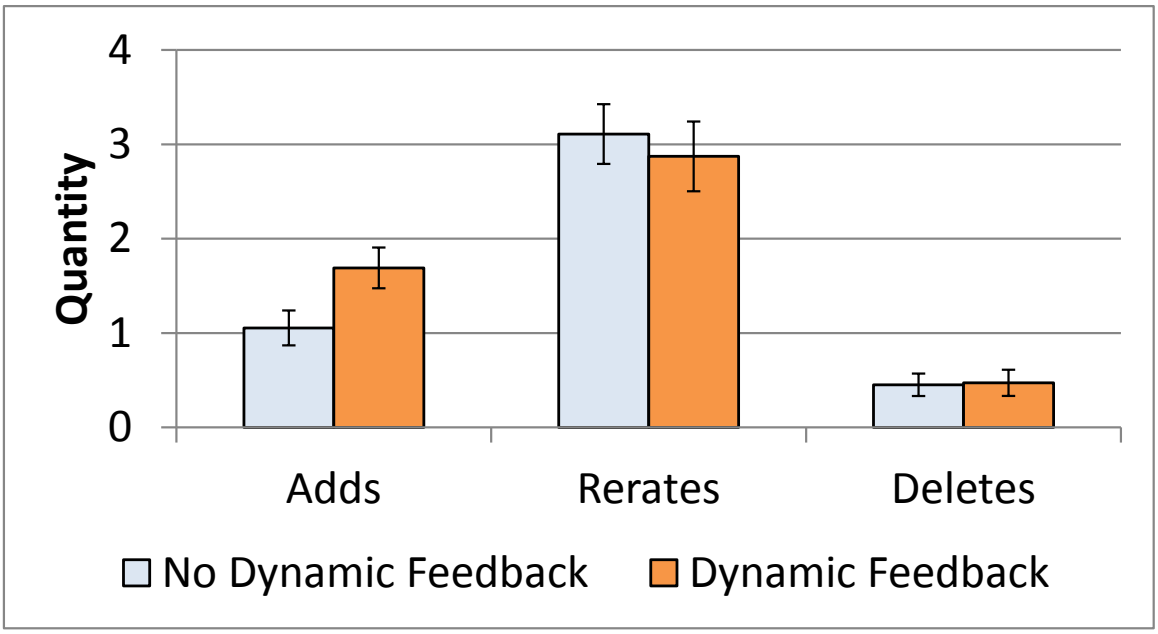


Figure 4.2: Frequency of each type of manipulation for each treatment. Error bars show one standard error below and above the mean. The most common action was re-rating an old item, and deletion of an old item was much more rare in comparison.

experimental design did not enforce any minimum number of manipulations. After the initial profile collection phase, many users did not make enough updates to their profile, so could not be used in our analysis. Furthermore, some users indicated that their profile no longer required updates to accurately reflect their preferences, therefore implicitly indicating that the data was not truly 'stale' when the task was started. After removing these participants, data from 129 users (73 for the no feedback treatment, 55 for the feedback treatment) was analyzed. The average rating over initial recommendations for these users was 3.88 (out of 5) while the average rating for final recommendations was 3.93.

Effect of Dynamic Feedback on Profile Updates

After the user's profile was gathered, we allowed them to make an arbitrary of manipulations to update their profile and get hypothetical recommendations. A breakdown of

the manipulation behavior, by treatment is shown in Figure 4.2. Re-rating a previously added item was the most common behavior in both conditions, followed by addition of a new item and deletion of an item respectively. Between both treatments, participants were 2.18x more likely to re-rate than add ($p < 0.01$), and 2.97x more likely to add than to delete ($p = 0.01$). Participants in the dynamic feedback treatment were also 1.6x more likely to add items than participants in the no feedback treatment with low presumption ($p = 0.108$).

Effect of Updates on Recommendation Error

Since we are interested in understanding the impact of each type of profile update on recommendation error, the error was measured with both the initial and final profiles so the two could be compared. Here, recommendation error is defined as mean absolute error ($MAE(p)$) for each participant p, or the difference between a participant's rating for an item and the predicted rating for that item:

$$MAE(p) = \frac{1}{n} \sum_{i=1}^{n} |p_i - r_i| \tag{4.1}$$

Where n is the total number of movies rated by participant p, p_i is the rating given by participant p to movie i, and r_i is the rating the system predicted participant p would give to movie i. Now we can define an error shift between the initial and final profiles of participant p by looking at the recommendations for each:

$$\delta error_p = MAE_{final}(p) - MAE_{initial}(p) \tag{4.2}$$

We realized one difficulty with our methodological approach is that users who initially received high quality recommendations are likely to exhibit different manipulation behavior from those with poor quality initial recommendation. Accordingly, we hypoth-

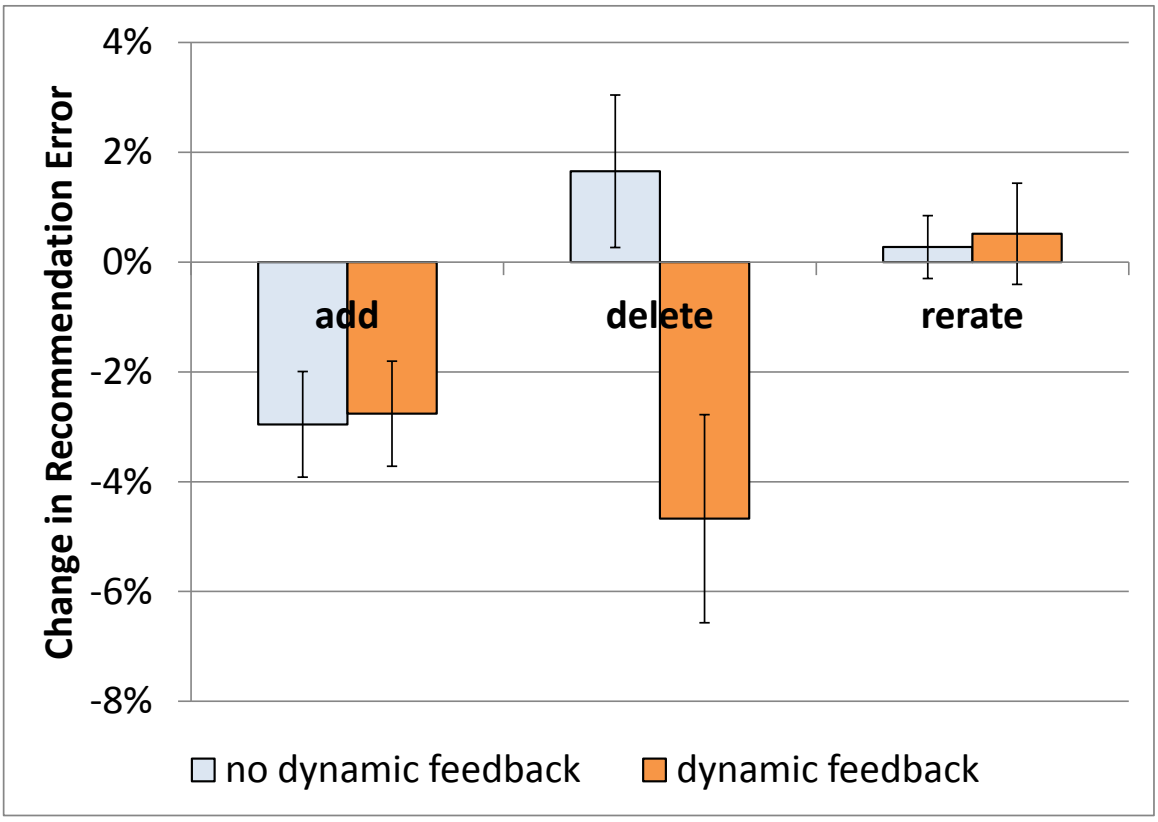


Figure 4.3: This graph shows the change in recommendation error that occurs from each type of manipulation in each treatment. Error bars are one standard error above and below the mean. These values were found by fitting a linear model to the manipulation patterns of participants that initially received poor recommendations. Error bars show one standard error below and above the mean. Adding new items was productive in both treatments, while deleting items was productive only in the dynamic feedback treatment.

esized that initial recommendation error and the resulting shift in error would have a significant interaction effect. We compensated for this by performing an analysis of the error shift based on the initial recommendation error. A linear regression showed that error shift was highly dependent on the initial recommendation error ($p < 0.01$). When the data is split on the average initial recommendation error (0.214), we find that users below the mean saw a 6.73% decrease in recommendation error after manipulation, while users above the mean saw a 1.14% increase ($p < 0.01$). In other words, users that had good initial recommendations could not do much to improve them, and in some cases

manipulations caused increase in error, despite the fact that dynamic feedback was given during this process.

Given the above, we fit the following linear regression models to each treatment of users that saw initial recommendations with an error below the mean (no feedback: N=27, feedback: N=21):

$$\delta error(p) = adds(p) + rerates(p) + deletes(p) \tag{4.3}$$

Where $adds(p), rerates(p), deletes(p)$ return the quantity of those manipulations for participant p. The coefficients of the model indicate the impact of each type of manipulation had on recommendation accuracy for participants that had below average initial recommendations. We fit the regression model to both treatment groups and the resulting model coefficients are shown in Figure 4.3. Note that the model for the dynamic feedback group was accurately able to explain variability in the data set ($p = 0.016, R^2 = 0.45$) vs. the model for the group without dynamic feedback ($p = 0.68, R^2 = 0.062$). The resulting models show that profile additions are the most effective manipulation for both treatments in terms of recommendation error, but deletes in the dynamic feedback group were the most effective manipulation overall. Deletes in the no-feedback treatment as well as re-rates in either treatment were either not effective or somewhat harmful to recommendation accuracy.

Effect of Dynamic Feedback on Perception

As stated before, perceptual metrics (overall satisfaction with recommendation, overall trust in the recommender, and perceived accuracy of recommendations) were taken after the final profile manipulation phase during the post-study test. Figure 4.4 and Figure 4.5 show a breakdown of the initial and final reports from each participant for these

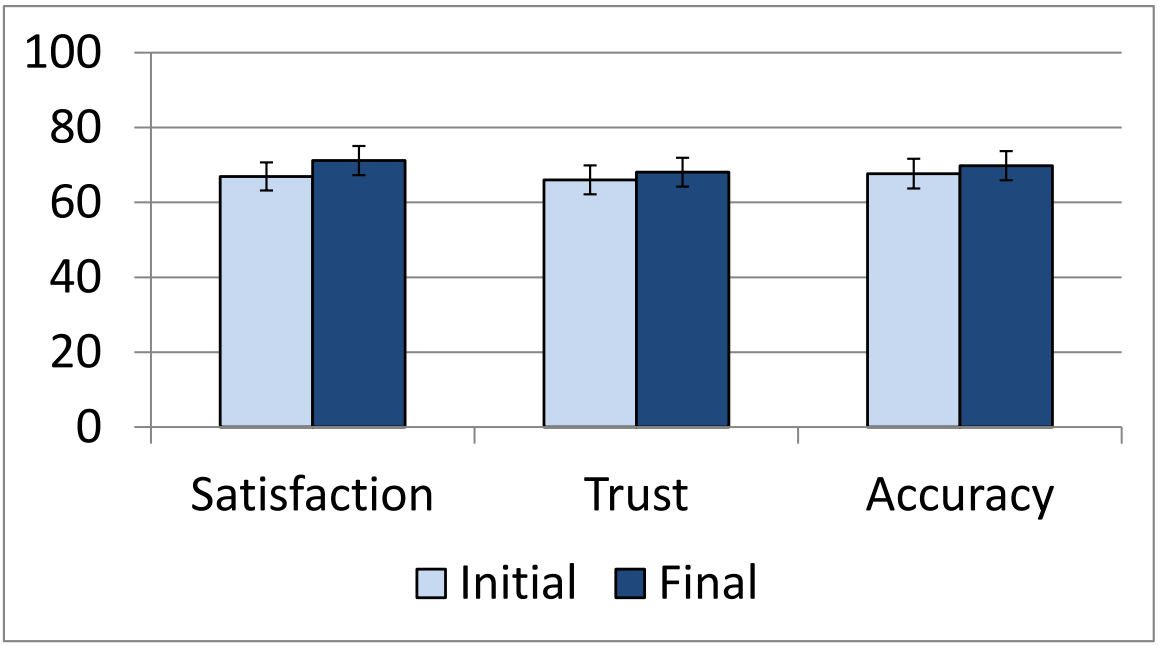


Figure 4.4: This graph shows participant responses to questions about satisfaction with recommendations, overall trust in the recommender, and perceived accuracy of recommendations for the no feedback treatment. Error bars show one standard error below and above the mean.

questions. We found that perceived trust and accuracy significantly increased for the dynamic feedback condition; this was verified by repeated-measures ANOVAs ($p = 0.0095$, and $p = 0.0158$ respectively). No significant change was found when dynamic feedback was not present. However, a mixed-measures ANOVA comparing the two treatments showed that the presence of dynamic feedback did not account for most of the change, as the before-and-after effect was more significant. Keep in mind this also applies to all participants in the dynamic feedback treatment, not just the ones that received poor initial recommendations, and that, overall, actual recommendation accuracy did not change for these participants.

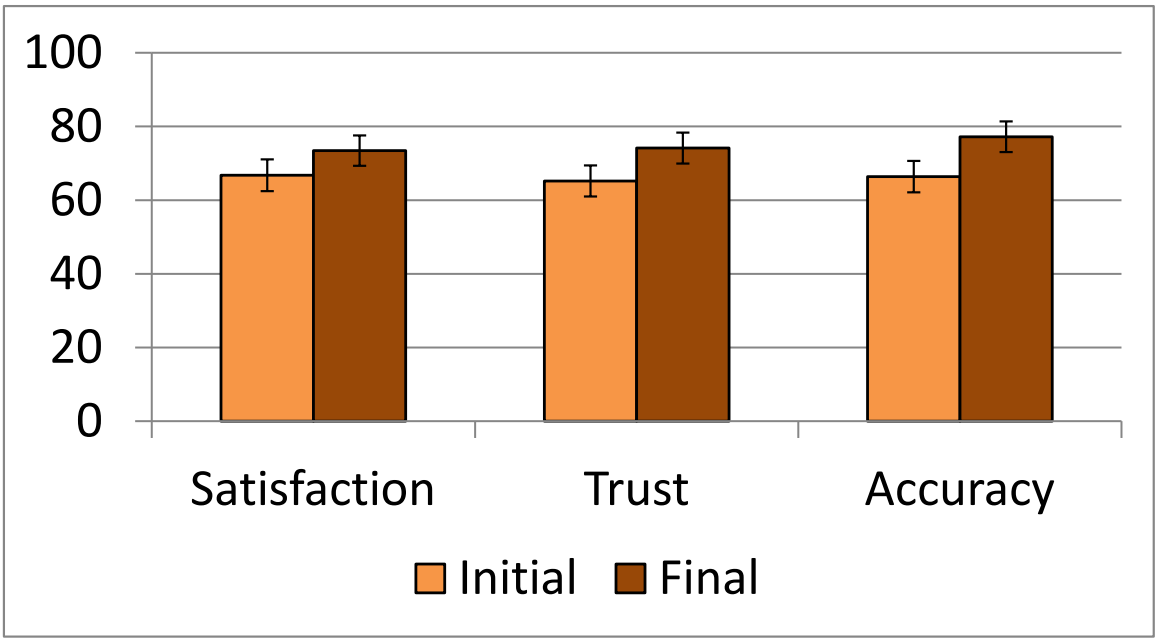


Figure 4.5: This graph shows participant responses to questions about satisfaction with recommendations, overall trust in the recommender, and perceived accuracy of recommendations for the dynamic feedback treatment. Error bars show one standard error below and above the mean.

4.5 Analysis and Discussion

Here we discuss the potential broader impacts of the results, limitations of the experiment, and plan for a follow-up study.

Re-rates are the most frequent type of profile manipulation, but have the least impact. To explain this effect, we posited that users did not change their rating very much from the initial to final profile. To verify this, we obtained the average difference between the original rating and any arbitrary re-rate, and found that most re-rates are within 1 point of the original rating on the 5 point scale provided. This supports the idea that, for movies at least, user tastes do not change very much over time. Visual recommenders in similar settings might consider the option of foregoing the functionality to re-rate, or perhaps put more emphasis on adding items or removing them altogether.

Dynamic feedback and visual explanation let users identify sources of bad recommendation and remove them. In our dynamic feedback treatment, delete actions improved recommendation accuracy by more than 4% on average for each individual delete performed. The most likely explanation is that users identified bad recommendations and were able to use the interface to determine and remove the item causing the correlation.

Users overvalue their profile updates. When users updated their profiles, they perceived that the overall accuracy of the recommendations and trust in the system increased significantly, even though the actual recommendation error stayed the same on average. While our mixed-measures ANOVA showed no significant effect of dynamic feedback on perception, participants in the no feedback condition reported on average that there was more or less no difference in accuracy and trust from the initial and final profiles. Thus, we still recommend that if a service requires a user to perform a profile update task (such as the first time a user accesses the system after a long period), dynamic feedback should be utilized.

4.6 Limitations

The authors note several points where this study could have been improved. First, there was difficulty in acquiring truly stale profiles from users and obtaining quality manipulations. As stated before, many users that started the study copied their profile into our system and then skipped the manipulation phase of the study. We considered enforcing that participants make some threshold number of manipulations, but any such enforcement would prevent our measurement of true profile manipulation behavior. A second limitation was that we only used a single recommendation strategy (collaborative filtering) in this experiment. It is not clear whether the findings about the manipulation

types would apply to other recommendation algorithms or to different data domains. Additionally, the choice of MovieLens 10M was for ease of implementation, and we note that many participants requested more up-to-date movie recommendations. Finally, our rating system could have been improved by considering list-based satisfaction, since disjoint ratings do not fully capture user satisfaction.

4.7 Summary

At the beginning of this chapter we stated that 1) feedback explanations encourage users to interact with a recommendation algorithm (Figure 4.2), 2) increased explanation can aid users when choosing control parameters (Figure 4.3), and 3) when feedback was present, users reported increased trust and accuracy in recommendations after a profile update task.

Chapter 5

Complex Algorithms and Human Beliefs about Data: Microblog Traffic Analysis

In this chapter, we'll show increasing levels of explanation from a complicated pattern detection algorithm affected a user's cognitive load, task effectiveness, and quantitative data understanding during a search and discovery task in traffic data. Users interacted with two computational models (one search/filter interface, and the "anomaly detector") through a single user interface. The results from this experiment support the following claims:

- Increased explanation facilities can negatively impact a user's cognitive load, specifically decreased confidence and decreased enjoyability.
- Decision support systems can quickly boost data understanding in foreign domains, but only if users understand the limitations of automated analyses

This work was previously published at ACM Intelligent User Interfaces in 2015 [133].

5.1 Introduction to Exploratory Search Sessions

In adaptive information systems, users are typically kept at a distance from the the underlying mechanisms used to generate personalized content or predictions, for example, personalized product recommendations (e.g. from Amazon) or movie recommendations (e.g. from Netflix). Google's search result lists are another popular example of dynamically adapted content, based on the search history of a target user. In this work, we use the term 'recommender' to refer to complex prediction algorithms, data mining algorithms, intelligent systems, or any other algorithms which produce ranked lists of 'interesting' data items, but are complex enough that they would not commonly have their mechanisms explained to the user. Note that recommendations from these systems are often presented through an otherwise static interface such as a list of search results or a in grid. Moreover, traditional browsing mechanisms such as text search, data overviews, and sorting mechanisms are often presented separately from, and operate independently of, recommendations (such as Amazon's product catalog).

During an exploratory search session with an interface, users perform iterated cycles of exploration, hypothesis, and discovery - a process often employed in scientific research, statistical analysis, and even catalog browsing. Users may start with very vague parameters, for example: "What are the most interesting movies in this genre?"; "What interesting things do Twitter users say about this topic?". Exploration can yield hypotheses that can then be answered with targeted search, for example, "is this Amazon product cheaper from another seller? Perhaps yes?"; "is this Twitter user familiar with this topic? Probably not."; "is there a higher-rated Netflix movie that is similar to this one? There has to be!". Each answer that the user finds may create new questions, prompt additional exploration, or cause the user to change his or her search strategy. Recommenders can be extremely valuable during these iterated cycles of exploration, but there

is not yet a complete understanding of the interaction between recommender and user search strategies. Recent research in conversational or critiqued recommender systems [134, 135] go some way towards adapting to rapidly changing user needs. Research on explanation interfaces [136, 137, 138] shows that explanations can bias the user towards system predictions, but can also help the user understand why the system is predicting particular content, resulting in a better experience with the system and increased trust in its predictions. However, recommendation algorithms are not perfect as they struggle with noisy and unreliable user-provided content. We believe that providing more transparent and interactive recommenders to users performing exploratory search tasks can result in better consolidation of search strategies and thus improved performance.

Our research results that indicate that varying the degree of available information *about the underlying recommender* significantly affects the trade-offs between 1) information discovery (amount of interesting/useful information found), 2) general insight into the underlying data set, and 3) user experience with the system. By understanding these trade-offs, better interfaces that show the right content to the right user at the right time can be developed. We describe a user experiment ($N = 197$) designed to provide insight on three general research questions: 1) how can an interface be adapted to consolidate user and recommender search and exploration strategies? 2) how do recommendation algorithms change user perception of an underlying data set? 3) what are the positive and negative effects of explaining recommendation algorithms in this context?

As an example task for our experiment, we chose analysis of commuter traffic reports on Twitter in the San Francisco Bay area. This application scenario was chosen because of the large amount of potentially noisy user-provided content (Twitter postings) and associated metadata. The volume of data (22,580 messages) was large enough to make visual scanning of the messages inefficient, necessitating use of traditional search and recommendation functionality within our tool. An automated anomaly detector and

recommendation algorithm [139], from here on referred to as the *anomaly recommender*, was used to generate recommendations of anomalous messages in the data set. This system serves as an example of a prediction algorithm - the results from this experiment should reasonably apply to similar systems. As previously motivated above, the study design focuses on open-ended data exploration. Users were simply tasked with finding a broad set of interesting, potentially useful information related to traffic blockages. In agreement with [6], we believe that this methodology reduces evaluation biases that occur when users are assigned very specific search parameters, and is a good representation of many real world exploratory search tasks.

5.2 Related Work in Traffic Analysis

Daly et al. [140] also study the domain of commuter traffic, with a view to increasing user understanding of a large corpus of real time Twitter messages. The approach in this case contrasts with our research in that they evaluate a system using a novel combination of Linked Data and Twitter messages to inform users about anomalies, rather than studying how interfaces might best support explanation facilities in this context.

5.3 Approach

This section describes the the anomaly recommender, Clarisense, and the interactive interface, Fluo, in more detail. The Twitter tweets and related metadata shown to participants were collected between July 12, 2014 and August 24, 2014. Tweets were filtered by looking at the keyword 'traffic' near San Francisco, California, USA. These tweets were then fed to the content recommender for summary, and a provenance view of this operation was shown in the Fluo interface. In most treatments, participants were

also given the original unfiltered dataset alongside the Clarisense recommendation.

In our experiment, we compromise between a truly open-ended protocol and a benchmark protocol by giving users a set of high-level parameters, which are qualified as 'interesting,' and allowing them to explore the dataset in any way they choose. Additionally, we measure performance by comparing participant discoveries against a benchmark that was created post-hoc by examining all discoveries made by participants. We believe this methodology sidesteps the difficulties of longitudinal studies while still representing a realistic task.

5.3.1 Clarisense Architecture

Clarisense is a Twitter-targeted automated anomaly detection algorithm developed at UIUC. To detect relevant information in the dataset, Clarisense employed a search strategy of examining the frequency of topics over time in the Twitter microblog. Once provided with the complete set of raw data, the input goes through a series of intermediate steps to before a final report is generated. The first step in this pipeline is the division of the input data into 24 hour period time chunks followed by clustering within each chunk in order to retain only unique tweets and remove any redundant information if present. The 24 hour period parameter was determined based on the percentage of delay between the retweets and the original tweet that was observed, corresponding to a value of about 70%. In each cluster, only one tweet is chosen as relevant and passed to the next step in the pipeline. At this point, the goal is to find the events within each cluster that stands out from the normal ongoing events during that period of time. We opted for an approach which uses keywords from each tweet as the primary purpose of identification. A pair of keywords, rather than single keywords or n-tuples, showed the highest correspondence between independent events and their keyword signatures.

Then, information gain, expectation maximization, and spatial analysis were utilized in order to find the discriminative keyword pairs in a current window as opposed to previous windows. For the final step in the pipeline, we just look at the top discriminative keyword pairs in each window and report all tweets that contain these pairs ranked by their spatial-credible information gain value. A full description of Clarisense can be found in [141].

5.3.2 Fluo

This section introduces our interactive interface (and experimental platform), Fluo, and briefly describes its methodology.

Design

Fluo (Figure 5.1) is a provenance visualization that was designed for exploring the top-N results from a ranking algorithm. It is part of ongoing work in the inspectability and control of recommenders and data mining algorithms at UCSB [121]. In the interface, data items or intermediate calculations are represented as nodes and organized into columns, which can be placed serially (creating an upstream/downstream relationship) or in parallel (to represent that multiple sources are weighted together). Each node in the visualization may have a corresponding "score" which is shown as a gauge and can be mapped to any corresponding value in the underlying algorithm (e.g., Pearson Correlation for collaborative filtering). A mapping from interaction techniques to commonly recognized user intents [142] is shown in Table 5.1.

Since Fluo groups nodes into lists which only contain items belonging to one semantic category, reconfiguration for each experimental treatment becomes simplified. A breakdown of which metadata is shown in which condition is provided in Table 5.2. Moreover, interaction techniques and visualizations of each type of item remain consistent across

User Intent	User Action and Response
Select	User selects an item in a list, the system highlights the item and keeps it at the top its list.
Explore	User scrolls a list, the system shows new items along a fixed parameter (time, frequency, relevance to search term)
Reconfigure	For the purpose of evaluation, the types of items represented and the sort parameters were fixed by the experimenters ahead of time.
Encode	For the purpose of evaluation, the color, size, and shape of items was fixed by the experimenters ahead of time.
Abstract /Elaborate	The user mouses over an item in a list, the system provides additional details about the item in a panel.
Filter	The user selects a time bin, only items from that time are shown. The user enters a search term, only items matching that term are shown.
Connect	The user selects an item, connected items (friends) are brought to the top of their respective list. The user can then expand the selection to show even more connected items (friends of friends).

Table 5.1: User intents supported by the interactive interface for this experiment.

configurations allowing for easier interpretation of results.

Explanation of Clarisense

Clarisense's search strategy was simplified for users and presented through the interface, as shown in Figure 5.1. For users in the full explanation condition, intermediate steps of the algorithm and their values (time chunks and topics or keyword pairs) are exposed as provenance metadata. Users can inspect this metadata like any other list, revealing relationships between the original data set, the extracted keywords, and their frequency on vary time chunks.

Explanation Level	**Treatment**	**Description**
Baseline	Tweet Metadata Only (Figure 1, A)	Twitter metadata (source, tweet, hashtags, time) shown. Text search over message body content, filter by time. Different selections of messages, sources, and hashtags unveil different relationships through edges on-demand.
Level 0	Clarisense Only (Figure 1, H)	Clarisense's summarized reports with text search, filter by time. The 'what' of Clarisense's reports are summarized but not the 'how' (no provenance). Twitter metadata and messages are not present.
Level 1	Clarisense in Context (Figure 1, A+H)	A combination of the two previous conditions. Additionally, users can see the relationship between the original tweets and the reports, making this a partial provenance view.
Level 2	Clarisense in Context w/ Explanation (Figure 1, A+B)	Similar to the previous condition, but Clarisense's selected time intervals and topic modeling were exposed to the users, making this tool a full provenance view of Clarisense's anomaly calculation.

Table 5.2: Description of Experimental Conditions.

5.4 Experiment Setup

In this experiment, we wanted to examine how varying levels of explanation from the recommender affect the entire human-recommender system's ability to 1) find relevant, interesting data items and 2) generate an overall understanding or accurate perception of the data, especially when data items are too large to be browsed sequentially. We also measured the effects of various levels of explanation on the user's confidence, perception of the tool, and enjoyment of the task. Four experimental conditions were tested, as described in Table 5.2 (See also Figure 5.1).

5.4.1 Metrics

In this experiment, users collected 'interesting' evidence (Tweets or anomaly reports from Clarisense) from the Twitter traffic data set, based on a loose criteria that was given to them. Afterwards, they made estimations of the dataset and reported their perceptions of the tool in a questionnaire. An overview of the metrics that were utilized in this study are shown in Table 5.3.

Event Recall

Once all participant data was collected, an analysis of evidence yielded a list of benchmark discoveries shown in Table 5.4. Each event $e \in E$ was chosen based on the frequency that participants reported it, as well as accuracy to which the event can be described by the data set. Non-descriptive tweets that mention traffic but do not mention at least the what or the where were not included in the final benchmark, nor were events that were reported by fewer than 3 participants. During the task, we allowed a participant to claim a discovery if they included at least one message in their evidence, $v \in V_p$, from that event's ground truth in our post-hoc benchmark, $g \in G_e$, with the rest of the submitted evidence being classified as noise $n \in N_p$, $N_p \subseteq V_p$. Recall is simply defined as the total number of events detected by a participant:

$$recall = \sum_{e \in E} \begin{cases} 1 & : |V_p \cap G_e| > 0 \\ 0 & : |V_p \cap G_e| = 0 \end{cases} \tag{5.1}$$

For recall, there was one exception in our benchmark. Due to the large size of the 'Web Traffic' event and because participants were not told that this event is interesting during the start evidence collection, we assessed if participants made this discovery during the following questionnaire. Note that recall is **not** normalized, and can fall between 0

and 22.

We also considered precision, which in this case measures the amount of noise the participant reported N_p:

$$precision = \frac{|V_p| - |N_p|}{|V_p|} \tag{5.2}$$

Precision allows us to understand the quality of evidence the participant submitted, and was useful for detecting outliers.

Quantitative Understanding

After the evidence collection task ended, we requested that the participant estimate of the number of blockages that were actually represented in the data set which pertained to a particular type of incident. Disabled vehicles, damaged infrastructure, police/riot/protest, and planned public events were chosen for these questions due to practical limitations on generating ground truth for events that are likely to have two instances occur simultaneously in time (construction, traffic accidents). Participants entered their answer in plain text boxes. These metrics gave us the participant's qualitative judgment the impact of each type of incident on traffic.

Usability

Participant perceptions of the tool were collected after the evidence collection and estimation tasks. Participants provided answers on a Likert scale (1-7) for each question on a web form. Participants were asked "How confident were you using the tool to complete the task?", "How much did you like the interface tool?", "How much did you like the training portion of the task?", and "How much did you enjoy the evidence collection portion of the task?"

Metric	Description
Event Recall	Total number of events the participant discovered through evidence submission during the fixed-time phase of the task.
Quantitative Understanding	Error in estimation when guessing the quantity of events related to specific types of blockages (disabled vehicles, damaged infrastructure, police/riot/protest,planned public events.
Usability	Participant's confidence, enjoyment, and perceptions of the tool, taken on a Likert scale in the post-study.

Table 5.3: Description of dependent variables.

Hypothesis

Evaluating features of each treatment separately, then in combination enables systematic assessment of the value added by each feature as well as allows for the identification of synergistic value gained by the combinations. The following hypotheses were evaluated during this experiment:

- H_0: interface type does not impact event recall
- H_1: interface type impacts event recall
- H_0: interface type does not affect overall insight and understanding
- H_1: interface type affects overall insight and understanding
- H_0: interface type does not affect usability
- H_1: interface type affects usability

Event	Description	Score	Size
gas leak	On 7/11, a gas leak closed 7th St and Broadway	0	4
drake/eliseo light	On 7/22, a traffic signal broke on Eliseo Dr	0.21	4
08/12 oil spill	Oil from a truck was spilled (San Mateo bridge)	0.82	10
08/21 oil spill	Oil from a truck was spilled on Magdalena Ave	0.34	6
cesar light	On 08/18, a traffic light malfunctioned on Cesar Chavez	0.29	4
quake	On 08/24, a major quake damaged multiple roads in Napa, Vallejo, and Sonoma	0.13	17
andy lopez	On 07/12, A protest demanding justice for Andy Lopez blocked highway 101	0.26	6
07/20 market st protest	A protest caused severe congestion on Market St	0	1
07/26 market st protest	A protest caused severe congestion on Market St	0.10	2
ferguson	On, 08/22, a protest of the Ferguson shooting caused traffic to stop near Civic Center Plaza	0.05	3
coliseum	On 07/25, a bomb threat at Coliseum Station caused highway 880 to become blocked	0.46	10
lombard	On 07/12 and 07/19, Lombard St was closed to the public by city officials	0.14	3
49ers	On 08/03, a 49ers game at Levi Stadium caused severe congestion	1.00	30
obama	On 07/23, an Obama visit caused multiple blockages/road closures near downtown	0.48	33
mccartney	On 08/14, a Paul McCartney concert caused severe congestion near Candlestick theater	0.48	30
soccer	On 07/26, a soccer game at UC Berkeley caused severe congestion	0.29	9
marathon	On 07/26, the San Francisco Marathon resulted in multiple road closures	0.18	13
japan	On 07/19, a J-Pop Festival in Japantown resulted in road closures and severe congestion	0.65	12
terminator	On 08/03, the Golden Gate Bridge was closed for the filming of Terminator 5	0.08	2
st francis constr	On 07/16, part of St Francis Dr was closed all day to traffic	0.03	2
slurry seal	on 08/07 and 08/08, construction caused delays and closures near Ralston Ave	0.08	6
web traffic	A significant percentage of the messages in the dataset related to web traffic	0.08	-

Table 5.4: The post-hoc benchmark - events discovered by participants during the task.

5.5 Experiment Protocol

For convenience, an overview of the terms used in this experiment are shown in Table 5.5. A list of events that were found by the participants is shown in Table 5.4. In this table, 'Score' indicates Clarisense's recommendation for tweets associated with this event. 'Size' indicates the total number of distinct Tweets that identified the what, where, and when of the event.

The experimental toolkit was deployed as a web service and the link was made available on Amazon Mechanical Turk (AMT). The AMT web service is attractive for researchers who require large participant pools and low cost overhead for their experiments. However, there is valid concern that data collected online may be of low quality and require robust methods of validation. Numerous experiments have been conducted, notably [130] and [131], that have attempted to show the validity of using the service for the collection of data intended for academic and applied research. These studies have generally found that the quality of data collected from AMT is comparable to what would be collected from supervised laboratory experiments, if studies are carefully set up, explained, and controlled. Previous studies of recommender systems have also successfully leveraged AMT as a subject pool [121, 128]. We carefully follow recommended best practices in our AMT experimental design and procedures.

5.5.1 Overview

After accessing the experimental system through AMT, participants were presented with a pre-study questionnaire using the Qualtrics survey tool [1], collecting basic demographic and background information. Next, they were directed to one of the variations of our online tool, instructed about its operation, and introduced to Clarisense through ex-

[1] www.qualtrics.com

ample queries and questions designed to test understanding. Once training was complete, the open-ended search task was described and participants were informed they would be limited to fifteen minutes for the next phase of the task. Users then used the interface to explore the tweets and Clarisense's reports into a list of evidence that they believed to be relevant to the high-level search terms. Once time was up, the interface was removed and we asked them several questions related to key quantities in the data set that were related to the exploration they performed. The training and evidence collection protocols are talked about in more detail below.

5.5.2 Training

Since the experimenters could not verbally direct the participants, a complex training module was created which walked the participant through key concepts before the evidence collection portion of the task. The participant was required to answer a series of targeted search questions, the answers to which could only be known once the participant identified which parts of the interface were providing what information. An unlimited number of attempts were given for each question. Easier questions were chosen as multiple choice with fewer than 4 options, while the hardest questions had blank response forms that required the entry of quantities. After informal pilot testing in the lab, it was decided that hints for every question were needed to alleviate participant fatigue during this portion of the study. During data collection, few participants encountered difficulties completing the training in any condition (or simply were not vocal if they chose to drop out of the task).

5.5.3 Evidence Collection

Once training was completed, participants were prompted that the evidence collection phase was about to begin. They were told that the data set contained numerous traffic blockages and that we were interested in studying blockages related to construction, infrastructure damage, broken or disabled vehicles, police activity (riots, protest), and planned public events. Participants were actively told to ignore traffic accidents, and distinctions were made between planned public events such as sports games or concerts and other events like riots. The active prompt for the task is shown in Figure 5.1 (J). Participants were told to look for these events and collect evidence in a list (by dragging and dropping either Tweets or anomaly reports), and that they would be paid a bonus for finding more interesting evidence related to blockages.

5.6 Results and Analysis

This section presents the statistical analysis that was done on participant data. Participant age ranged from 18 to 65, with an average of 25 and a median of 27. 52% of participants were male while 48% were female. 197 data points were utilized for analysis after outlier detection.

Table 5.6 shows the precision and recall for participants across treatments on the post-hoc benchmark (Table 5.4). The relative proportion of values within precision and recall was roughly the same, indicating that an increase in recall corresponded to an increase in precision. This result indicates that participants that were more careful about what evidence they chose during the task were also more successful at finding more evidence that was relevant to the broad search parameters. To simplify reporting of results, this section focuses on recall rather than precision.

Figure 5.2 shows recall across the four conditions with standard error shown. The

vertical axis indicates how many events in our benchmark (Table 5.4) each participant detected. Note that we measure recall non-normalized as to best represent the magnitudes of the quantity of discoveries. A large increase in recall is seen between the Twitter Only' condition and the conditions where Clarisense was present. A slight drop in total discoveries seems to occur between the 'Clarisense Only' condition and the conditions with Clarisense AND the original Twitter data.

Based on our design, these results indicated that the recommender was indeed useful overall when exploring the dataset. Participants were grouped by whether the recommender was available(no=60,yes=137) and a single factor analysis of variance was run, showing a significant decrease of 60% when Clarisense was not present ($F = 92.87, p < 0.01$) This result is not too surprising given that it Clarisense presented several key discoveries to participants even without any interaction. However, when we compare the 'Clarisense Only' condition with the conditions that provided context and explanation facilities, we see a 25% drop in discovery rate ($F = 19.54, p < 0.01$). Participant records indicate that they drew their evidence from Clarisense with 50% likelihood in these conditions (remaining evidence came from the original Twitter data). We can only conclude that when the participants spent time employing their own search strategy on the original data, it detracted from the rate at which they considered and incorporated the recommender's discoveries.

To assess if participant search strategies using Fluo were able to reliably contribute to the overall discovery process, we then considered for our recall plot only the discoveries which we determined that Clarisense 'missed' or underrepresented due to its filtering and reporting mechanism (Figure 5.3). To qualitatively determine what Clarisense had 'missed', we decided that a lenient anomaly score threshold should be chosen that would give Clarisense a precision of at least that of the worst participant in the Twitter Metadata only condition (0.034). We settled on a score threshold of one standard deviation

above the mean (0.15), which corresponded to a precision of 0.035. Referring to Table 5.4, this means that Clarisense reported 12 events total (slightly more than half) from 339 pieces of evidence total, which results in a list of 10 events that Clarisense missed. In the 'Clarisense Only' condition, participants only appear to be half as likely to discover one of these events.

Based on our design, these results indicated that presenting the original Twitter data in the Fluo interface allowed the participants to develop search strategies that yielded different discoveries than Clarisense. To verify this, we again grouped the participants by whether they had the original Twitter data available (no=51,yes=146) and ran another single-factor ANOVA between conditions that contained the original Twitter data set and the 'Clarisense Only' condition, finding a 63% decrease ($F = 33.65, p < 0.01$) in underrepresented discoveries when only Clarisense was present. To reiterate, two of these events were not represented in the 'Clarisense Only' condition (gas leak, 07/26 market st protest), and the remaining 8 were classified as significantly less interesting by the recommender. Of particular note is the 'terminator' event, which saw remarkably higher probability of recall when both the Twitter data and Clarisense reports were present. Evidently, even our novice participants were able to develop search strategies that consistently contributed at least a few novel discoveries to the analysis process.

Figure 5.4 shows the overall estimation error from our insight questionnaire. The vertical axis shows the average difference between actual and estimated distinct blockages for each treatment. Since the scales of the ground truth were similar (disabled vehicles: 29, damage: 6, police/riot/protest: 9, planned public events: 12) we aggregated these results into one graph. A value of 0 indicates perfect accuracy. Participants were much more likely to overestimate than underestimate. A large increase in estimation accuracy can be seen between the condition where Clarisense was absent (Twitter Only) and the other three conditions. Another drop can be seen between the Clarisense explanation

condition and the conditions where less explanation is given.

The results indicate that recommender presence had a positive impact on overall understanding of the data. We ran a single factor analysis of variance between the 'Twitter Only' condition and the conditions with the recommender, showing an estimation error decrease of 60% ($F = 4.8, p = 0.030$). This comparison shows that the data set may have simply been too large to gain a good understanding in the limited time frame, but that the recommender was able to provide a better introduction in a shorter time. We also investigated the additional 56% drop in estimation error for 'Clarisense with Explanation' against the other two Clarisense conditions, finding it fell just short of the 0.05 significance level despite its notable effect size ($F = 2.99, p = 0.087$).

To further investigate the decrease in estimation error for the full explanation condition, we plotted the estimation parameters for each type of blockage individually (Figure 5.5). The most notable of these was a large difference in planned public events - there was a 31% decrease in estimation error in the 'Clarisense with Context' condition and a 75% decrease in estimation error in the 'Clarisense with Explanation' condition. For the latter, we ran another single-factor ANOVA and found ($F = 4.10, p = 0.046$).

The drop in estimation error in planned public events becomes much more meaningful when we consider Clarisense's search strategy for interesting anomalies in the data set. On Twitter, large public events usually have distinct key terms and usually hashtags associated with them that only appear in conjunction with the event. As such, Clarisense is much more likely to view these as anomalous than other types of events in the data set. From Table 5.2, it can be seen that planned public events dominate the top 5 most anomalous events from Clarisense's perspective. It seems as though providing the original Twitter data set seemed to help the participant understand the context in which Clarisense was working, and further explanation and exposure of the topics seemed to caused additional improvement in the participant's perception.

Finally, we assessed the impact of an increasingly complicated interface and explanation of Clarisense on the participant. Figure 5.6 shows the results from our post study questionnaire. The answers were provided on a Likert scale (1-7). From left to right, the questions were 'How confident were you using the tool to complete the task?' (confident), 'How much did you like the interface tool?' (like), 'How much did you enjoy the training portion of the task?' (training), 'How much did you enjoy the evidence collection portion of the task?' (task). Presence of the original Twitter data appeared to decrease both confidence and enjoyment of the task, with the largest drops (27% and 29%) being between the 'Clarisense with Explanation' and the 'Clarisense Only' conditions. Across treatments, the participant's fondness of the tool and enjoyment of the training session did not appear to vary much.

To verify the large drop in confidence between 'Clarisense Only' and 'Clarisense with Explanation', we ran two more single-factor ANOVAs yielding ($F = 15.28, p < 0.01$) for the 27% confidence drop and ($F = 7.074, p < 0.1$) for the 29% enjoyment drop during the task phase. Several single factor ANOVAs were run between treatments for likeability, but nothing below the widely accepted significance level was found. This was also true for participant enjoyment of training, which was surprising due to the varying length based on the treatment. In sum, these results indicate that explanation facilities can potentially drop both a user's confidence and make the process of discovery more stressful.

In conclusion of our statistical analysis, we reject all null hypotheses and conclude that the Fluo-Clarisense configuration had a significant impact on event recall, participant perception and understanding of the data, and usability of the tool.

5.7 Discussion

In light of the results from this experiment, we derive three recommendations for recommender systems and search tools, especially for interfaces that attempt to consolidate both. Limitations of the experiment and directions for future work are also discussed.

5.7.1 Key Takeaways

Recommend First, Search Second: Interfaces should highlight results from a recommender when a user begins the process of data exploration, but general search and exploration tools should always be available. The participants in our experiment benefited greatly from the recommender presence, consistently reporting better estimations of the content of the data over those that received no recommender. Participants that could search over the original data set were still apt to do so, and through their own search strategy made discoveries that the recommender missed at the same rate as those who didn't interact with the recommender at all. The recommendations themselves may also serve as catalysts for initial searches and strategies, which can greatly help new users, novices, or those working with new data sets.

Contextualize and Explain Recommendations: Both the introduction of the original Twitter data and more explanation facilities appeared to help participants understand and contextualize Clarisense's search strategy, which greatly decreased their estimation error with respect to the specific topic that Clarisense over-represents (public events). Explanation facilities should carefully explain the search strategy of a recommender to users when this is appropriate and put the recommendations in context to avoid these errors. Though not every recommendation system is the same, in domains where decisions are costly, perception biases can be disastrous. For example, analysts of epidemics might ask: what is the relative severity of illness x and hazard y at a specific location - which

problem should more resources be allocated to? In other examples, a storekeeper may want to avoid creating misconceptions about the variety of items that his store has available, or a library might want to emphasize the impression of diversity among its titles.

Recommend to New Users, Explain to Returning Users: In this experiment, full explanation of the recommender decreased user confidence and enjoyment of the search and discovery task. The presence of the daunting Twitter dataset also appeared to contribute. While most of the participants in this task could be classified as novices in the field of information analysis, they were also new to the tool and some were new to the concept of Twitter. By creating and maintaining models of users, different configurations of the recommender and search tool might be shown at different times. For instance, a digital shop could minimalize their storefront and only initially show recommendations until the user requests a targeted search. When regular customers are established, the store can begin explaining/contextualizing recommendations so that the user can synthesize the recommendations with their own search strategy, potentially finding new products.

5.7.2 Limitations

The evidence collection portion of the task was limited to fifteen minutes and all users were essentially novices with the Fluo interface. Given more time and more comprehensive training, it is possible that users would have reached a 'saturation point,' where all useful information from the recommender would have been exhausted and more discoveries from user-contributed search strategies would have emerged.

Experimental data was collected online on Amazon Mechanical Turk, which diversified our user base but made outlier detection tricky. In general, the quality of submission on AMT seemed to be comparable with those provided in comparative supervised settings,

however, most AMT workers expect tasks between 60 seconds and 5 minutes on average. Longer tasks may catch users off guard, fatiguing them and increasing tendency for satisficing. While we took detailed timing metrics for all interactions, for some time windows it is difficult to tell if a user is merely thinking or, e.g., went to use the restroom. This makes outlier removal difficult to justify in some cases. Additionally, if a participant suffers from a key misconception, we cannot correct or account for it. Fortunately, larger sample sizes and quicker uptake help mitigate some of this noise inherent in AMT experiments.

5.8 Summary

At the beginning of this chapter, we stated that 1) increased provenance facilities can negatively impact a user's cognitive load and 2) explaining how decision support systems operate increases the accuracy of the user's perception of a dataset. From the experiment, we saw that more explanation from the anomaly detector decreased confidence and task enjoyability (Figure 5.6). Participants in one of the treatments where Clarisense was present had significantly higher recall rates over the treatment that did not have Clarisense (Figure 5.2). Finally, more explanation resulted in improved perception, or insight into, the dataset (Figures 5.4, 5.5). From these figures we can also see that the mere presence of Clarisense significantly improved the participant's ability to estimate the dataset parameters.

[H]

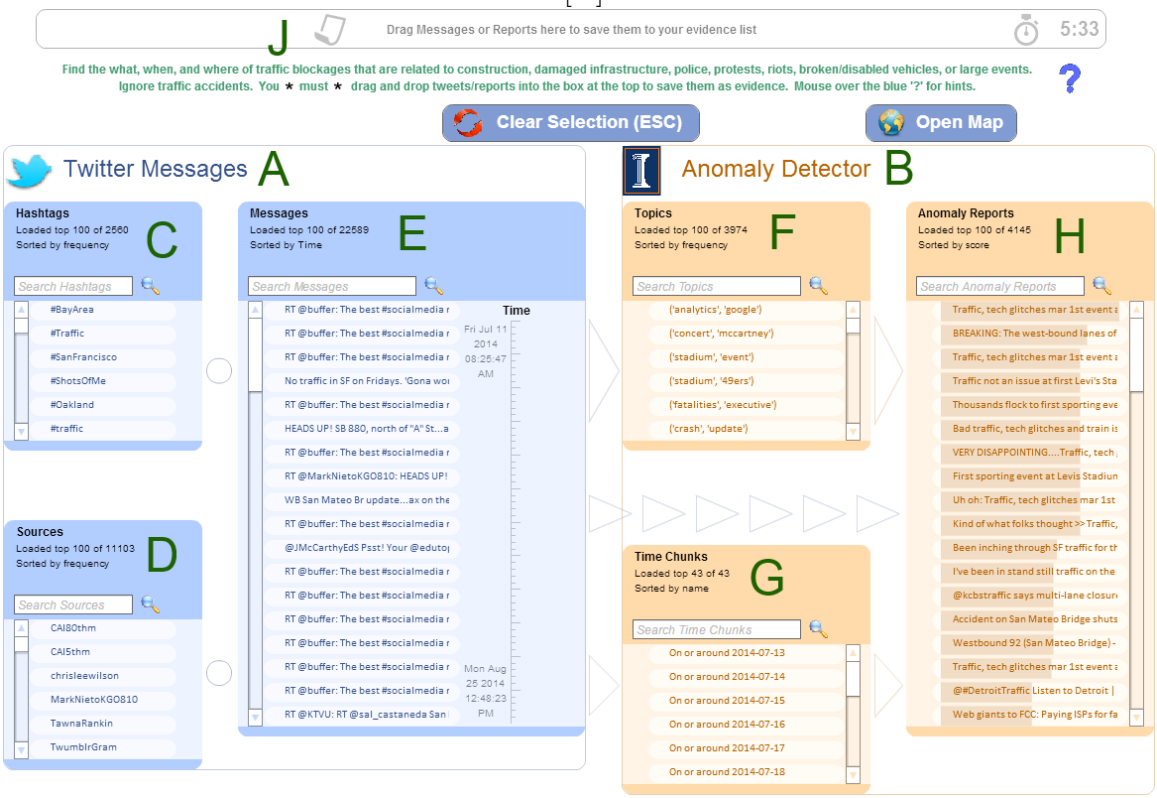


Figure 5.1: The provenance and data visualization tool, dubbed Fluo, showing A) the original dataset and B) provenance and results from Clarisense. Metadata from Twitter or Clarisense are organized into separate lists C) hashtags in the data, D) Twitter sources, E) the tweets themselves, sorted by time, F) topics that Clarisense utilized for anomaly detection, G) intervals of time that Clarisense utilized for anomaly detection, and H) the final anomaly reports, ranked by Clarisense's anomaly score. At the top of the interface (J), the evidence box and remaining task time is displayed. The participant's task prompt (in green) is also shown during the duration of evidence collection as a reminder.

Term	Definition
Fluo	The interactive interface used in this experiment. It was configured in one of four different ways to test the effects of recommender presence and explanation.
Clarisense	Also referred to as the anomaly recommender, an automated algorithm developed at UIUC to explain sensor anomalies using microblog data (in this case, Twitter feeds).
Event	A traffic blockage in the Twitter dataset which can be explained by major construction, broken vehicles (engine failures, fuel shortage), damaged infrastructure (broken roads and lights, police activity (riots, protest), or planned public events (sports games, concerts, festivals). Traffic accidents were considered uninteresting in this task due to their frequency of occurrence and noise level.
Evidence	Either a Tweet from the original dataset or a summarized group of Tweets from the anomaly recommender (see: Anomaly Report)
Anomaly Report	A group of Tweets summarized by the anomaly recommender and scored by its level of anomaly.
Discovery	If a participant's evidence list contained a Tweet that indicated at least the when and where or the when and what, we decided that the participant had discovered the event.
Insight	A participant's ability to estimate the total number of blockages that could be explained in the dataset, for a particular type of event.

Table 5.5: Definition of major terms in the experiment

Condition	Precision	Recall
Twitter Only	0.14	2.77
Clarisense Only	0.60	8.23
Clarisense with Context	0.44	6.44
Clarisense with Explanation	0.50	5.85

Table 5.6: Mean precision and recall for each interface configuration

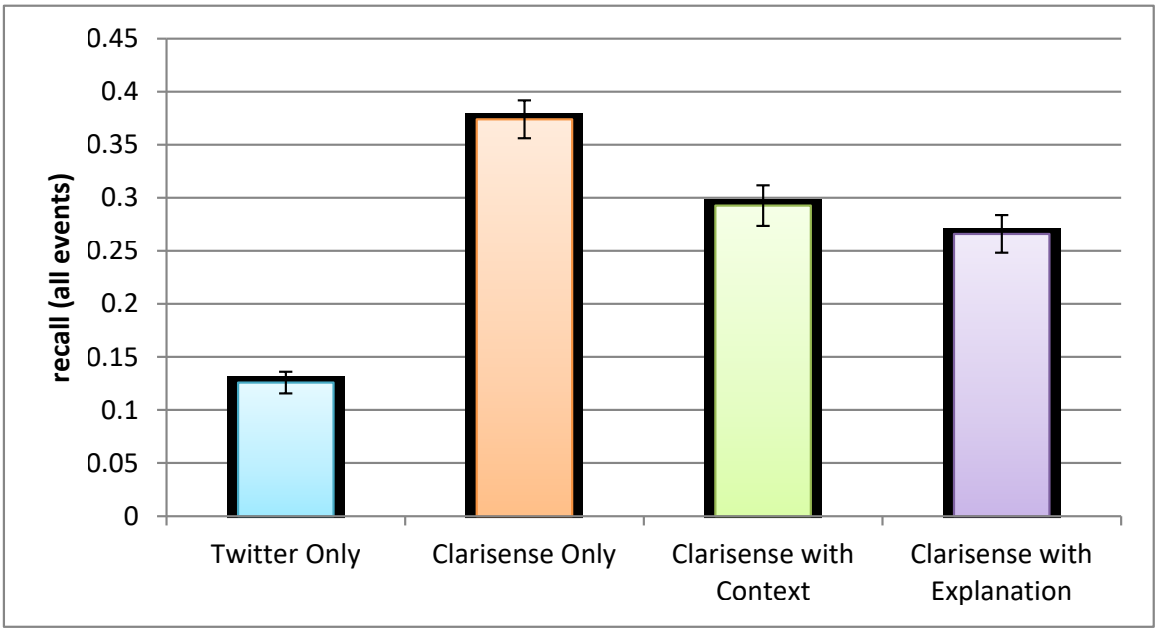


Figure 5.2: Participants that only interacted with the anomaly recommender were able to incorporate more of its discoveries in the same time period. Error bars show one standard error below and above the mean.

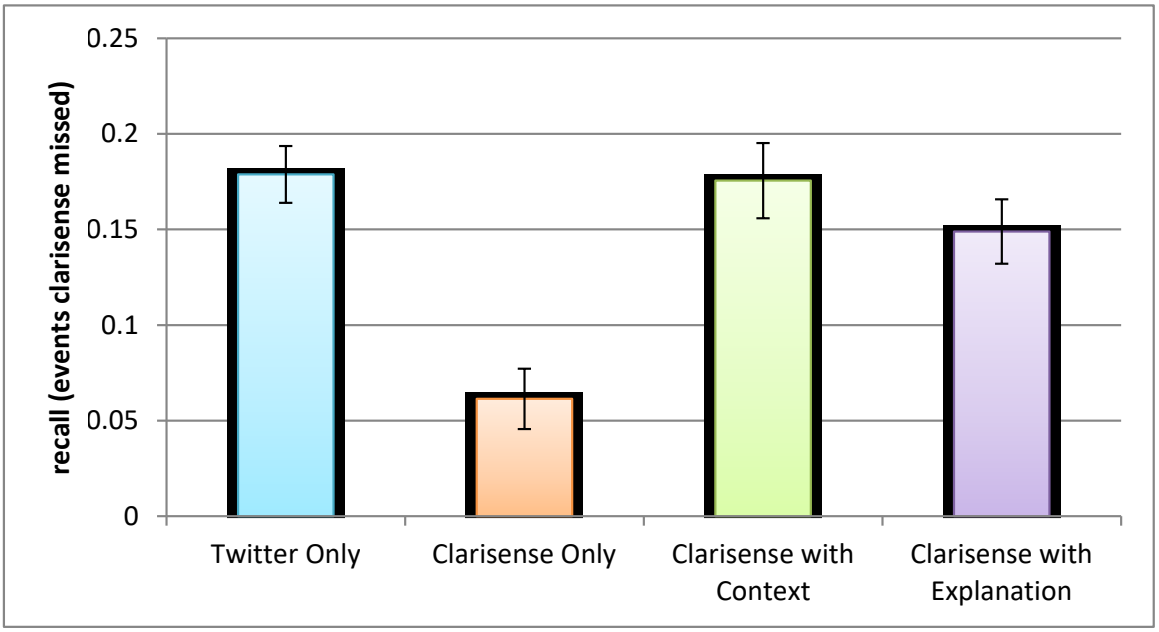


Figure 5.3: Participants that interacted with the original data were able to consistently find discoveries that the anomaly recommender missed or classified as relatively less interesting. Error bars show one standard error below and above the mean.

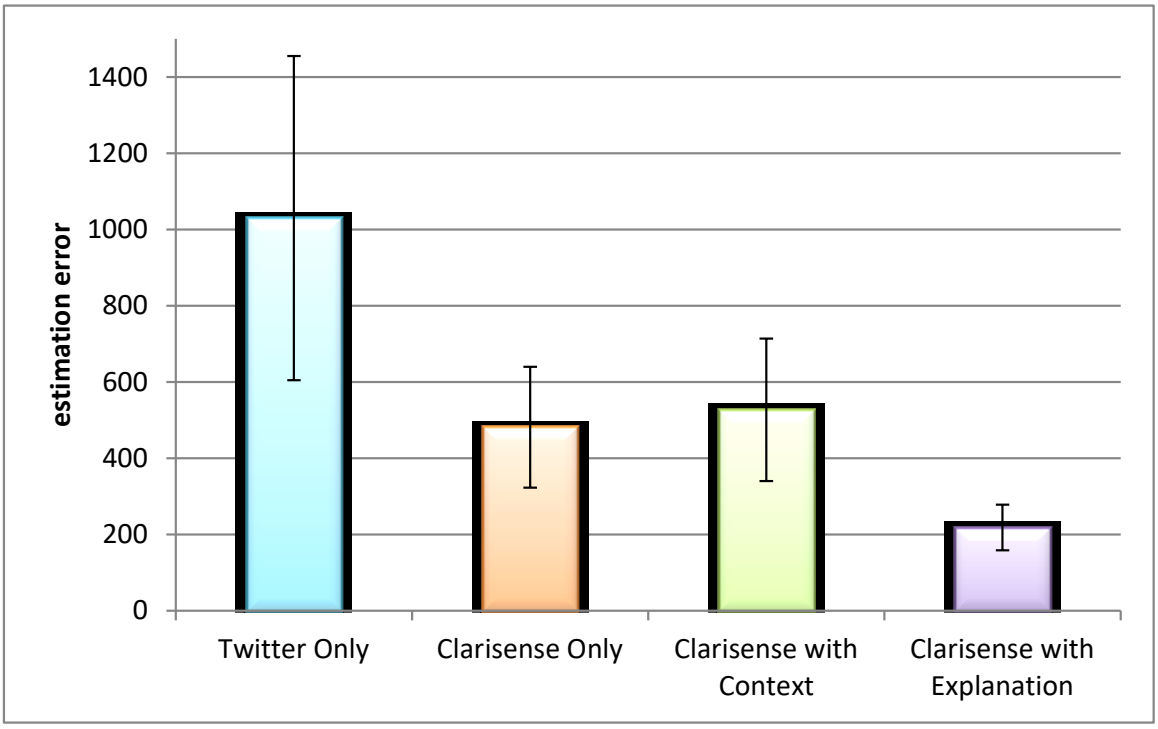


Figure 5.4: After the discovery process, participants that interacted with the anomaly recommender consistently had a better understanding of the data. The presence of explanation and provenance improved understanding further. Error bars show one standard error below and above the mean.

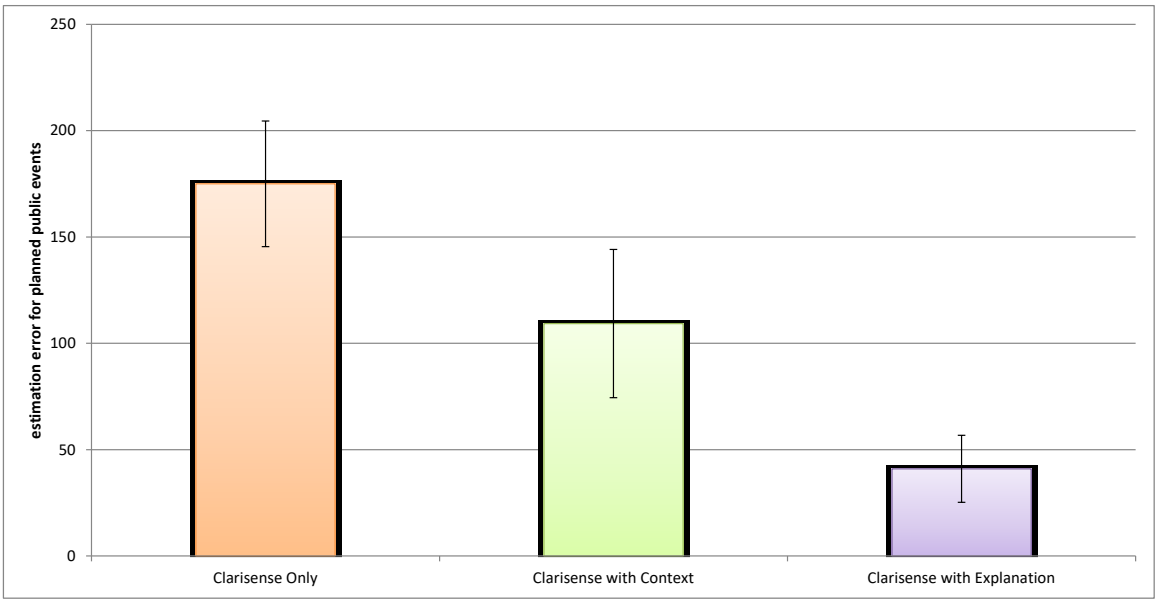


Figure 5.5: The anomaly recommender was much more likely to report major public events. The level of explanation greatly decreased the participant's error in perception for the frequency of these types of traffic blockages. Error bars show one standard error below and above the mean.

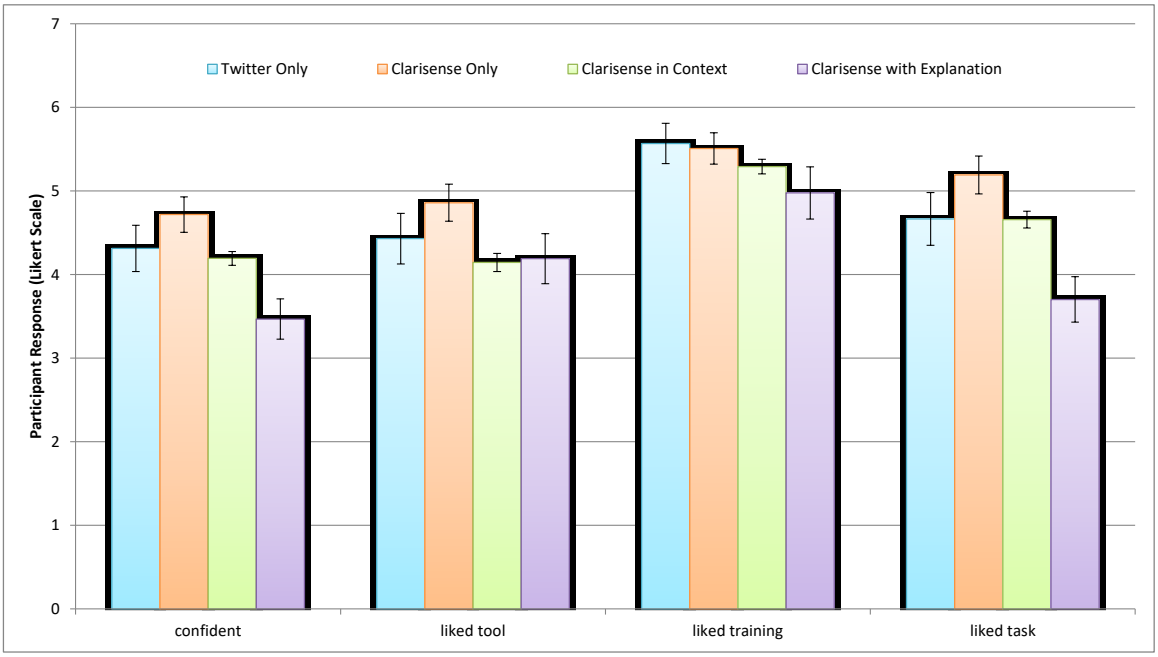


Figure 5.6: Presence of the original dataset and an increase in explanation corresponded to a decrease in confidence and enjoyment of the task.

Chapter 6

A Measurement Ontology for Human-Agent Interaction in Information Analysis

In this chapter, we describe a high-level, theoretical measurement framework for modeling human cognition responses to information systems during information analysis. The framework measurements were chosen to be as general as possible. We describe the explanation, control, and error (ECR) profile for information tools, four latent parameters for profiling users (insight, cognitive reflection, trust propensity, reported expertise), and four latent parameters for modeling inter-task cognition (cognitive load, SAT, insight, and system trust). These eleven factors are used to explain interaction decisions, adherence decisions, and domain decisions. We show how measurements of the general parameters can be concretely specified while still remaining true to the general framework. We propose the use of structural equation modeling (SEM) to fit experimental data to the causal model, due to its ability to handle latent variables and error. An evaluation of the measurement framework is given in Chapter 9.

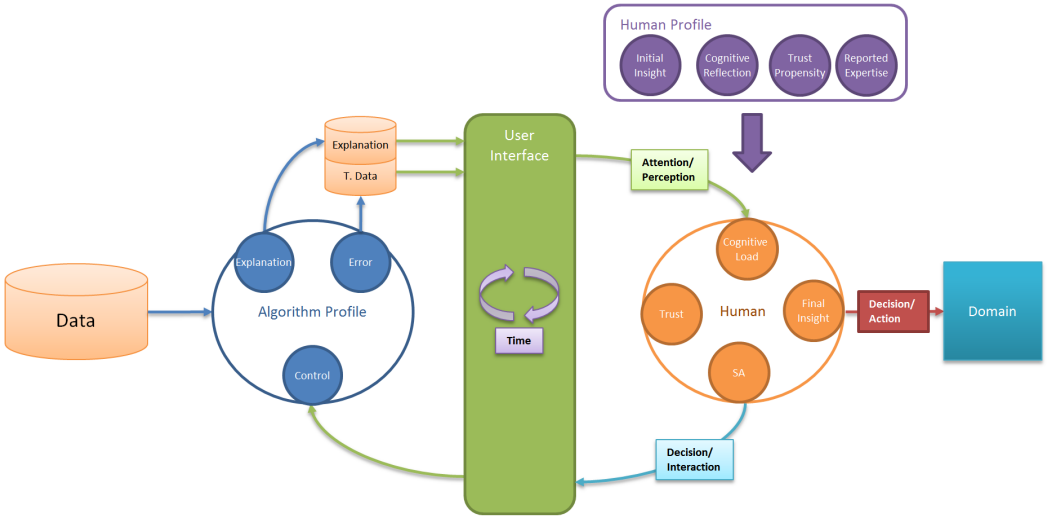


Figure 6.1: Measurement ontology, with labeled human profiling factors (purple), algorithm profiling factors (blue), and inter-task cognitive factors (orange).

6.1 Measurement Framework

Our framework draws large inspiration from models recently proposed in visual analytics [7], is expanded to accommodate the complexity of some real world situations, where users may be interacting with multiple user interfaces (and thus multiple representations of data) to control multiple information tools. We also incorporate concepts from Norman's theories of usability [5] and Mica's theory of situation awareness [11] that identify parameters that play significant roles in the interaction and decision making process.

The framework shown in Figure 6.1 applies to a user interacting with a single information tool through a single user interface. Other information tools are not shown in this perspective.

We'll make several arguments for the adoption of this measurement ontology. Previous visual analytics models have included insight into their models [7], but they typically do not include measurements for success of domain decision making, and it is implicitly

assumed that improved accuracy of information tools and usability of user interfaces will increase insight, which translates into better decisions made in the larger domain. However, changes to designs of models can increase the time that users interact with the system, which in many domains (such as live operation of a vehicle) directly translates to worse decisions and failed goals. For that reason, measurements of domain decision making need to be included in the model and ideally measured experimentally, possibly with hypothetical scenarios or using abstract games where decision quality can be directly observed. Second, we believe it is important to incorporate a user's beliefs about an information tool, as it may have a significant relationship with trust (people are more trusting of transparent tools), adherence (people like using tools when they feel in control), and insight (people will better understand raw sensor data by viewing transformed data if they understand the transformations). This second point is illustrated in Figre 6.2. Third and finally, we show how user interfaces and information tools are interdependent. User interfaces cannot expose control or explanation features to users without those features already being present (or possible) in algorithms. Information tools may not be able to convey these features effectively without proper user interface design.

In many data analysis tasks, multiple information tools are invoked. Measurements in this framework can be applied to these more complex environments by repeating measurements for each information tool defined. Adherence thus measures the proportion of information incorporation from a specific information tool. For different contexts, adherence could be measured in different ways based on the goals of each information tool, e.g. the distribution of results accessed by the user, the ratio of user actions as spread among the models, or the agreement between model output and domain action. Adherence has been measured in expert systems literature , and has been shown to have a strong relationship with explanation level [34].

In practice, the multi-tool framework is the one to apply new user tasks, tool config-

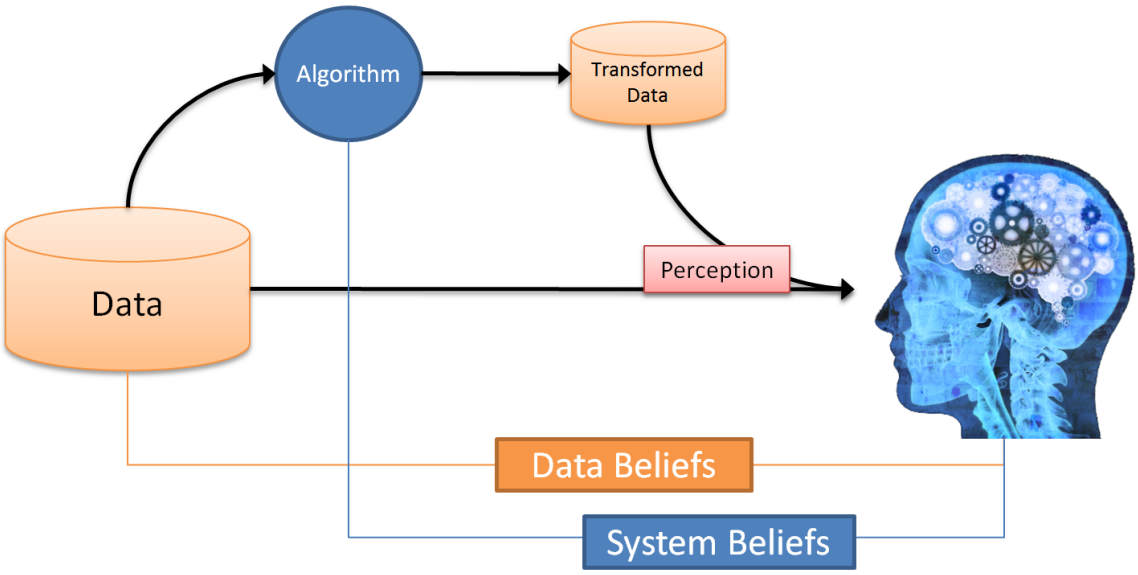


Figure 6.2: This study hypothesizes that user beliefs about data will be less accurate when user beliefs about complex algorithms are not accurate.

.

urations, and domains. For each information tool, parameters will have to be assigned and measured separately. It is conceivable to aggregate some parameters across all tools (for instance, to see how the *global* situation awareness of a user changes as more tools are added to the system, or to see how much users trust the system as a whole), but this work is focused on understanding how provenance for *individual* tools affect outcomes and we compare our assessed tools against simple, well-established alternatives.

6.2 Statistical Modeling

Different tool designers might have different goals, e.g. just get the user to use the system (recommender designer), maximize the user's understanding of the data (visual analytics designer), or improve decision making (cognitive agent designer). Here, we hypothesize that all such parameters may be interrelated, and a complete set of measurements as described by our framework will create the best understanding of the effects

Parameter	**Latent?**	**Description**
Explanation Level	No	Higher levels of explanation mean a larger amount of output bandwidth of the information tool is allocated to indicating algorithm operation.
Control Level	No	Higher levels of control mean a larger amount of input bandwidth of the information tool is allocated to the user.
Error Level	No	The algorithm's accuracy and relevance when satisfying information requirements.

Table 6.1: Information tool/algorithm profiling factors.

of different design decisions. For this reason, we use structural equation modeling (SEM) to uncover relationships between variables. SEM has increased in popularity due to its ability to model complex systems of variables while taking error into account [116]. Below, we outline the factors that need to be specified, and discuss how indicator variables can be crafted for each latent parameter.

From the framework in Figure 6.1, there are eleven system/user factors. There are also three important system measurements: user decisions to interact, user adherence to information tool output, domain decision-making success. Designers of information tools have the ability to affect the explanation, control, and error of their tool. User profiling factors and cognitive responses are assumed to be out of the control of the system designer.

6.3 Factors in Abstract

Explanation Level: The "level" of explanation. Higher levels of explanation means a relatively larger ratio of the tool's output bandwidth (for instance, measured in bytes) is dedicated to generating an understanding of the algorithm operations rather than satisfying the user's information requirements. Algorithms that have more *effective* explanation generate higher level of situation awareness in the user in a smaller amount of

Parameter	Latent?	Description
Cognitive Reflection	Yes	The user's ability to engage type II thinking as indicated by a cognitive reflection test (CRT).
Reported Expertise	Yes	The user's self-reported expertise as indicated by responses to questionnaire items.
Trust Propensity	Yes	The user's self-reported propensity to trust information tools and technology as indicated by responses to questionnaire items.
Initial Insight	Yes	The user's pre-task beliefs about the information space and whether those beliefs are accurate.

Table 6.2: User profiling factors.

Parameter	Latent?	Description
SAT	Yes	The user's beliefs about the information tool and whether those beliefs are accurate.
Cognitive Load	Yes	The total amount of mental effort exerted by the user.
Trust and Perceptions	Yes	The degree to which the user trusts or thinks highly of the tool and will rely on its output.
Final Insight	Yes	The user's inter- or post-task beliefs about the information space and whether those beliefs are accurate.

Table 6.3: Inter-task cognition factors.

Parameter	Latent?	Description
Adherence	No	The degree to which information provided by a tool is incorporated into user decision-making.
Interaction	No	The quantity, type, and quality of interaction with an information tool.
Decision Success	No	Domain action effectiveness, measured in domain terms.

Table 6.4: User decision-making factors.

time. Different tool configurations could easily be ordered in terms of explanation level, and user feedback in the form of questionnaires could provide extra validation.

Control Level: The amount of input bandwidth dedicated to the user, for instance, measured in the number of parameters the user has control over. More effective control features improve the ability of a user to affect the tool's output (either positively or negatively). Again, different configurations of tools could easily be ordered in terms of control level, and user feedback could provide extra validation.

Error Level: The accuracy of an algorithm, with respect to its stated goals. Relevance is also a concern. A GPS algorithm that solves a shortest-route problem cannot help a user looking for the easiest route. User feedback could be elicited to judge relevance.

Insight: The user's beliefs about data and whether those beliefs are correct. An insight can be expressed as a qualitative statement about the data that involves all or most of the data, might be unexpected, is framed in domain terms, and is possibly built on top of other insights [6]. Can be measured via a SAGAT Freeze with various quantitative and qualitative estimation questions.

Trust Propensity: The user's self-reported trust in information systems and tools. These can be likert scale questions such as "I trust automation" or "I would take advice from a cognitive agent if one were available."

Reported Expertise: The user's self-reported domain expertise. These can be likert scale questions such as "I am an expert on <x>" or "I am familiar with <x>."

Cognitive Reflection: The user's response to the cognitive reflection test [67]. The questions are:

1. A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost? (5 cents)

2. If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? (5 minutes)

3. In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? (47 days)

Due to the popularity of the cognitive reflection test, some participants may know the answers to these questions. Recent research has explored alternative versions of these questions [143].

1. If youre running a race and you pass the person in second place, what place are you in? (second)

2. A farmer had 15 sheep and all but 8 died. How many are left? (8)

3. Emilys father has three daughters. The first two are named April and May. What is the third daughters name? (Emily)

4. How many cubic feet of dirt are there in a hole that is 3 deep x 3 wide x 3 long? (none)

Situation-awareness based Agent Transparency (SAT): The user's beliefs about the model and whether those beliefs are correct. Specifically, we define the three levels of SAT for information tools:

- **Level 1 SAT: Perception**: The user develops an understanding of the information tool, perceived through a user interface, as being distinct from other information tools invoked by the interface, and is aware of its goals, the scope of its operations, and the time during which its operations are carried out.

- **Level 2 SAT: Comprehension**: The user develops an understanding of the procedural knowledge by which an information tool operates, and understands previous output by the tool, including when and how the tool's algorithm operated on a particular piece of data.

- **Level 3 SAT: Projection**: The user is able to predict how changes in control parameters will affect future outputs of the model.

SAT can be measured via a SAGAT Freeze [144] with various quantitative and qualitative estimation questions.

Cognitive Load: The total amount of mental effort exerted by the user when using the system. This is directly measurable by observing pupil dilation or with ECG equipment, but indirect methods such as questionnaires can work as well. Subjective cognitive load measures can be taken throughout an experiment using a SAGAT-style freeze.

Trust and Perception: The user's self-reported trust in an information tool and perceptions of its usefulness. Pearl and Knijnenburg have conducted useful quantitative investigations into user perceptions of recommender systems [18][19]. As evidenced in Chapters 7,8, and 9, user perceptions of a tool and trust are very highly correlated.

Adherence: The number of times a user's domain decision agrees with information tool output, or the number of times that a user interacts with a tool, based on the type of tool (does it merely inform, or does it provide a solution to a goal?)

Interaction: user choices about which control features to use, when to use them, and how often. This will be directly observable. It may also be necessary to determine if user interaction choices are *effective*, that is, they may increase or decrease algorithm error and relevance, based on the skill of the user.

Decision Success: The degree to which user decisions lead to success in the domain. This will generally be directly observable, but still must be tailored to each domain.

Goals, requirements, and satisfaction criterion for decision making are likely known at the time of system interaction. For instance, an intelligence analyst might be involved in the task of capturing military targets, so domain action effectiveness could be measured as the ratio of targets captured divided by the amount of time it took to capture those targets.

6.4 Constraints and Recommendations for Specifying Measurements

Setting up an experimental methodology using the above framework requires the specification of the three observable variables (interaction, adherence, decision success) and the creation of new indicator variables for each of the eleven factors. Constraints and guidelines will be given here.

First, **error level** should capture the degree to which the information tool satisfies its stated goal and to what degree that goal matches the user's goal. For this reason, multiple indicator variables might be needed, such as the algorithm's F-measure and accuracy measures (if applicable), the user's satisfaction with the tool's output, a statement from the user saying that the tool matches their intended goal, and the time required for the algorithm to produce output given a change in input. For instance, a GPS system that takes 6 hours to calculate a route, chooses surface streets over freeways, and occasionally makes an error is not going to be an effective choice for a user trying to reach a hospital. Information tools should be designed to be as effective as possible, but in theory we might want to know if a moderately effective tool will get more use than a highly effective one if the former was, for instance, more predictable or easier to explain. A hypothetical scenario could be created and noise could be added to the results from a mock algorithm

to independently control the quality of results. Recall from SEM theory that independent latent variables should have as many indicators as possible, and these variables should all be correlated for best fit.

Explanation level, like the effectiveness of results, is a parameter of the information tool and would have to be treated as an independent variable during experimentation. Provenance level, then, needs to order different configurations of the information tool by the degree to which they attempt to expose their operation. In many situations, this could be a simple matter of measuring the number of bytes the tool outputs that are for this purpose, but getting participant feedback in the form of questions such as "how well did the algorithm attempt to justify its output?" or similar questions, based on application, could mediate explanation level manipulations. Furthermore, these types of questions could be repeated for new systems (outside of experimentation) to make a comparison in provenance level, and thus draw conclusions about how well users will understand the system.

Control level is another parameter that would be controlled experimentally. Experiment designers will again have to rank different configurations of the model in terms of the number of parameters or input bandwidth they assign to users. Again, feedback questions on Likert scales such as "I felt in control of the tool (1-7)" could be used to mediate the effects of control treatments.

Insight should capture the degree to which the user understands the contents of the data, its distributions, and its implications for domain decisions. The only practical way to measure insight is with a test of knowledge, administered during or after interaction with the system, that captures relevant aspects of the data. This means that any data used experimentally will have to be well understood in advance, and tests must be crafted carefully as to reflect goals in the domain. North's characterization of insight [6] is a good starting point on requirements for the types of questions or tests used.

SAT should capture the degree to which the user understands the operation of the information tool. It might be best to measure this in a similar way to insight, with a short battery of understanding questions. Question items should attempt to capture all levels of SAT and would ideally be able to place the user in one of the three levels of SA. Endsley has shown how the SAGAT freeze can be used effectively in measuring the SA of pilots [13] since the interruption does not affect measurement.

As outlined in a previous chapter, measuring **trust** in information tools has been well studied. Trust in models is usually recorded with a series of questions, again on a Likert scale, such as "I trust the output from the model (1-7)" or in recommender systems "I trust the recommendations given to me (1-7)." Hypothetical scenarios could also be used in some domains, such as "If a friend were in danger I would adhere to the advice from this system (1-7)."

Cognitive load has been well studied in education science, and many measurement methodologies exist that remain domain independent, such as the use of ECG equipment and observation of pupil dilation. In experiment methodologies where this might not be possible, questions following a user task that ask for a stress level on likert scale might be appropriate for getting feedback. Question items could also be taken during the inter-task SAGAT freeze. Based on the type of information tool, these questions could be similar to "the task was easy to complete" or "I got frustrated during the task."

Chapter 7

Human Cognition and Recommender Systems: Effects of Manipulating ECR

In this chapter, we will discuss the results of a study that manipulated the ECR profile of a collaborative filtering algorithm Here, users interacted with a tool dubbed "Movie Miner" in order to construct a watchlist of 5-7 movies, which represented a domain decision. Movie Miner provided the users with two separate computational tools: a browser tool providing simple search/rank/filter operations and a more complex recommender. Results support the following claims:

- Measurement of user beliefs about an algorithm (SAT), cognitive reflection (CRT), and insight are important factors for explaining decision satisfaction variance and adherence
- In the measurement ontology, SAT is the most significant variable affecting adherence to recommendations

Figure 7.1: A screenshot of "Movie Miner," a movie discovery interface, which was used in the study. The browsing tool is shown on the left (blue) and the recommendation tool is shown on the right (brown). Users could search, rank, and filter on all movies in the data set using the browser, and the recommendations on the right interactively updated as rating data was provided (center, yellow). In the task, users were asked to find a set of interesting movies to watch (center, green) using whichever tool they most preferred.

- Interaction with a recommender system causes users to form incorrect beliefs about underlying data, but this can be mitigated through system explanations

- Explanation, control, and recommender error are contributing factors to user acceptance and satisfaction with selected items

7.1 System Design

This section describes the design of the interface in more detail. In designing the system for this study, we kept the following two goals in mind: a) to make the system as familiar to modern web users as possible and b) to make the system as similar to

currently deployed recommender systems as possible. The use of novelty in any design aspect was minimized so that results would have more impact on current practice.

Participants were presented with a user interface called *Movie Miner* (Figure 7.1). The interface was closely modeled after modern movie search and recommendation tools (such as IMDb or Movielens) and distinctive features were avoided. On the left side, the system featured basic search, rank, and filter for the entire movie data set. The right side of the interface provided a ranked list of recommendations derived from collaborative filtering, which interactively updated as rating data was provided.

The "Movielens 20M" data set was used for this experimental task. Having up-to-date movie references and ratings is important for the tasks in our experiment, as it is less likely that participants have seen many of the newly released movies when compared with older ones. Moreover, the Movielens dataset has been widely studied in recommender systems research [145][146][147]. Due to recommender speed limitations, the data set was randomly sampled for 4 million ratings, rather than the full 20 million.

7.1.1 Generating Recommendations

A traditional collaborative filtering approach was chosen for the system. Details for this can be found in Resnick et al [118]. Collaborative filtering was chosen due to the fact that it is well understood in the recommender systems community. The results from this study should generalize reasonably well to other collaborative-filtering based techniques, such as matrix factorization and neighborhood models [148]. Results from this study can inform the UI design of other recommendation algorithms, but only to the degree that they are similar to collaborative filtering. In this experiment, user-user similarity was used. We made two minor modifications to the default algorithm based on test results

from our benchmark data set: Herlocker damping and rating normalization[1].

Our user-user similarity function is Pearson correlation over the user profile of ratings, which is specified as.

$$sim(u,v) = \frac{\hat{u} \cdot \hat{v}}{||\hat{u}||||\hat{v}||} = \frac{\sum_i \hat{r_{ui}} \hat{r_{vi}}}{\sqrt{\sum_i \hat{r_{ui}}^2} \sqrt{\sum_i \hat{r_{vi}}^2}} \tag{7.1}$$

Where $r_u i$ is the rating given by user u to item i, $i \in I_u \cap I_v$ (I_u is the set of items rated by u), and $\hat{r_u} i$ is the normalized rating $r_{ui} - \mu_u$ (μ_u is the mean item rating for user u). Once $sim(u,v)$ was calculated, Herlocker damping [149], which penalizes popular items. Finally, the predicted rating, r_{ui}, $i \notin I_u$, was calculated as:

$$r_{ui} = \mu_u + k \sum_{v \in U} sim(u,v) \hat{r_{vi}} \tag{7.2}$$

Where k is a normalizing factor defined on all users as $k = 1/\sum_{v \in U} |sim(u,v)|$. We also slightly improved recommendations in our initial benchmark by multiplying r_{ui} by another normalizing factor over all predicted ratings $b = \max_{i \notin I_u} r_{ui}$, which spreads all predicted ratings the over full 0.5 to 5 star range, rather than assume there is a maximum rating for the user.

7.1.2 User Interface Design

General functionality that applied to the entire interface included the following: mousing over a movie would pop up a panel that contained the movie poster, metadata information, and a plot synopsis of the movie (taken from IMDb); for any movie, users could click anywhere on the star bar to provide a rating for that movie, and they could click the green "Add to watchlist" button to save the movie in their watchlist (we questioned

[1]Our approach was nearly identical to: http://grouplens.org/blog/similarity-functions-for-user-user-collaborative-filtering/

users about their chosen movies at the end of the task). Clicking the title of any movie would take a user to the IMDb page where a trailer could be watched (this was also available during the watchlist feedback stage, when decision satisfaction was measured).

Browser Side

On the left (browser) side of this interface, users had three primary modes of interaction which were modeled after the most typical features found on movie browsing websites:

1. **SEARCH**: Typing a keyword or phrase into the keyword matching box at the top of the list returned all movies that matched the keyword. Matches were not personalized in any way (a simple text matching algorithm was used).
2. **RANK**: Clicking a metadata parameter (e.g. Title, IMDb Rating, Release Date) at the top of the list re-sorted the movies according to that parameter. Users could also change the sort direction.
3. **FILTER**: Clicking "Add New Filter" at the top of the list brought up a small popup dialog that prompted the user for a min, max, or set coverage value of a metadata parameter. Users could add as many filters as they wanted and re-edit or delete them at any time.

Recommendation Side

The recommender features varied based on the treatment that was assigned to the user, but the movies that appeared on this side came from the same set of movies that were the basis for the filtering results on the left, with some differences in tool features. The first distinct difference is that the list was always sorted by predicted rating and the user could not override this behavior (even when maximum control was provided).

The keyword matching tool was also not available on this side. When maximum control was given, users could provide a filter in the same manner as on the search side. They also had the option to tell the recommender they were "Not interested" in a particular recommendation with a red button that appeared on each movie. When pressed, this button would permanently hide that movie from the recommendation list.

7.2 Experiment Design

Relationships between cognitive factors in the model were tested by constructing an SEM through a bottom-up process. Additionally, we investigated how much each factor can explain decision satisfaction variance by building different statistical models that incorporate or omit some factors and comparing the R^2 of decision satisfaction in each model.

7.2.1 Independent Variables

Two levels of control, two levels of explanation, and two levels of recommendation error were manipulated. All manipulations (3 parameters, 2 values taken, $2^3 = 8$ manipulations) were used as between-subjects treatments in this experiment. Note that since we are testing the effects of recommender configuration rather than the effects of recommender presence, this experiment's "baseline condition" corresponded to the treatment where control, explanation, and error are absent. An alternative baseline was considered where the recommender itself was absent, but this treatment was not tested. This choice allowed us to allocate more participants to each condition while still allowing us to indirectly tease out the effects of using either tool by measuring interaction and adherence.

Two alternatives were considered to vary the control level. The first alternative was

to take a similar approach to some visual recommendation algorithms [150][18][117] and allow users to override algorithm values. The second alternative was to allow users to define filters on the list of recommendations. The latter approach was chosen due to more similarity with real-world systems that are currently deployed on Movielens, IMDb, and so on.

Control Manipulation

- **Partial Control** The partial control configuration allowed users to manipulate a profile (with adds, deletes, or re-rates) to get dynamic recommender feedback.
- **Full Control** On top of the partial control features, users were allowed to define custom filters on recommender results to narrow the recommendations. Additionally, users could remove individual movies (indicating they were "not interested") from the recommendation list.

Text-based explanations were chosen due to their similarity to real world systems such as Netflix and Amazon.

Explanation Manipulation

- **Opaque** The opaque recommender simply provided the recommendations without any explanation.
- **Justification** The justification recommender explained how ratings were calculated with the following blurb: "Movie Miner matches you with other people who share your tastes to predict your rating." This was followed by a list of the items in the user's profile that most affected the recommendation (calculated via an intersection with the rated item sets of the user's profile and the top 3 most similar users).

Two alternatives were considered to vary recommendation error. The first alternative was to use two different algorithms and confirm a difference in accuracy post-hoc. The second alternative was to use the same algorithm with varying levels of noise added. The latter was chosen due to concerns about differences in speed between two different algorithms and ease of implementation. The approach was validated by verifying that the random noise was reducing accuracy by performing a 5-fold cross validation on our ratings data set. The error-free recommender achieved an MAE of 0.144, while the noisy version did considerably worse at 0.181 (nearly a 26% difference).

Error Manipulation

- **Collaborative Filtering** Collaborative filtering: user-user similarity, Herlocker damping, and normalized across the 0.5-5 star rating scale.
- **Collaborative Filtering w/Noise** A vector of noise (of up to 2 stars difference) was calculated at session start and the vector was added in to the recommendation vector before normalization. From the participant's perspective, the list of recommendations thus appeared to be reordered as affected by this noise.

7.2.2 Dependent Variables

Dependent variables consisted of observations of system interactions and more complex factors, which were collected through questionnaires. Basic dependent variables measured in this study were quantity (and type) of interaction with each tool. An important dependent variable, adherence, was measured as the percentage of items in each participant's watchlist that originated from the recommender side of the interface. Dependent variables are shown in Table 8.6. For the more complex dependent variables,

Factor	Item Description	R^2	Est.
Trust Prop.	I think I will trust the movie recommendations given in this task.	0.81	1.17
$ALPHA$: 0.92	I think I will be satisfied with the movie recommendations given in this task.	0.83	1.18
AVE : 0.80	I think the movie recommendations in this task will be accurate.	0.75	1.15
Movie Exp.	I am an expert on movies.	0.77	1.40
$ALPHA$: 0.82	I am a film enthusiast.	0.63	1.16
AVE : 0.61	I closely follow the directors that I like.	0.45	1.14
CRT	If it takes 5 machines 5 minutes to make 5 widgets...	0.54	0.37
$ALPHA$: 0.79	A bat and ball together cost $1.10, and the bat costs $1.00 more than the ball...	0.51	0.35
AVE : 0.55	In a pond there is a patch of lily pads that doubles in size every day...	0.59	0.38

Table 7.1: Factors corresponding to user metrics. R^2 reports the fit of the item to the factor. Est. is the estimated loading of the item to the factor. Items that were removed due to poor fit are not shown.

confirmatory factor analysis (CFA) was used to eliminate measurement error when possible. Structural equation modeling (SEM) was then used to test the relationships between the confirmed factors in our HAI model. A list of the subjective factors is shown in Tables 7.2 and 7.1, which includes the factors covered in the related work section in addition to two more user profiling factors: trust propensity and reported expertise in movies. All of these items were taken on a Likert scale, except for when ratings were elicited, where a 5-star rating bar was used. Additionally, for decision satisfaction, answers were averaged over the 5-7 movies chosen by the participant.

"User Experience" was intended to be split into subjective system aspects (SSA, similar to [19]) such as perceived transparency, perceived control, perceived usefulness, and trust in the recommender (this is reflected in the questions that were chosen). Although item fit was acceptable for these sub-factors, very high correlations among them indicated they were better represented as a single scale (i.e. the participants had a uni-dimensional "good" or "bad" impression of the recommender) and collapsing the items onto one fac-

Factor	Item Description	R^2	Est.
User Exp.	How understandable were the recommendations?	0.51	1.04
$ALPHA$: 0.93	Movie Miner succeeded at justifying its recommendations.	0.73	1.32
AVE : 0.68	The recommendations seemed to be completely random.	0.41	-1.09
	I preferred these recommendations over past recommendations.	0.64	1.27
	How accurate do you think the recommendations were?	0.77	1.35
	How satisfied were you with the recommendations?	0.84	1.45
	To what degree did the recommendations help you find movies for your watchlist?	0.65	1.26
	How much control do you feel you had over which movies were recommended?	0.62	1.14
	To what degree do you think you positively improved recommendations?	0.60	1.09
	I could get Movie Miner to show the recommendations I wanted.	0.67	1.27
	I trust the recommendations.	0.85	1.42
	I feel like I could rely on Movie Miner's recommendations in the future.	0.83	1.48
	I would advise a friend to use the recommender.	0.72	1.43
Cognitive Load	There was too much information on the screen	0.48	1.11
$ALPHA$: 0.82	I got lost when performing the task.	0.36	0.79
AVE : 0.55	Interacting with Movie Miner was frustrating.	0.67	1.23
	I felt overwhelmed when using Movie Miner.	0.64	1.17
Decision Sat.	How excited are you to watch <movie>?	0.78	0.66
$ALPHA$: 0.93	How satisfied were you with your choice in <movie>?	0.89	0.70
AVE : 0.83	How much do you think you will enjoy <movie>?	0.92	0.67
	What rating do you think you will end up giving to <movie>?	0.57	0.34

Table 7.2: Factors determined by participant responses to subjective questions about task experiences. R^2 reports the fit of the item to the factor. Est. is the estimated loading of the item to the factor. Items that were removed due to poor fit are not shown.

Factor	Item Description
Situation Aware. all-item parcel	1. What is the recommender trying to predict? 2. Are the recommendations I see just for me? 3. What are the recommendations affected by? 4. What are the recommendations based on? 5. When does the recommender update? 6. What happens if I delete all drama movies from my ratings? 7.What if I were to highly rate movies in the Sci-Fi genre? 8.What happens if I rate more movies according to my tastes? 9. What happens if I remove accurate ratings?
Insight all-item parcel	1. Online, which genre has the highest current average audience rating? 2. Online, which of these genres tends to be the most common among the movies with the highest average audience rating? 3. Online, which of these genres has the highest current popularity? 4. Generally, which of these genres has the most titles released, for all time periods? 5. Online, which of these decades has the highest current average audience rating? 6. How many movies have an average audience rating great than 9/10? 7. Popular movies tend to have an average rating that is lower average|higher? 8. Movies with an average rating of 9/10 or higher tend to have fewer|average|more votes?

Table 7.3: Factors determined by participant responses to insight and situation awareness questions. Multiple choice answers were given. Insight was measured at the beginning (initial insight) and the end of the study (final insight). Situation awareness was measured 8 minutes into the study, during the watchlist phase.

Variable Name	Description	μ	σ
Browser Int.	Number of interactions with the browser tool	37	23
Recommender Int.	Number of interactions with the recommender	14	29
Adherence	Proportion of the final watchlist that were recommendations	0.67	0.36
Initial Insight	Score on initial insight questionnaire (out of 8)	3.60	1.30
Final Insight	Score on final insight questionnaire (out of 8)	3.44	1.27
SAT	Score on recommender beliefs questionnaire (out of 9)	5.99	1.75

Table 7.4: Observed dependent variables in the study.

tor both improved factor and final model fit. This is reflected in Cronbach's alpha of the scale (0.93). After this modification, none of the latent factors in Tables 7.2 and 7.1 had a co-variance higher than 0.5 which indicated good discriminant validity between the factors.

We used a SAGAT-style freeze [144] during the movie selection task to assess situation awareness-based agent transparency (SAT), which showed 9 questions related to ground truth of recommender behavior. Insight was measured twice - before and after the user finished interacting with Movie Miner. The test was a set of eight questions which relates to knowledge of the movie metadata space. The questions were chosen so that someone who had a lot of experience searching for movies online would be able to answer correctly. We used all-item parcels for insight and SAT and the variance was fixed to the variance of the sample population. Both the SAT and insight metrics are shown in Table 7.3.

7.2.3 Procedure

Participants were recruited on Amazon Mechanical Turk (AMT). AMT is a web service that gives tools to researchers who require large numbers of participants and are capable of collecting data for their experiment in an online setting. AMT has been studied extensively for validity, notably Buhrmester [114] has found that the quality of data collected from MTurk is comparable to what would be collected from laboratory experiments [115]. Furthermore, since clickstream data can be collected, satisficing is easy to detect.

Participants made their way through four phases: the pre-study, the ratings phase, the watchlist phase, and the post-study.

The pre-study and post-study were designed using Qualtrics[2]. Items related to trust propensity, movie expertise, and cognitive reflection (CRT) (also, see Toplak et al [69])

[2]https://www.qualtrics.com/

were collected during this phase using the question items shown at the top of Table 7.1. Questions related to insight were shown following these first three items (and were shown again after the watchlist phase, before the post-study).

Next, in the "ratings phase," participants accessed Movie Miner and were shown only the blue *Movie Database* list and the ratings box (refer back to Figure 7.1). We asked participants to rate *at least* 10 movies that they believed would best represent their tastes, but many participants rated more than the minimum.

In the "watchlist phase," participants were shown the brown *Recommended for You* list and the watchlist box. Instructions appeared in a popup window and were also shown at the top of the screen when the popup was closed. Participants were told to freely use whichever tool they preferred to find some new movies to watch. They could add movies to their watchlist with the green button that appeared on each individual movie (regardless of the list that it appeared in). We asked them not to add any movies that they had already seen, required them to add at least 5 movies (limited to 7 maximum), and we required them to spend at least 12 minutes interacting with the interface. At the end of this phase, they were asked about each of the movies they had added to their watchlist to measure decision satisfaction.

Finally, the questions related to insight were shown again. Then, we showed questions related to perceived transparency, perceived recommendation quality, cognitive load, and trust in the recommender.

The use of a minimum time limit allowed us to do several things. First, we did not want to force the participants to interact with either system since doing so would not allow us to make any observations about what they would choose on their own. Second, we wanted to understand how insight would change over time when interacting with the recommendation system and/or movie browser. Attempting to changes to insight with a protocol that freely allowed participants to move to the next step would have been

problematic. Third, we wanted the task to mirror real-world situations as closely as possible and thus the session needed to be exploratory. A twelve minute session in which 5-7 items are selected was also sufficient time to select quality items, given that people only browse Netflix for 60-90 seconds to find a single item before giving up [151].

7.3 Results

We collected more than 526 samples of participant data using AMT. Participants were paid $1.50 and spent between 25 and 60 minutes doing the study. Participants were between 18 and 71 years of age and were 45% male. Participant data was checked carefully for satisficing and these records were removed, resulting in the 526 complete records.

7.3.1 All-Factor SEM

Regressions in the SEM and fit of the all-factor model are shown in Table 7.5. Model co-variances are shown in Table 7.6. Note that the change in insight is not explained by this model and we initial insight is not regressed onto final insight (a model in the following subsection explains the change). An illustration of the SEM was omitted due to visual complexity. Model fit: $N = 526$ with 99 free parameters = 5 participants per free parameter, $RMSEA = 0.041$ ($CI : [0.038, 0.044]$), $TLI = 0.94, CFI = 0.94$ over null baseline model, $\chi^2(815) = 1543.588$. The model was built using R 3.0.3, lavaan 0.5-17.

7.3.2 Raykov Change Model

A Raykov structural model [152] was built to identify factors that predict changed belief between the initial and final insight tests. The final model is shown in Figures

Regressand	Regression (←)	Coeff.	P(> |z|)
User Experience R^2**=0.17**	← Trust Propensity	0.35	***
	← Initial Insight	-0.08	*
	← Error	-0.34	***
Cognitive Load R^2**=0.03**	← Control	0.24	**
	← Trust Propensity	-0.14	**
SAT R^2**=0.11**	← Trust Propensity	-0.15	***
	← CRT	0.26	***
	← Explanation	0.30	*
Browser Int. R^2**=0.03**	← Trust Propensity	-0.15	**
	← Initial Insight	0.09	*
Recommender Int. R^2**=0.09**	← Movie Expertise	-0.16	***
	← Initial Insight	0.17	***
	← Control	0.38	***
Adherence R^2**=0.09**	← User Experience	0.10	*
	← Control	0.21	*
	← SAT	0.23	***
	← Browser Int.	-0.17	***
Final Insight R^2**=0.07**	← SAT	0.09	*
	← CRT	0.19	***
	← Explanation	-0.29	0.09
	← Control	-0.34	*
	← Error	-0.31	0.60
	← Expl. x Control	0.50	*
	← Expl. x Error	0.43	0.076
	← Control x Error	0.58	*
	← Expl. x Cont. x Error	-0.83	*
Decision Satisfaction R^2**=0.24**	← Trust Propensity	0.2	***
	← User Experience	0.29	***
	← Browser Int.	0.07	0.085
	← Recommender Int.	-0.20	***
	← Explanation	0.25	*
	← Control	0.30	**
	← Expl. x Control	-0.47	**

Table 7.5: Regressions in the fitted all-factor SEM, which attempts to explain decision satisfaction, final insight, and adherence to recommendations. Variables on the left are explained by variables on the right. Due to being non-normal, treatment variables take the value of 0 or 1 and coefficients reported are B values, which predict a change in standard deviation of the regressand when the treatment is switched on. All dependent and latent variables were standardized to have a mean of 0 and variance 1 and coefficients reported are β values, which predict a change in standard deviation of the regressand with a standard deviation change in the regressor. Significance levels for this table: *** $p < .001$, ** $p < .01$, * $p < .05$. Covariances are shown in Table 7.6.

Covariance (↔)	**Beta**	**P(> |z|)**
SAT ↔ Browser Int.	0.2	***
Trust Propensity ↔ Movie Expertise	0.46	***
Trust Propensity ↔ CRT	-0.23	***
Trust Propensity ↔ Initial Insight	-0.11	**
Movie Expertise ↔ CRT	-0.161	***
CRT ↔ Initial Insight	0.29	***
User Experience ↔ Cognitive Load	-0.54	***

Table 7.6: Covariances in the fitted model. Regressions are shown in Table 7.5. User experience and cognitive load were negatively correlated. A covariance was tested between decision satisfaction and adherence but none was found.

8.3 and 8.4. In each graph, the y-axis indicates the fitted, standardized beta parameter (mean 0, variance 1) for insight. The x-axis shows the difference between the initial and final tests. The slope indicates the rate at which beliefs became more correct or incorrect.

Figure 8.3 indicates that participants that interacted with the recommender more were predicted to have higher initial insight but also formed incorrect beliefs during the task. Participants with higher CRT also scored significantly better on the initial insight test but were also predicted to form incorrect beliefs. Users that reported having a better experience with the recommender had lower initial insight but were also predicted to learn the most during the task.

Figure 8.4 predicted that explanation, control, and noise all caused incorrect beliefs to be formed. Users in the "control only" condition actually formed incorrect beliefs about two of the eight insight questions (β=-0.25). Formation of incorrect beliefs was mitigated somewhat when noise and explanation were present.

7.3.3 False Discovery Rate Analysis

Multiplicity control was enforced in our chosen SEM using the Benjamini-Hochberg procedure with $Q = 0.10$ [153], which is recommended for exploratory SEM analysis [154]. This procedure indicates how many of the tested relationships in the All-factor

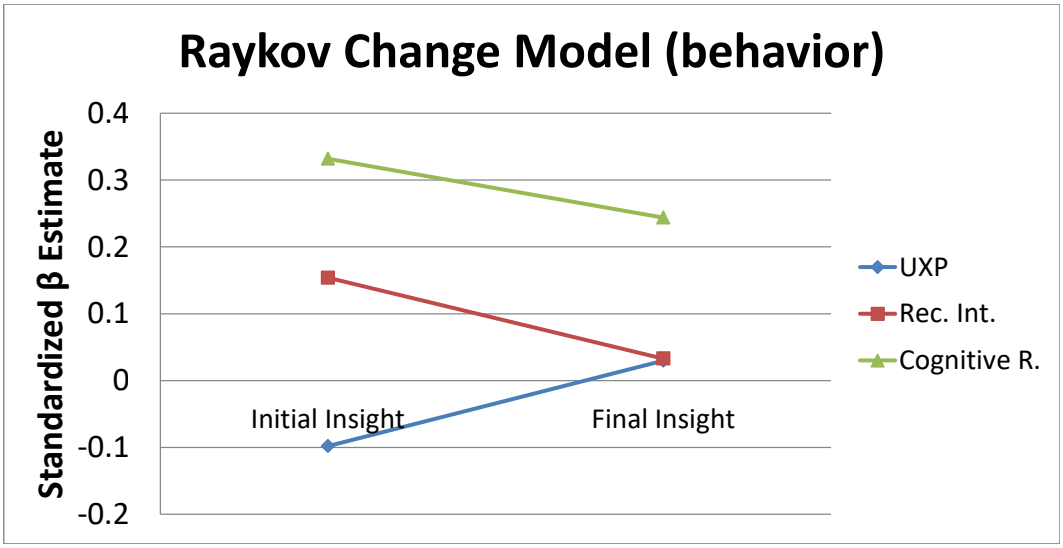


Figure 7.2: A Raykov model explaining the rate at which beliefs about the data changed between the initial and final insight tests as predicted by user cognition and personal attributes.

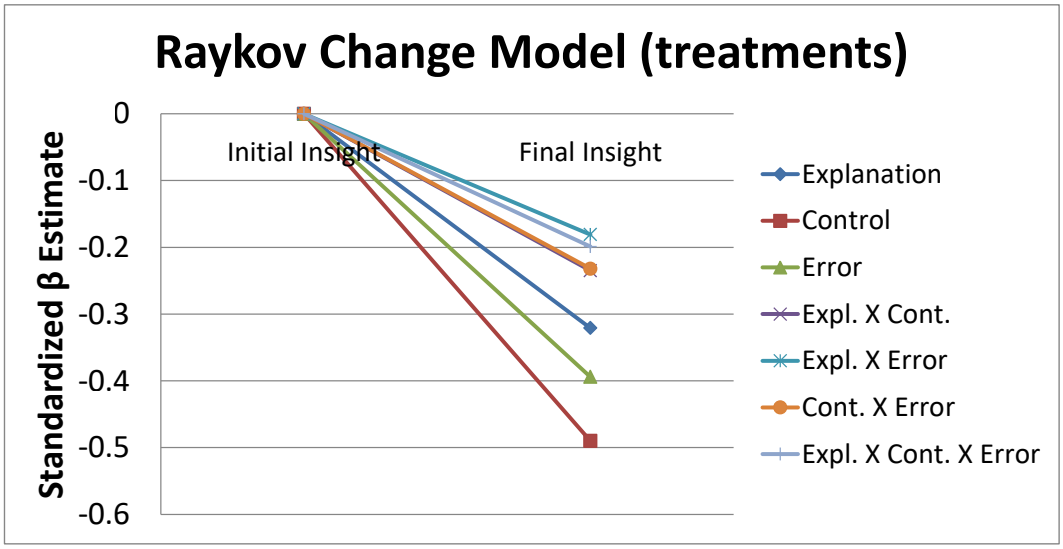


Figure 7.3: A Raykov model explaining the rate at which beliefs about the data changed between the initial and final insight tests as predicted by treatments

SEM are expected to be false positives. All reported effects in Tables 7.5 and 7.6 passed the FDR, with the most minor finding ($FinalInsight \leftarrow Explanation$) having $p = 0.085$ and $(i/m)Q = 0.10$.

7.4 Discussion of Results

Here we discuss four implications of the results of this work. The results discussed here are specific to this task domain. A meta-analysis of all studies in this dissertation is given in Chapter 9.

1. User experience only explained part of the decision-making process. First, we note that almost all factors that were chosen for the model were able to explain some part of decision satisfaction and adherence. This is evidenced in Table 7.5 and Figure 9.1. As indicated by the all-factor model, the exception to this was cognitive load, which does not correlate with either outcome (adherence or decision satisfaction). However, cognitive load and user experience were strongly negatively correlated. An alternative model that uses cognitive load to explain adherence and decision satisfaction fits the data almost as well as using user experience. This result reinforces the idea that cognitive load and user experience have an inverse relationship (see Jung [146]).

As Figure 8.10 illustrates, user experience was an important mediating variable for decision satisfaction. User experience fully mediated initial insight, trust propensity, and error with respect to decision satisfaction. However, it may be possible that users that are "easy to satisfy" reported higher satisfaction with both the recommender and their selected movies. The authors recommend controlling for this in future recommender evaluations.

Second, splitting user experience into SSA (similar to the ResQue framework [19] and the model in Knijnenburg et al [18]) decreased model fit, despite each sub-aspect

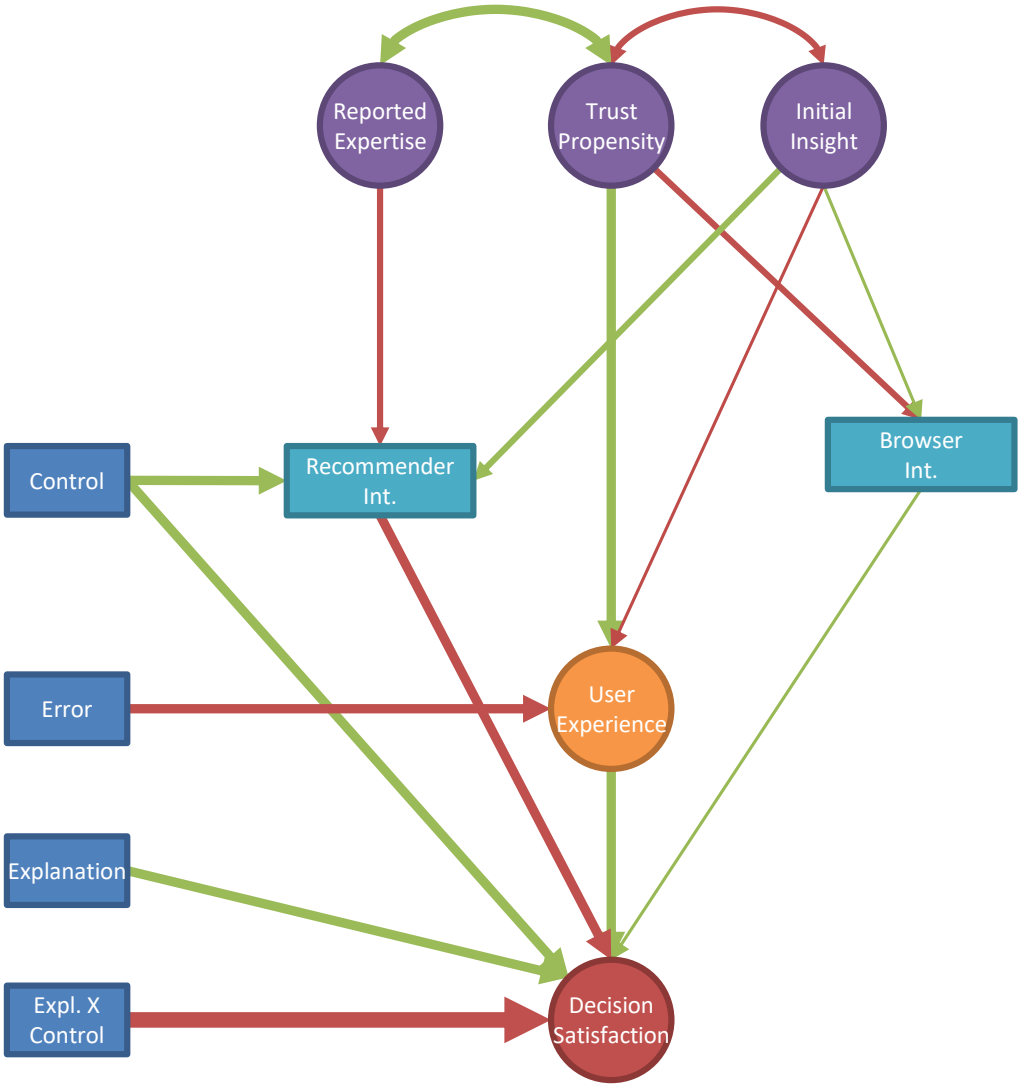


Figure 7.4: A visualization of all factors that directly or indirectly affect decision satisfaction. Line thickness indicates magnitude of β or B parameter. Green lines indicate a positive effect. Red lines indicate a negative effect. Note that this is just a visualization of a portion of the model reported in Table 7.5, and does not represent a tested statistical model.

(perceived transparency, perceived control, perceived quality, and trust) having items with acceptable fit but high inter-correlation (about 0.95). Generally, high correlations among factors are undesirable due to the decreased questionnaire item-to-information ratio. For instance, in this study, a 3-item scale for "trust" would have captured nearly the same signal as the 12-item SSA model that was used. This may have occurred because participants had a unidimensional perception of the recommender (i.e. "I like this" or "I don't like this"), which was a surprising finding. We considered it important to compare our results with Knijnenburg et al and the "ResQue" framework. The Knijnenburg data was available[3] and we examined the co-variances of perceived quality, satisfaction, control, and understandability. The scales in Knijnenburg's study were slightly better in terms of discriminatory power: about a 0.7 Pearson correlation between perceived overall system satisfaction, quality, and control, but this correlation level is still quite high. The transparency sub-construct, "understandability," is much more discriminating (0.34), perhaps due to the user-centric phrasings used. Unfortunately, discriminant validity between factors in the "ResQue" framework were not reported. In light of this analysis, we encourage other researchers to consider the inter-factor correlations and discriminant validity of their chosen factors.

We believe the results in this work help to demonstrate the value of insight, SAT, and cognitive reflection tests for recommender systems research. These constructs significantly increased the amount of explainable variance in decision satisfaction and adherence without affecting the order of complexity of the regression model. Moreover, their correlation with user experience constructs was quite low. Given that there were high correlations between user experience constructs in this experiment, it might be advisable to reduce the number of subjective user experience questionnaire items and instead use participant time to assess cognition. Many findings in this experiment would have been

[3]http://www.usabart.nl/QRMS/

missed if these measures had been omitted.

2. Users with correct beliefs about the recommender were more likely to adopt recommendations. Figure 8.5 visualizes all factors that affect adherence. SAT had the highest direct positive impact on adherence with a β coefficient of 0.23, followed closely by the presence of control. User experience did not predict adherence nearly as well as the SAT factor and the control treatment. Furthermore, the "perceived transparency" sub-construct was not nearly as effective at explaining adherence (the tested relationship was non-significant in all models). This highlights the need for the use of the objective SAT measure, instead of perceived transparency, within recommender systems research. Additionally, it highlights the need for recommender system designers to instill deep understanding of recommender operations to maximize usage.

3. Adding Explanation, Control, or Error to the Recommender Caused Incorrect Beliefs about the Data. Users in all non-baseline treatments were predicted to more inaccurate beliefs about the underlying movie data. Moreover, as Figure 8.5 illustrates, the all-factor SEM predicted that users in the baseline condition with high SAT had the most accurate beliefs about the data when the task was over (this can also be observed in Figure 7.7). The difference might be attributed to the particular way that the recommender "visualizes" the underlying data. Visualization theory predicts that users try to match their mental model with the information that is presented [155][24]. In this experiment, participants likely tried to reconcile their mental model with what the recommender displayed and made mistakes when their beliefs about the recommender were incorrect (especially in treatments where error was present). Control drew significant amounts of attention towards the recommender, with the control feature significantly increasing the amount of interaction (see Figure 7.9) and thus increasing the amount of time that the user was viewing recommendations. As Figure 8.3 indicates, increased recommender interaction also predicted the formation of more incorrect beliefs.

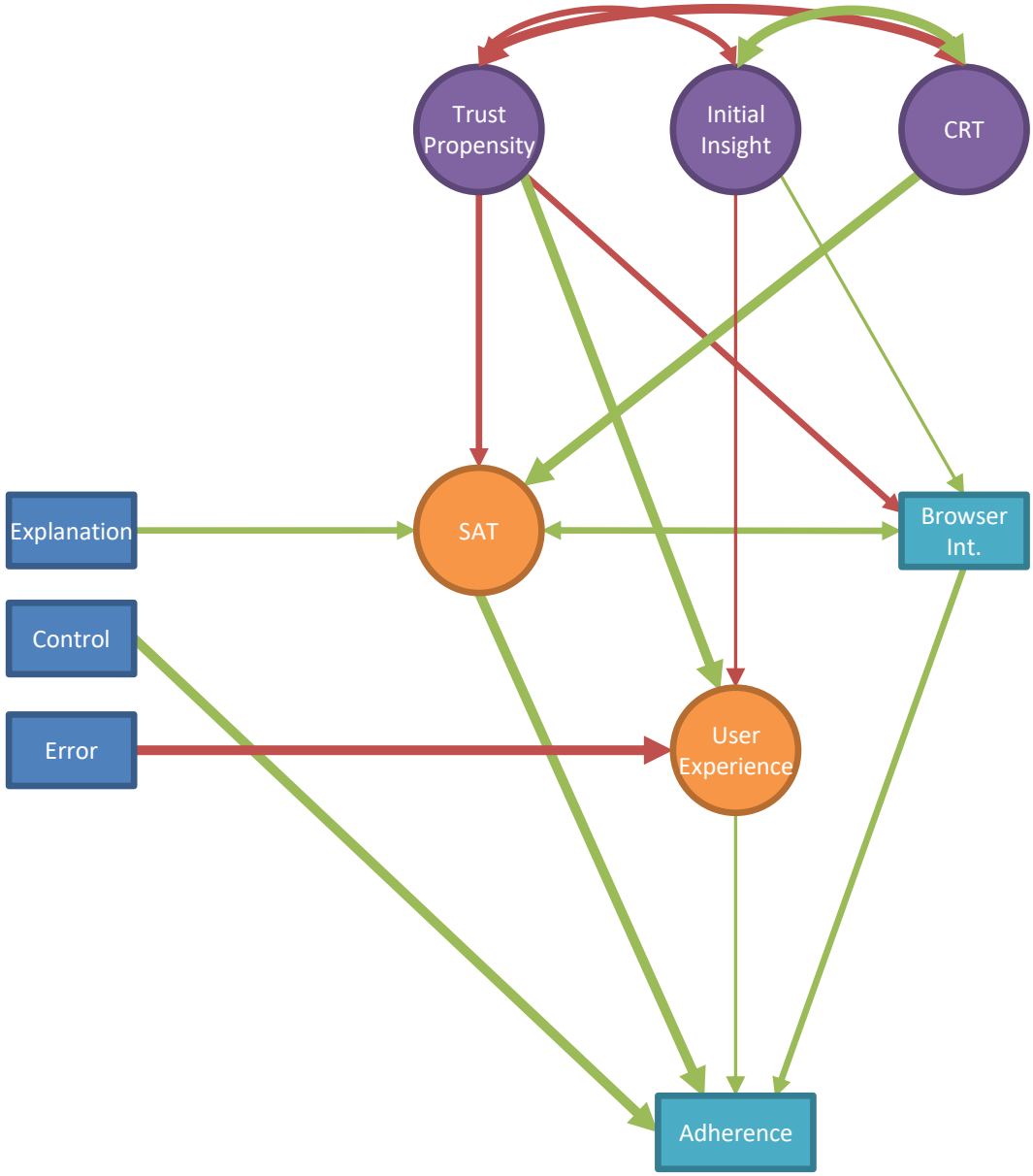


Figure 7.5: A visualization of all factors that directly or indirectly affect adherence. Line thickness indicates magnitude of β or B parameter. Green lines indicate a positive effect. Red lines indicate a negative effect. Note that this is just a visualization of a portion of the model reported in Table 7.5, and does not represent a tested statistical model.

Explanations also appeared to draw attention away from the browser tool, evidenced by lower interaction (Figure 7.8), but they also increased SAT to some degree, which in turn helped to maintain insight.

4. Explanation, control, and recommendation error steered the decision system towards different outcomes. Explanation played two roles. First, explanation improved SAT to a slight degree, which in turn correlated with increased adoption of recommendations. Second, explanation directly improved decision satisfaction regardless of participant interaction. However, increased interaction with the browser side of the interface was linked to increased SAT but also to decreased adherence. To explain this, we examined browser interaction in more detail. We found that, similar to the recommendation side, 50% of browser interaction were filter/sort/search actions and the other 50% were rating actions. What this might suggest is that participants were using the browser tool to find representative movies for their profile. As the participant found more representative items, there was more opportunity to get dynamic feedback from the recommender. Over time, this improved SAT but also increased the chance that the participant found satisfactory items from the browser tool (interesting titles were likely adjacent in metadata space to the targeted titles).

Control also played two roles. First, control (predictably) increased recommender interaction, which in turn correlated with increased cognitive load and decreased decision satisfaction. Second, the presence of control features increased adherence and recommender satisfaction regardless of interaction quantity. These findings reinforce the idea that users who interact more are harder to satisfy. Note that showing explanations and exposing control features together mitigated some of the benefit of doing either. The pop-up style explanations may have frustrated some users, affecting user experience and thus decision satisfaction.

The results from this study suggest that users benefit when a dynamic list of recom-

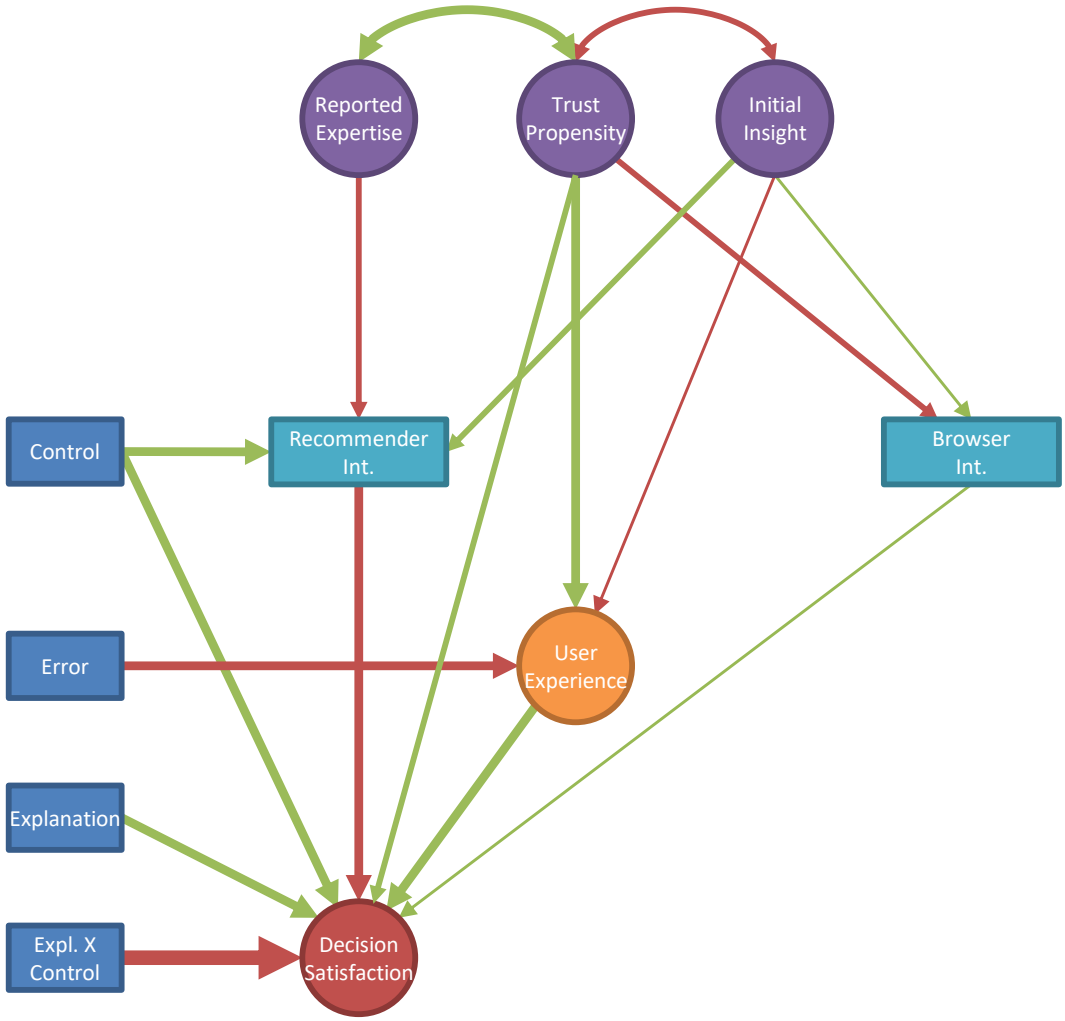


Figure 7.6: A visualization of all factors that directly or indirectly affect final insight. Line thickness indicates magnitude of β or B parameter. Green lines indicate a positive effect. Red lines indicate a negative effect. Note that this is just a visualization of a portion of the model reported in Table 7.5, and does not represent a tested statistical model.

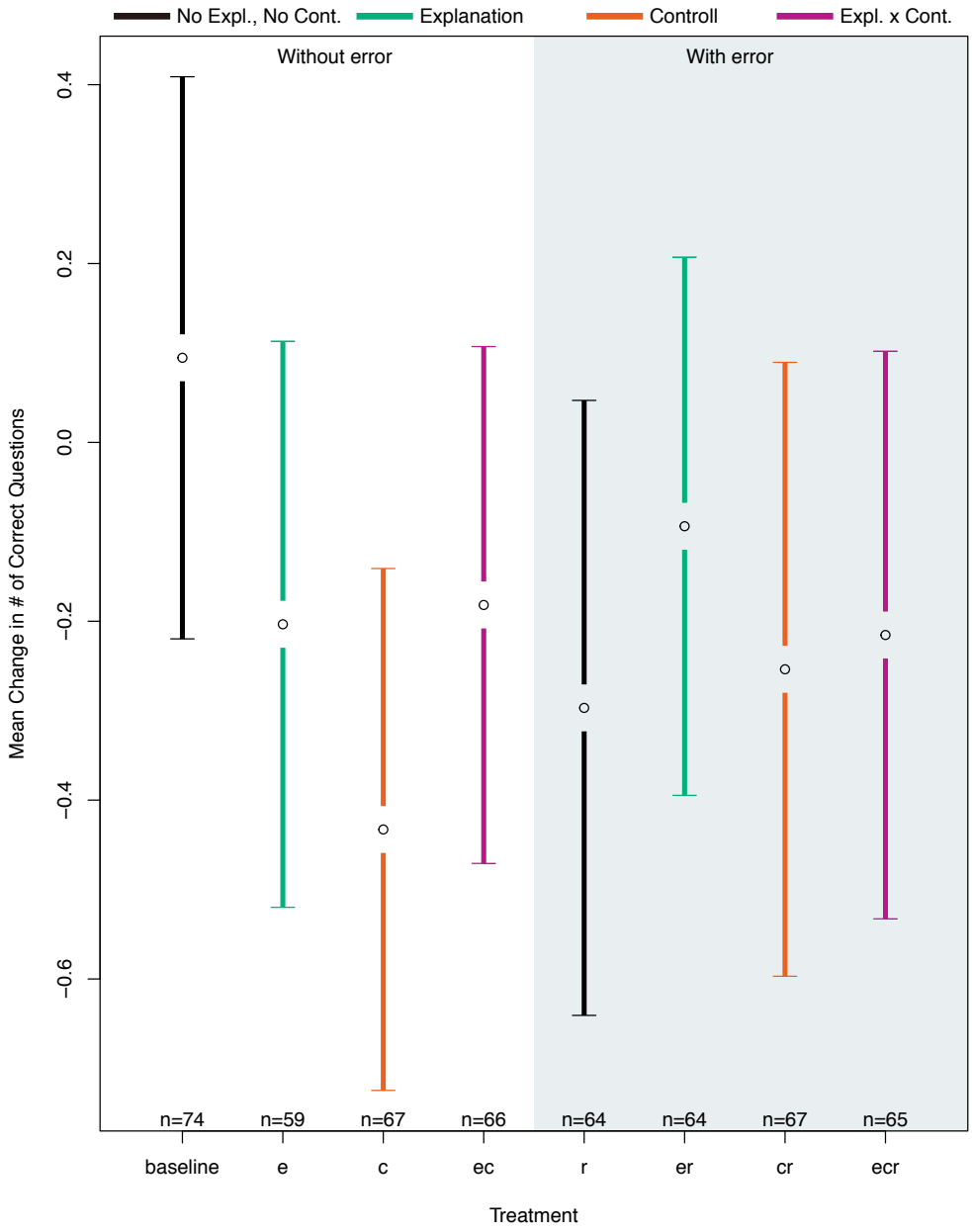


Figure 7.7: Mean change in insight between initial and final insight tests, broken up by treatment. Users in the baseline condition had the most positive change in mean. Error bars are 95% confidence interval.

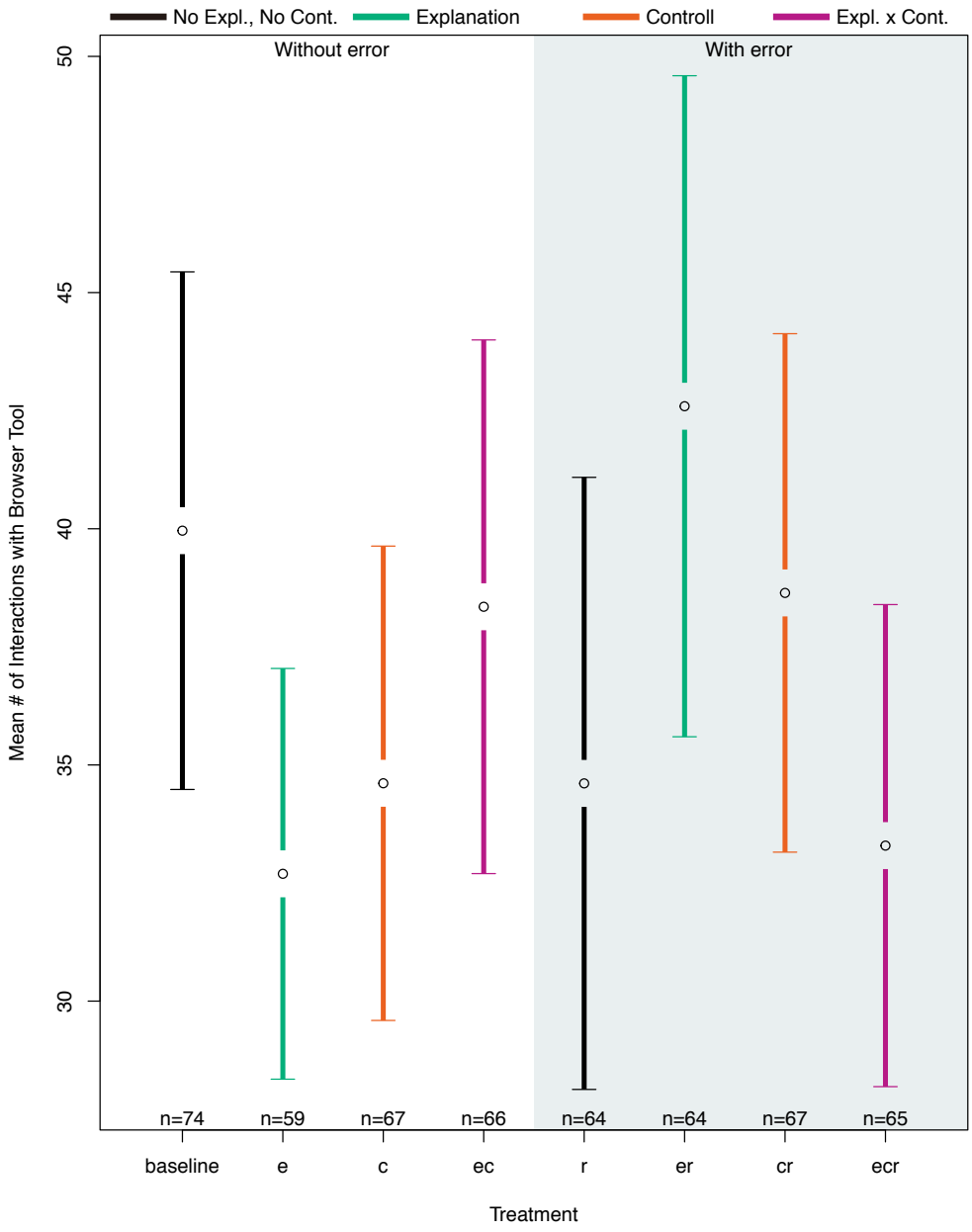


Figure 7.8: Quantity of interaction with the browser tool, broken up by treatment. Users in the baseline condition and users in the error condition with explanation had the highest mean interactions. Error bars are 95% confidence interval.

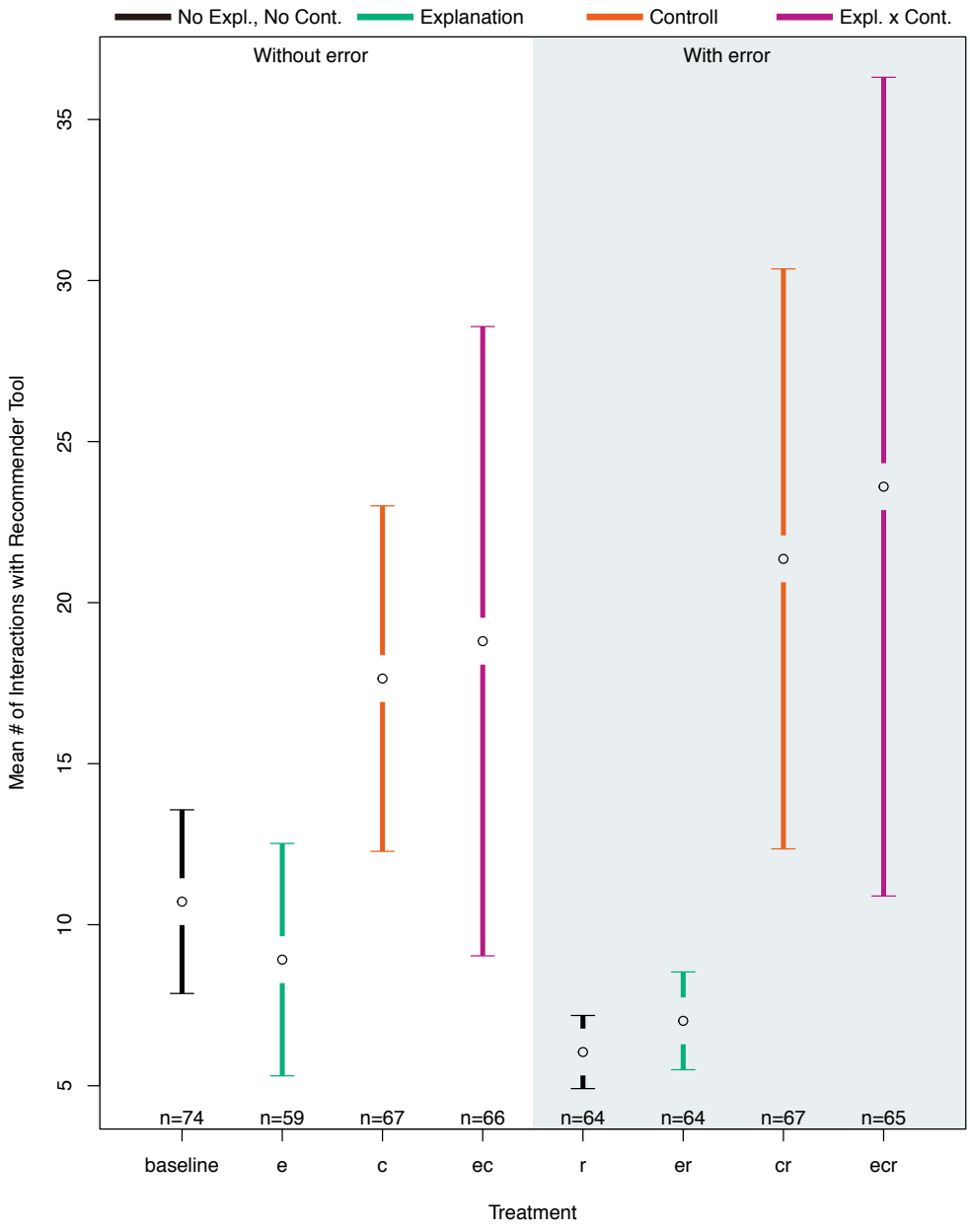


Figure 7.9: Quantity of interaction with the recommender tool, broken up by treatment. Users in control condition (unsurprisingly) interacted significantly more due to increased options for customization of the view. Error bars are 95% confidence interval.

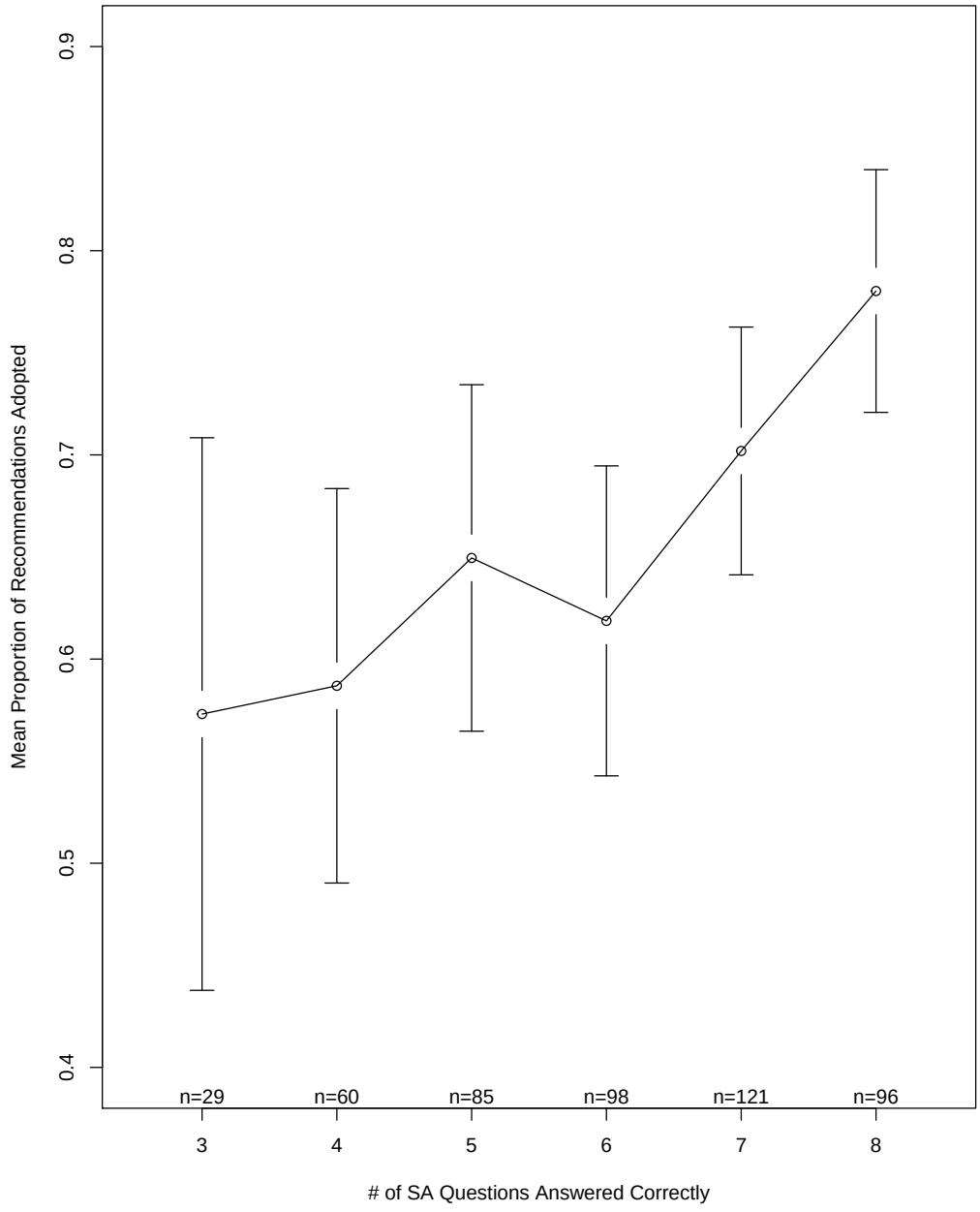


Figure 7.10: Relationship between adherence and understanding of the recommender. As understanding increases, users adopt more recommendations.

mendations is shown alongside a browser tool, but users should be encouraged to interact with the browser tool, not directly with the recommender. Explanations could be given to improve SAT. The data in this experiment suggests that this setup would maximize both adherence and decision satisfaction.

Finally, reductions in recommendation error had the largest impact on user experience but had no direct effect on decision satisfaction. Since an alternative to the recommender was available in this task, it is likely that users switched to the browsing tool when the recommender failed to produce satisfactory results. Our data also indicates that explanation and control have a bigger impact on the user's satisfaction with his/her final watchlist rather than recommender-related satisfaction and experience. More research where recommendation error is manipulated along with explanation and control would be needed to verify this finding.

7.5 Conclusion

We conducted a user study (N=526) on participants interacting with Movie Miner, –an interface that allowed users to choose between manual browsing and automated recommendation. Analysis of user cognitive metrics, observed participant behavior, task outcomes, and established user experience metrics revealed several key findings: 1) measurement of user beliefs about an algorithm, cognitive reflection, and insight increased the amount of decision satisfaction variance and adherence that could be explained by a statistical model, 2) SAT was the most significant variable affecting adherence to recommendations, 3) interacting with a recommender caused an insight reduction in users, but this can be mitigated by effecting higher situation awareness via system explanations, and 4) explanation, control, and recommendation error were all contributing factors to user acceptance of recommendations. This work is a step towards understanding user

cognition in recommender systems and we encourage other recommender researchers to adopt and improve the measurements described here.

Chapter 8

Decision Support in the Diner's Dilemma: Effects of Manipulating ECR

In this chapter, we will discuss the results of a study that manipulated the ECR profile of a decision support system for the Diner's Dilemma Here, users played the Diner's Dilemma game (introduced in Chapter 3, refer to that chapter for a complete description of the dilemma) with support from the "Dining Guru" - a decision support tool that recommended which choice to make in each round. Users made choices whether to access the Dining Guru and whether to take its advice. Results support the following claims:

- A simple, numeric-based explanation significantly increased adherence through a number of mediating factors as well as directly improved decision optimality by a significant amount.
- Requiring users to manually control the Dining Guru while providing explanations significantly decreased perceived control, trust, and correct beliefs about the Dining

Guru.

- High SA and low cognitive load were linked to correct beliefs about the game domain when the task ended

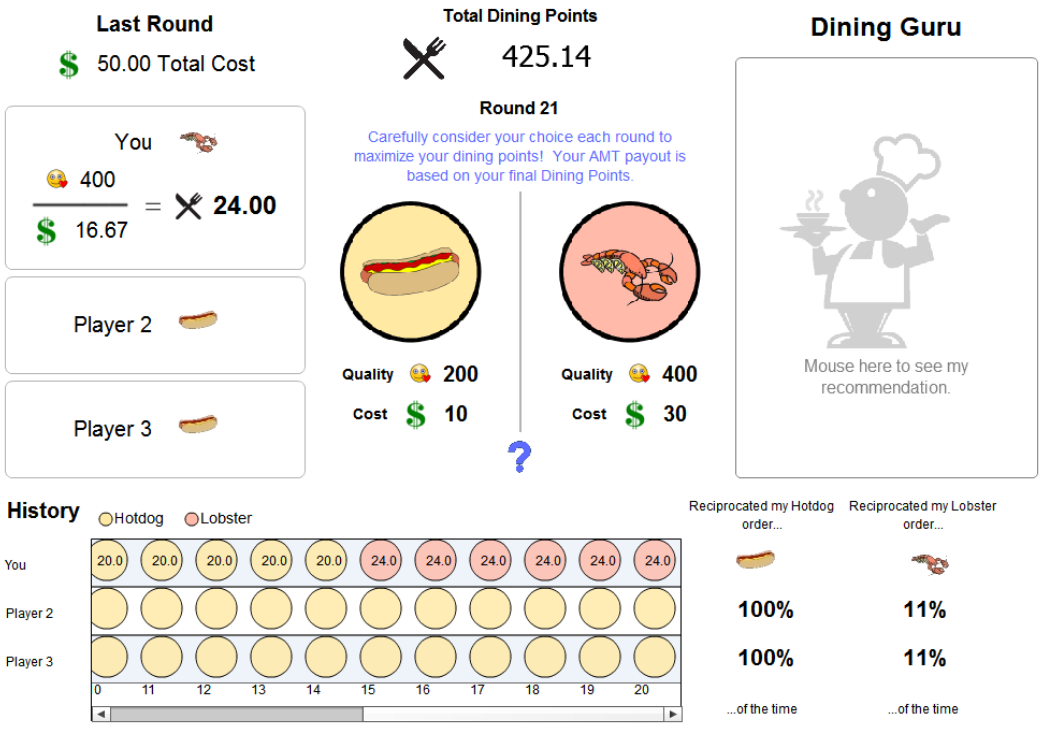


Figure 8.1: A screenshot of the Diner's Dilemma game with the Dining Guru. This is the same game as described in Chapter 3. Users were required to mouse over the Dining Guru to receive recommendations.

8.1 Game Design

Participants played the Diner's Dilemma game with the interface shown in Figure 8.1. This interface contains the level 1 and level 2 UI components from the experiment described in Chapter 3. Participants were also provided with a variation of the "Dining Guru," which took on several forms which are shown in Figure 8.2.

Participants played three different games against different co-diners. In the first game, co-diners betrayed often and the best strategy was to order lobster. In the second game, co-diners betrayed at a less rate and also forgave to some degree, which made hotdog the best choice. In the final game, co-diners were very forgiving and would rarely order lobster even when betrayed, which again made lobster the best choice. The second game was the most difficult to determine the optimal strategy due to a small difference between the expected value of ordering either item.

8.2 Generating Recommendations

The Dining Guru made a recommendation by calculating the expected value of ordering hotdog or lobster in the future, based on the maximum likelihood estimates of the true rate of forgiveness and betrayal from co-diners. This strategy caused the Dining Guru to make the "best possible" choice in each round, with most of the errors occurring in earlier rounds. When control was required from users, the Dining Guru took the hotdog and lobster reciprocity rates of opponents as input and only output a new recommendation when the sliders were changed.

8.3 User Interface Design

The Diner's Dilemma game was played through a similar interface as shown in Chapter 3.

8.3.1 Last Round Panel

All participants were shown their current dining points, the food quality and cost of each menu item, the current round, and the results from the previous round in terms of

dining points.

8.3.2 Game History Panel

On the lower portion of the screen a game history panel was provided. This panel contained all of the information that was necessary to generate an optimal strategy: information about who chose what in previous rounds and reciprocity rates.

8.3.3 Dining Guru

Users were required to mouse over the Dining Guru to access the information. This allowed us to understand when users were considering the Dining Guru's recommendation or using the alternative game history panel. The simple version of the Dining Guru calculated expected values and made a recommendation about which choice to make in each round. Adhering 100% to the Dining Guru's advice in this treatment would result in near optimal performance.

8.4 Experiment Design

This section describes the experimental methodology, including the study procedure, treatments, and dependent variables.

8.4.1 Independent Variables

Two levels of control, two levels of explanation, and three levels of recommendation error were manipulated. All manipulations (3 parameters, 2/3 values taken, $3 * 2^2 = 12$ manipulations) were used as between-subjects treatments in this experiment. Note that since we are testing the effects of recommender configuration rather than the effects

of recommender presence, this experiment's "baseline condition" corresponded to the treatment where control, explanation, and noise are absent.

Control Manipulation

In the control treatments, the Dining Guru did not update until prompted by the user, who was required to provide estimates of reciprocity rates from co-diners. Users could freely experiment with the sliders, which means that they were able to completely understand the data space if they chose to do so.

Explanation Manipulation

In the explanation treatments, the Dining Guru provided the estimate for the expected points per round of each choice along with the recommendation. This appeared as a text number for each item as well as two bars so participants could directly compare the values at a glance.

Error Manipulation

Three levels of error were manipulated. In the no-error treatment, the Dining Guru produced correct recommendations, which if followed would result in almost complete optimality in game performance. The first level of error (weak error) would randomly adjust the reciprocity estimates up or down by up to 25%. For instance, if the true hot-dog reciprocity rate was 65%, the Dining Guru would use a value anywhere between 40 and 90%. This would occasionally cause the Dining Guru to "flip-flop" between recommendations. However, if followed, recommendations in the weak error condition would still result in relatively high performance. Finally, the "full" error condition adjusted reciprocity estimates by up to 50% in either direction. A practical consequence of this is that the Dining Guru would flip it's recommendation almost every round.

Game	Opt. Choice	Rounds	DG/No Err	DG/Weak Err	DG/Err	Users
1	Lobster	55	92.3%	75.2%	63.5%	70.7%
2	Hotdog	60	65.5%	62.8%	56.1%	46.5%
3	Lobster	58	79.1%	69.3%	60.6%	62.4%
All		173	78.6%	68.9%	60.0%	59.1%

Table 8.1: Performance of the Dining Guru across all games, compared with participants. Game 1 was the easiest, followed by game 3 and 2. On average, participants performed the same as the Dining Guru with Error.

The error in the recommendations was reasonably hidden from participants, as only the estimates for Hotdog/Lobster were adjusted, and the error values would only change between rounds when the Dining Guru was hidden (only one participants remarked in comments that he/she noticed something wrong with the Dining Guru). Explanation and control made the error more easy to detect by exposing the expected values for hotdog and lobster. If participants only accessed the Dining Guru occasionally, it might be difficult to notice. However, if a participant was checking the Guru every few rounds, explanation and control features would make these changes more detectable.

Table 8.1 shows a breakdown of how the Dining Guru performed in each game. The optimality percentage for the Dining Guru without error represents the "best possible rational strategy," in that the only way to do better than this strategy would be to know what the simulated co-diners were doing in advance or to "guess" and be lucky. Overall, the Dining Guru with full error did about the same as the participants on average. This means that players with a good understanding of the game were better off not adhering to the Dining Guru's recommendation.

8.4.2 Dependent Variables

Dependent variables consisted of observations of system interactions and more complex factors, which were collected through questionnaires.

Basic dependent variables measured in this study was the quantity of interaction

Factor	Item Description	R^2	Est.
Trust Prop.	Based on past experience, I think I would trust an AI adviser if one were available.	0.80	1.12
$ALPHA$: 0.91	Based on past experience, I think I would be satisfied if I adhered to advice from an AI adviser.	0.73	1.07
AVE : 0.51	Based on past experience, I think AI advisers give accurate information.	0.67	1.01
	I trust automation and AI in general.	0.65	1.03
Game Exp.	I am familiar with abstract trust games.	0.53	1.28
$ALPHA$: 0.80	I am familiar with the Diner's Dilemma.	0.48	1.16
AVE : 0.36	I am familiar with the public goods game.	0.75	1.58
Cog. Reflection	If it takes 5 machines 5 minutes to make 5 widgets...	0.54	0.33
$ALPHA$: 0.73	A bat and ball together cost $1.10, and the bat costs $1.00 more than the ball...	0.30	0.27
AVE : 0.54	In a pond there is a patch of lily pads that doubles in size every day...	0.60	0.34

Table 8.2: Factors determined by participant responses to subjective questions. R^2 reports the fit of the item to the factor. Est. is the estimated loading of the item to the factor. Items that were removed due to poor fit are not shown.

and accesses of the Dining Guru, the total percentage of rounds where optimal decisions were made, and adherence to recommendations. Decision optimality was calculated as the percentage of rounds where an optimal choice was made. An optimality score of 1 indicates the player ordered 100% lobster in games 1 and 3, and 100% hotdog in game 2. Adherence occurred for each round where the user choice matched the last recommendation given by the Dining Guru. The final adherence measurement was scaled between 0 and 1, where 0 indicates no adherence and 1 indicates complete adherence. Some users never accessed the Dining Guru, which caused their adherence score to become 0.

For the more complex dependent variables, confirmatory factor analysis (CFA) was used to eliminate measurement error when possible. Structural equation modeling (SEM) was then used to test the relationships between the confirmed factors in our HAI model. A list of the subjective factors is shown in Tables 8.3 and 8.2. All of these items were

Factor	Item Description	R^2	Est.
Trust	I trusted the Dining Guru.	0.78	1.35
$ALPHA$: 0.90	I could rely on the Dining Guru.	0.75	1.30
AVE : 0.58	I would advise a friend to take advice from the Dining Guru if they played the game.	0.74	1.36
Cognitive Load	It was hard to keep track of all of the information needed to play the game.	0.45	1.19
$ALPHA$: 0.75	I got lost while playing the game.	0.49	1.15
AVE : 0.26	I got frustrated during the game.	0.58	1.3
Perceived Control	I had control over the Dining Guru.	0.71	1.33
$ALPHA$: 0.96	I could affect what the Dining Guru recommended.	0.75	1.02
AVE : 0.34	I had no control over the Dining Guru.	0.74	-1.2

Table 8.3: Factors determined by participant responses to subjective questions. R^2 reports the fit of the item to the factor. Est. is the estimated loading of the item to the factor. Items that were removed due to poor fit are not shown.

taken on a Likert scale except for cognitive reflection, where free response was used.

In this study, we included items for perceived transparency, control, effectiveness, and trust. However, during the confirmatory factor analysis, inter-item correlations indicated we only had two factors: perceived control and trust. Items for perceived transparency, effectiveness, and trust all correlated strongly, so we collapsed this factor and used only the three items for trust. The items for perceived control had a "low" (0.2-0.4) correlation with items for trust and the other collapsed factors, indicating good discriminant validity.

We used a SAGAT-style freeze [144] during game 2 to assess situation-awareness based agent transparency (SAT), which showed 7 questions related to ground truth of recommender behavior. This questionnaire was originally 10 questions, but 3 questions were very similar to questions on our insight metric. When they showed high inter-correlations during the exploratory factor analysis, they were discarded. Insight was measured twice - just after the training and just after game 3. This test was a set of eleven questions which relates to knowledge of the game and whether the participant possessed the ability to map from the current game state to the optimal decision. We

Factor	Item Description
SAT all-item parcel	1. The Dining Guru updates automatically every round (T/F) 2. When the Dining Guru is updated, it predicts the choice I should make in the next round (T/F) 3. When the Dining Guru is updated, it predicts the choice I should make in all remaining rounds (T/F) 4. When does the Dining Guru recommend Hotdog? 5. How does the accuracy of the Dining Guru change as the game progresses? 6. Generally, I can maximize the Dining Points I get per round by ordering a mix of Hotdog and Lobster, regardless of what the Dining Guru recommends (T/F) 7. Generally, I can maximize the Dining Points I get per round by only ordering what the Dining Guru recommends (T/F)

Table 8.4: Situation-awareness based agent transparency (SAT) questions designed to measure understanding of the Dining Guru. Multiple choice answers were given. SAT was measured after game 2 and before game 3.

used all-item parcels for domain knowledge and SAT and the variance was fixed to the variance of the sample population. The SAT and insight metrics are shown in Tables 8.4 and 8.5, respectively.

8.4.3 Procedure

Participants were recruited on Amazon Mechanical Turk (AMT).

Participants made their way through four phases: the pre-study, training, the game phase, and the post-study.

The pre-study and post-study were designed using Qualtrics[1]. Items related to trust propensity, game expertise, and cognitive reflection (Toplak et al [69]) were collected during this phase using the question items shown in Table 8.2.

Next, in the "training phase," participants accessed the Diner's Dilemma game and were introduced to game concepts and the Dining Guru. Several testing questionnaires, which could be resubmitted as many times as needed, were used to help participants

[1]https://www.qualtrics.com/

Factor	Item Description
Insight all-item parcel	1. How much does a hotdog cost? (slider response) 2. How much does a lobster cost? (slider response) 3. What is the quality of a hotdog? (slider response) 4. What is the quality of a lobster? (slider response) 5. In a one-round Diner's Dilemma game (only one restaurant visit), you get the least amount of dining points when... (four options) 6. In a one-round Diner's Dilemma game (only one restaurant visit), you get the most amount of dining points when... (four options) 7. Which situation gets you more points? (two options) 8. Which situation gets you more points? (two options) 9. Suppose you know for sure that your co-diners reciprocate your Hotdog order 100% of the time and reciprocate your Lobster order 100% of the time. Which should you order for the rest of the game? (H/L) 10. Suppose you know for sure that your co-diners reciprocate your Hotdog order 0% of the time and reciprocate your Lobster order 100% of the time. Which should you order for the rest of the game? (H/L) 11. Suppose you know for sure that your co-diners reciprocate your Hotdog order 50% of the time and reciprocate your Lobster order 50% of the time. Which should you order for the rest of the game? (H/L)

Table 8.5: Insight was measured after the training protocol (initial insight) and was measured again after the third game to see if any changes had occurred as a result of interacting with the Dining Guru (final insight).

learn the game. The Dining Guru was introduced as an "AI adviser" and participants learned its basic parameters, including how to access it and what it was recommending. Participants were told that the Dining Guru was not guaranteed to recommend optimal decisions and whether they adhered to the advice was up to them. The initial insight test was taken just after training was completed.

In the "game phase," participants played three games of Diner's Dilemma against three configurations of simulated co-diners. SAT metrics were taken after game 2, and the final insight test was taken after game 3.

Finally, subjective metrics related to cognitive load, trust, and perceived control were collected during the post-study

Variable Name	**Description**	μ	σ
Recommender Int.	Number of times the Dining Guru was accessed	25	34
Adherence	Number of choices that matched the recommendation from the Dining Guru	0.33	0.25
Decision Optimality	Percentage of moves that were optimal	0.59	0.12
Initial Insight	Score on initial insight questionnaire (out of 11)	8.08	1.64
Final Insight	Score on final insight questionnaire (out of 11)	8.5	1.67
SAT	Score on recommender beliefs questionnaire (out of 7)	3.82	1.23

Table 8.6: Observed dependent variables in the study.

8.5 Results

We collected more than 529 samples of participant data using AMT. Participants were paid $3.00 and spent between 30 and 50 minutes doing the study. Participants were between 18 and 70 years of age and were 54% male. Participant data was checked carefully for satisficing and these records were removed, resulting in the 529 complete records.

8.5.1 All-Factor SEM

Regressions in the SEM and fit of the all-factor model are shown in Table 7.5. Model co-variances are shown in Table 7.6. An illustration of the SEM was omitted due to visual complexity. Model fit: $N = 529$ with 99 free parameters = 5 participants per free parameter, $RMSEA = 0.047$ $(CI : [0.043, 0.051])$, $TLI = 0.92, CFI = 0.91$ over null baseline model, $\chi^2(391) = 848.832$. The model was built using R 3.0.3, lavaan 0.5-17.

8.5.2 Raykov Change Model

A Raykov structural model [152] was built to identify factors that predict changed belief between the initial and final insight tests. The final model is shown in Figures

Regressand	Regression (←)	Coeff.	P(> |z|)
Trust R^2=**0.17**	← Trust Propensity	0.35	***
	← Game Expertise	0.12	*
	← Explanation	0.24	*
	← Expl. x Control	-0.21	0.07
Perceived Control R^2=**0.27**	← Game Expertise	0.11	*
	← Explanation	0.38	**
	← Control	1.22	***
	← Expl. x Control	-0.49	**
Cognitive Load R^2=**0.08**	← Initial Insight	-0.24	***
	← Expl. x Control	0.34	**
Recommender Int. R^2=**0.16**	← Game Expertise	-0.21	***
	← Trust	0.23	***
	← Control	-0.28	*
	← Error	0.41	***
	← Control x Error	-0.44	*
SAT R^2=**0.05**	← Initial Insight	0.11	**
	← Explanation	0.31	**
	← Expl. x Control	-0.54	***
Adherence R^2=**0.51**	← Recommender Int.	0.64	***
	← SAT	0.12	***
	← Game Expertise	-0.17	***
	← Perceived Control	0.11	***
	← Cognitive Load	-0.12	***
	← Error	-0.24	***
Final Insight R^2=**0.19**	← Cognitive Reflection	0.27	***
	← Game Expertise	-0.18	***
	← Cognitive Load	-0.23	***
	← SAT	0.08	0.06
Decision Optimality R^2=**0.18**	← Recommender Int.	0.13	***
	← Initial Insight	0.22	***
	← SAT	0.27	***
	← Explanation	0.28	**
	← Expl. x Error	-0.38	**

Table 8.7: Regressions in the fitted all-factor SEM, which attempts to explain decision satisfaction, final insight, and adherence to recommendations. Variables on the left are explained by variables on the right. Due to being non-normal, treatment variables take the value of 0 or 1 and coefficients reported are B values, which predict a change in standard deviation of the regressand when the treatment is switched on. All dependent and latent variables were standardized to have a mean of 0 and variance 1 and coefficients reported are β values, which predict a change in standard deviation of the regressand with a standard deviation change in the regressor. Significance levels for this table: *** $p < .001$, ** $p < .01$, * $p < .05$. Covariances are shown in Table 7.6.

Covariance (↔)	**Beta**	**P(> |z|)**
Perceived Control ↔ Trust	0.30	***
Trust ↔ Cognitive Load	-0.19	***
Trust ↔ SAT	0.14	**
CRT ↔ Game Expertise	-0.12	*
CRT ↔ Initial Insight	0.36	***
Trust Propensity ↔ Game Expertise	0.30	***
Trust Propensity ↔ Initial Insight	-0.09	*
Game Expertise ↔ Initial Insight	-0.25	***
Adherence ↔ Decision Optimality	0.20	***
Adherence ↔ Final Insight	0.08	0.07

Table 8.8: Covariances in the fitted model. Regressions are shown in Table 7.5. Trust and cognitive load were negatively correlated. All task outcomes (decision optimality, adherence, and final insight) positively co-varied.

8.3 and 8.4. In each graph, the y-axis indicates the fitted, standardized beta parameter (mean 0, variance 1) for insight. The x-axis shows the difference between the initial and final tests. The slope indicates the rate at which beliefs became more correct or incorrect.

Figure 8.3 indicates that participants with correct beliefs about the Dining Guru were more likely to have higher initial insight but also formed incorrect beliefs during the task. Participants that adhered to the Dining Guru more often also started with higher initial insight, but learned more during the course of the game. Participants that perceived they had more control were more likely to start with lower initial insight but learned more during the game.

Figure 8.4 predicted that explanation and control caused incorrect beliefs to form. Users in the "Explanation x Control" treatment formed the most incorrect beliefs during the task.

8.5.3 False Discovery Rate Analysis

Multiplicity control was enforced in our chosen SEM using the Benjamini-Hochberg procedure with $Q = 0.10$ [153], which is recommended for exploratory SEM analysis [154].

This procedure indicates how many of the tested relationships in the All-factor SEM are expected to be false positives. All reported effects in Tables 7.5 and 7.6 passed the FDR, with the most minor finding ($Trust \leftarrow ExplanationxControl$) having $p = 0.074$ and $(i/m)Q = 0.10$.

8.6 Discussion of Results

Here we discuss four implications of the results of this work. The results discussed here are specific to this task domain. A meta-analysis of all studies in this dissertation is given in Chapter 9.

1. Explanation indirectly causes increased adherence through a number of mediating factors. The complex visual in Figure 8.5 identifies all factors that directly affect or indirectly (through full/partial mediation) affect adherence. Explanation directly increases trust, SAT, and perceived control. SAT and perceived control both directly increase adherence. Additionally, refer to Figure 8.6 for the relationship between SAT and adherence. Trust indirectly increases adherence through full mediation by recommender interaction. It should be noted that, in this study, increased adherence due to recommender interaction is not surprising (users could not adhere to recommendations if they were not viewed!) Note also that while control had a significant positive influence on perceived control, it also caused users to access recommendations less - this indicates that the control feature did not have a noteworthy influence on adherence. Finally, when both explanation and control were shown, the positive effects of explanation almost disappear while cognitive load increases, indicating decreased adherence.

The mean number of times the Dining Guru was accessed in each treatment is shown in Figure 8.7. Error-prone recommendations prompted more interaction, due to a "flip-flop" behavior from the Dining Guru. It can be observed that users in the "control"

conditions did not exhibit increased access behavior, likely due to the increased effort it took to get a recommendation from the Dining Guru. Finally, note that while the error treatment significantly increased the number of times the recommendation was accessed, it had a negative effect on adherence (shown in Figure 8.5).

The results from this model indicate many positive benefits of incorporating explanations into decision support systems. Previously, explanations have been noted to increase trust [156], adherence [21], and perceived control [128]. This strengthens the validity of the new results found in this study: that explanations increase the number of correct beliefs about the remember, and that predicts an increase in adherence. Additionally, we have observed many important interaction effects between explanation, control, and recommendation error. The results suggest that in some situations, explanation and control given together may overload users of the system. Furthermore, control features may draw attention to flaws in the system predictions due to increased interaction. In some situations, this may lead to better decision making due to the rejection of incorrect predictions. However, making flaws more difficult to detect by withholding control features will improve adherence for those concerned.

2. High SAT and low cognitive load were linked with correct beliefs about the game when the task ended. Figure 8.8 shows a visualization of factors that directly or indirectly affect final insight. Explanation directly increased SAT. Additionally, when control and explanation were both present, SAT significantly decreased and cognitive load significantly increased, both of which predicted lower final insight. Additionally, the mean change in insight by treatment is shown in Figure 8.9. The mean for participants in the baseline condition (no explanation, control, or error) was the highest, followed by those that were just treated to weak noise.

Changes to user beliefs about the game can be explained with four points. First, the Raykov model (Figures 8.3 and 8.4) indicated that drawing attention to the Dining

Guru through explanations or control caused the formation of incorrect beliefs during the game. Second, the Raykov model also indicated that users with more correct beliefs (SAT) about the Dining Guru formed incorrect beliefs about the game, however, users that adhered to recommendations learned more. Finally, recall from the previous section that explanations increased trust (which co-varied with cognitive load), perceived control, and SAT, all of which were linked to adherence. These three points together suggest that users in the explanation conditions may have over-trusted the Dining Guru and relied on the recommendations rather than engaging with the game, however, users in the baseline condition were able to learn from the Dining Guru by carefully considering its advice. Due to fewer entities on the screen, these users would have also had more mental bandwidth with which to figure the game out for themselves.

3. Users that considered themselves experts were much less likely to interact with the Dining Guru and adhere to recommendations. These users knew less about the game and performed worse. Figure 8.10 shows factors directly or indirectly affecting decision optimality. From the visual, it can be seen that users that considered themselves experts were were also likely to have higher trust propensity and significantly lower insight on the initial test. These users reported more trust than average with the Dining Guru but less interaction and adherence (see Figure 8.5 as well). For these types of users, explanations would be appropriate, as they notably increase trust, SAT, and have a direct effect on decision optimality.

Meanwhile, less trusting users had higher initial insight, which predicted more correct beliefs about the recommender. These users performed well whether explanations were presented or not.

When both explanations and error were present, the model predicts that decision optimality drops below the mean. This can also be seen in Figure 8.11. This indicates that explanations allowed users to better detect the errors in the Dining Guru, which

may have steered them away from use in the noisiest conditions. However, adhering to the Dining Guru in even the noisiest condition would have put the user's performance at the mean, and adhering in the weak noise conditions would have put the users well above the mean (recall Table 8.1. This result implies that even relatively accurate decision support systems can be ignored if users are able to detect errors in the algorithm.

8.7 Conclusion

We conducted a user study (N=529) on participants playing the Diner's Dilemma game with help from an AI adviser – the Dining Guru. Analysis of user cognitive metrics, observed participant behavior, task outcomes, and user experience metrics revealed several key findings: 1) explanation significantly increased adherence through a number of mediating factors as well as directly improved decision optimality by a significant amount, 2) Requiring users to manually control the Dining Guru while providing explanations significantly decreased perceived control, trust, and correct beliefs about the Dining Guru, and 3) high SA and low cognitive load were linked to correct beliefs about the game domain when the task ended. The abstract nature of this task gives the results a wide impact on repeated decision-making contexts.

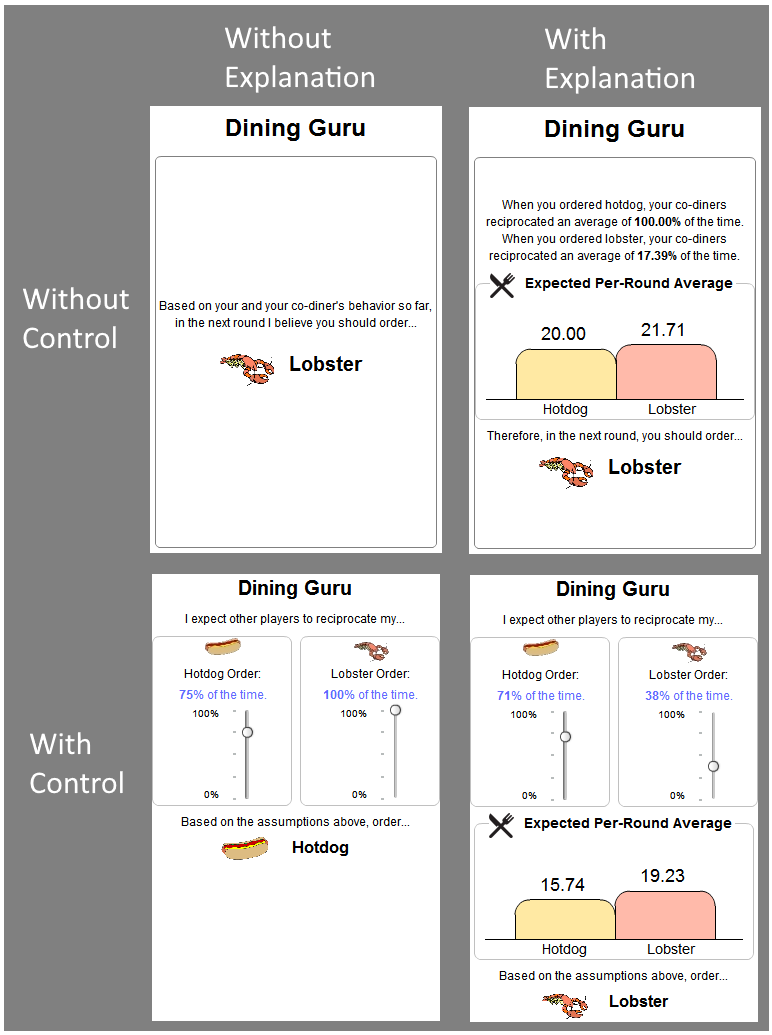


Figure 8.2: Variations in the ECR profile of the Dining Guru. The "Control" version of the Dining Guru is similar to the Level 3 Tool which was provided to participants in the experiment described in Chapter 3. Explanation came in the form of exposing the expected values of each choice to participants, which is the basis for the recommendation.

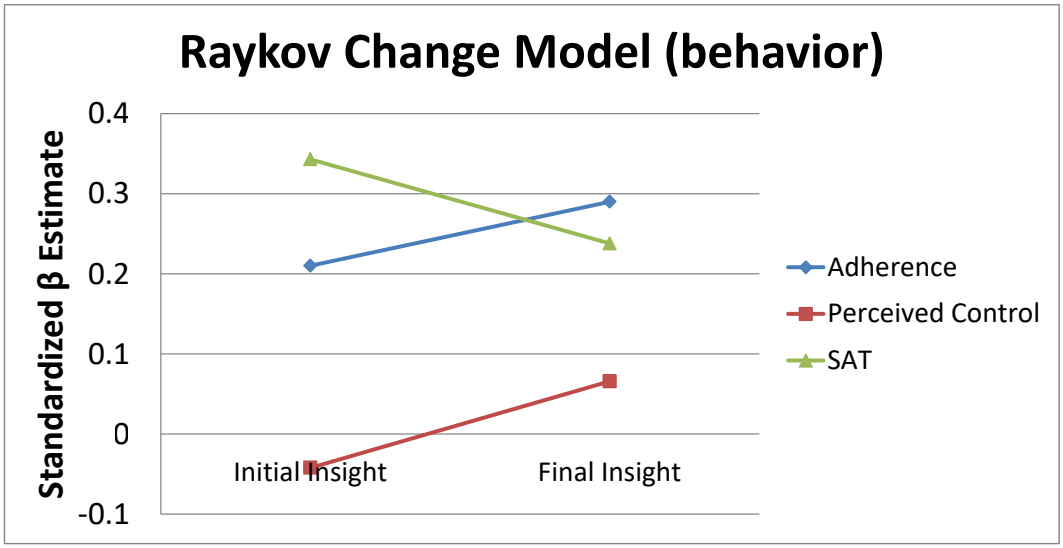


Figure 8.3: A Raykov model explaining the rate at which beliefs about the data changed between the initial and final insight tests as predicted by user cognition and personal attributes.

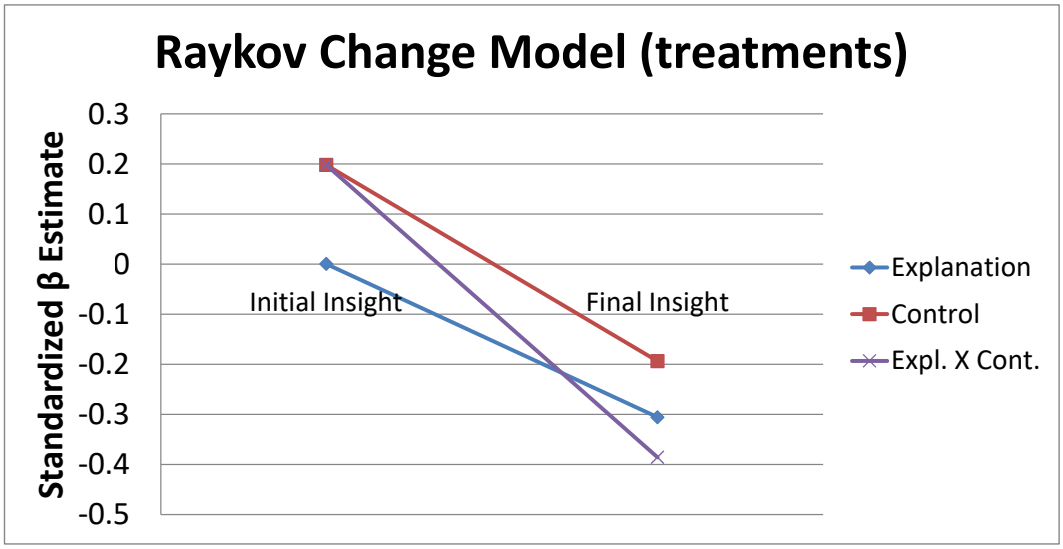


Figure 8.4: A Raykov model explaining the rate at which beliefs about the data changed between the initial and final insight tests as predicted by treatments

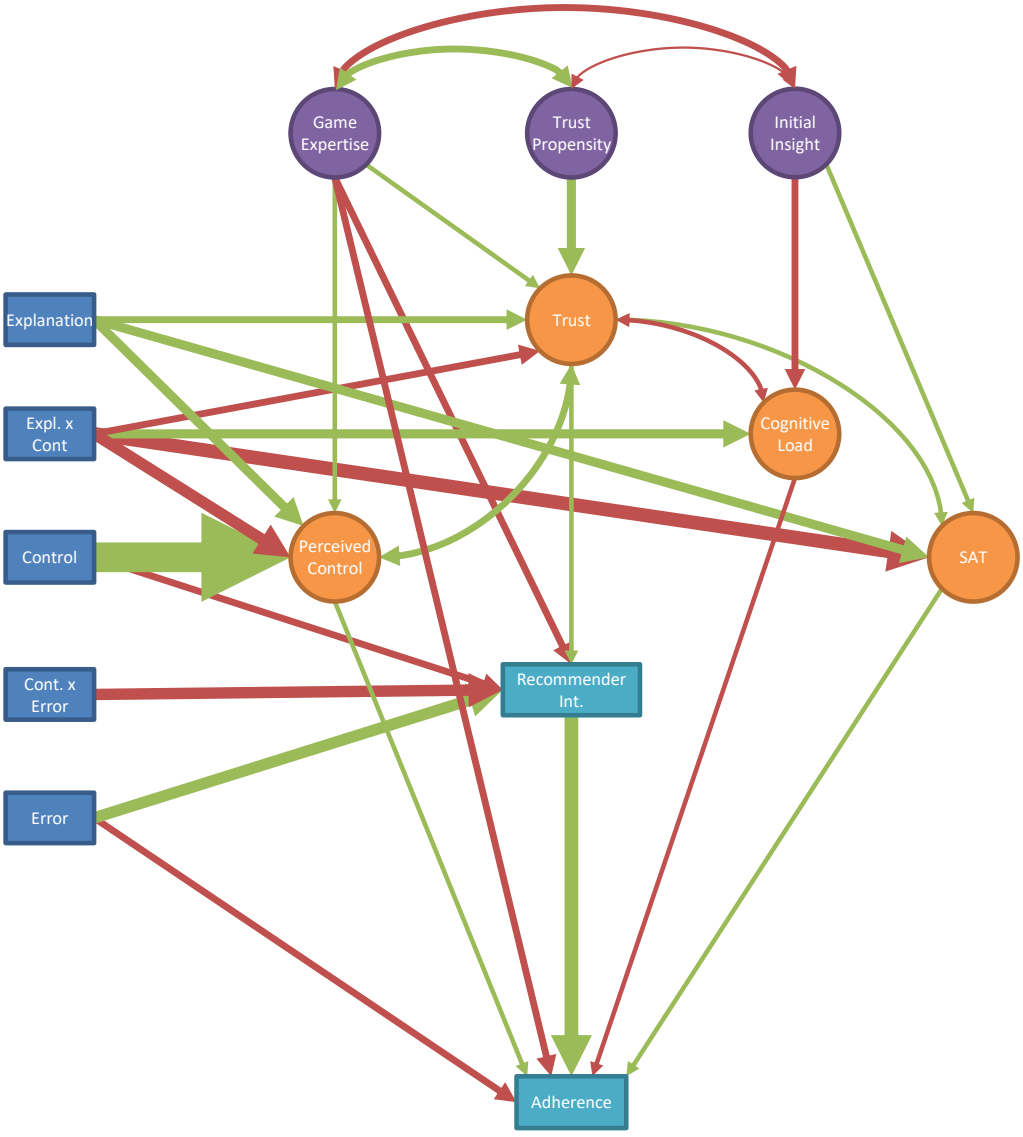


Figure 8.5: A visualization of all factors that directly or indirectly affect decision satisfaction. Line thickness indicates magnitude of β or B parameter. Green lines indicate a positive effect. Red lines indicate a negative effect. Note that this is just a visualization of a portion of the model reported in Table 8.7, and does not represent a tested statistical model.

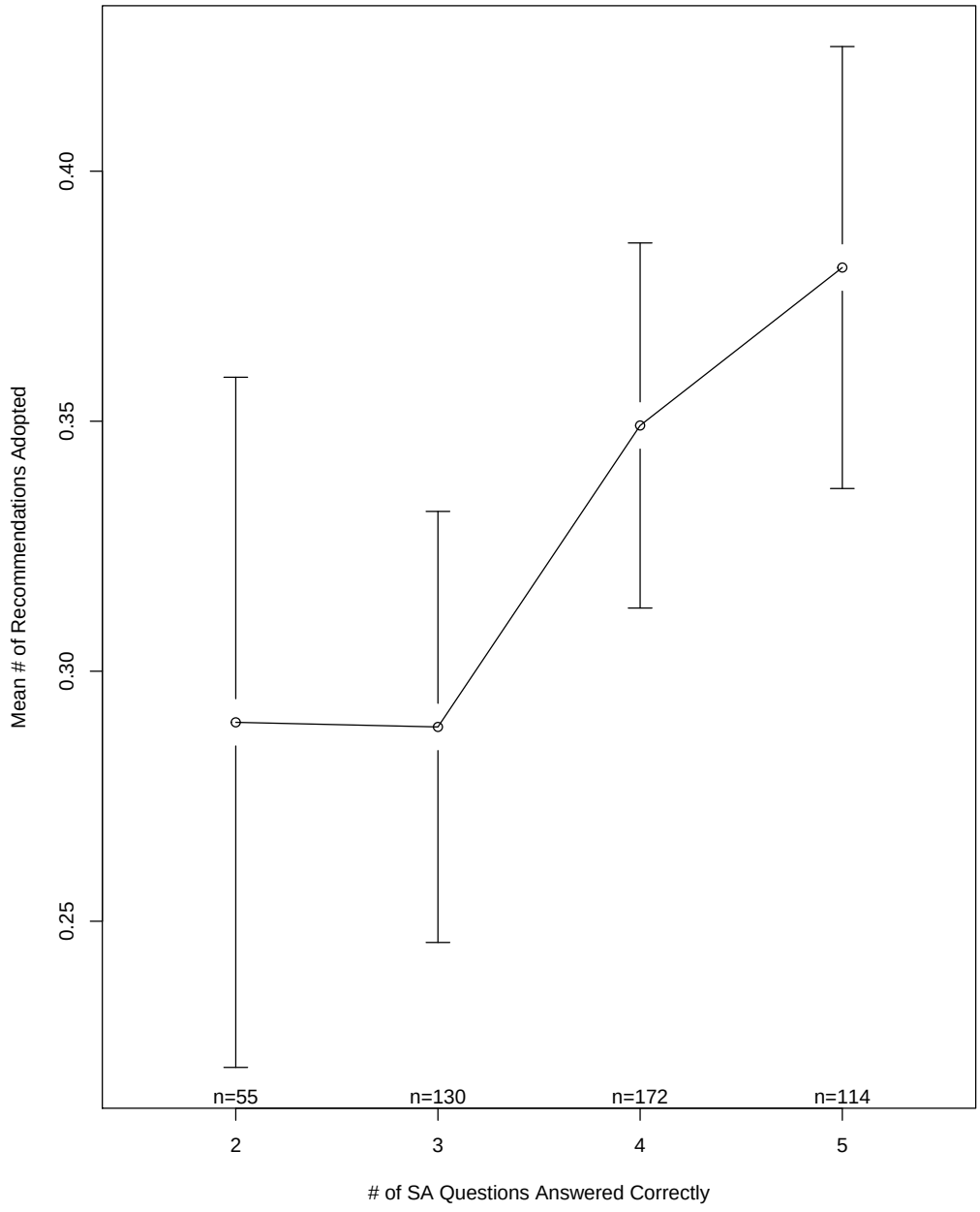


Figure 8.6: Relationship between adherence and understanding of the recommender. As understanding increases, users adopt more recommendations.

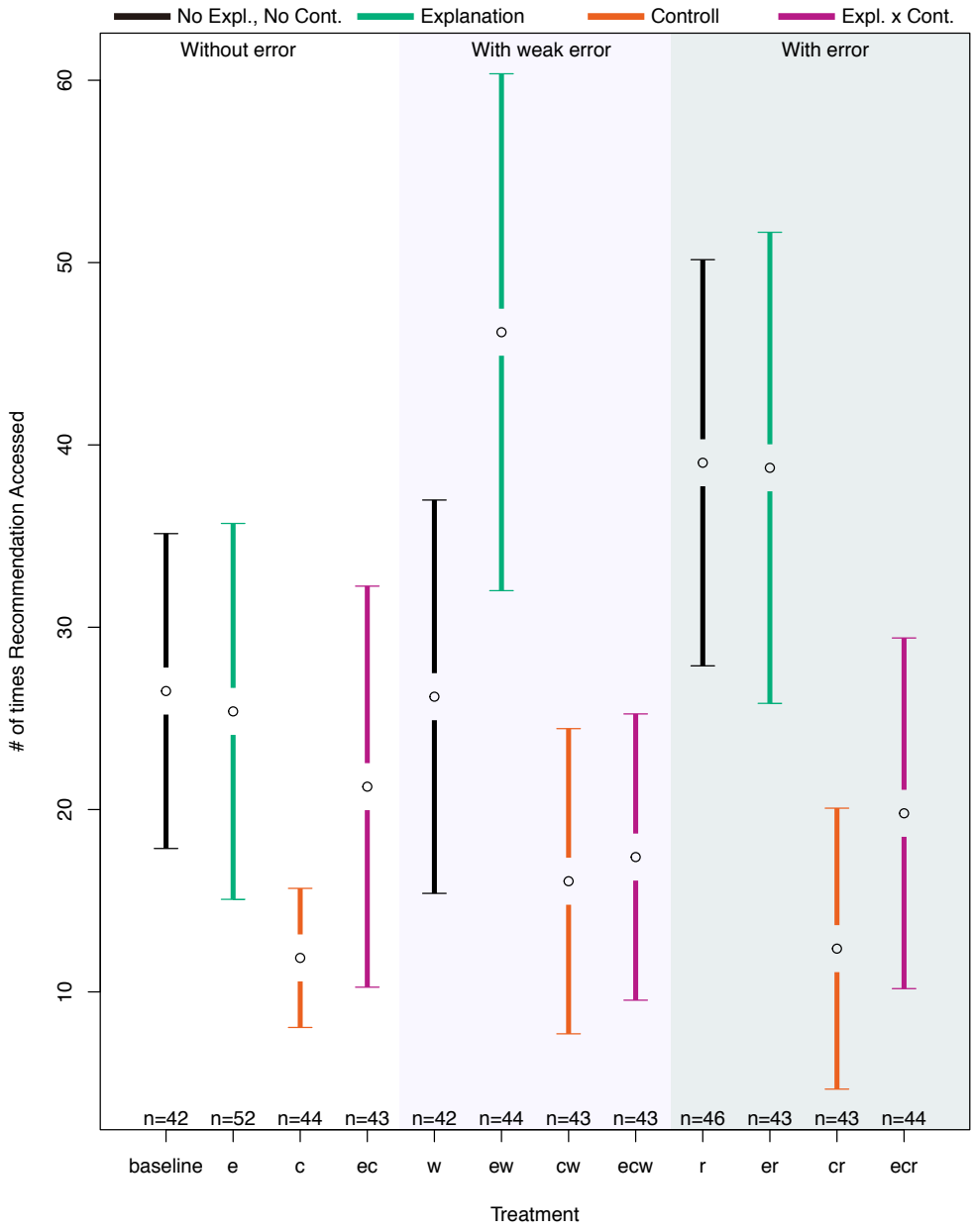


Figure 8.7: Quantity of accesses with the Dining Guru by treatment. Access significantly increases when noise is present but control is not. These users were likely responding to the Dining Guru's tendency to "flip-flop" between Hotdog and Lobster when noise was present. The changes were more noticeable when explanations were provided. Error bars are 95% confidence interval.

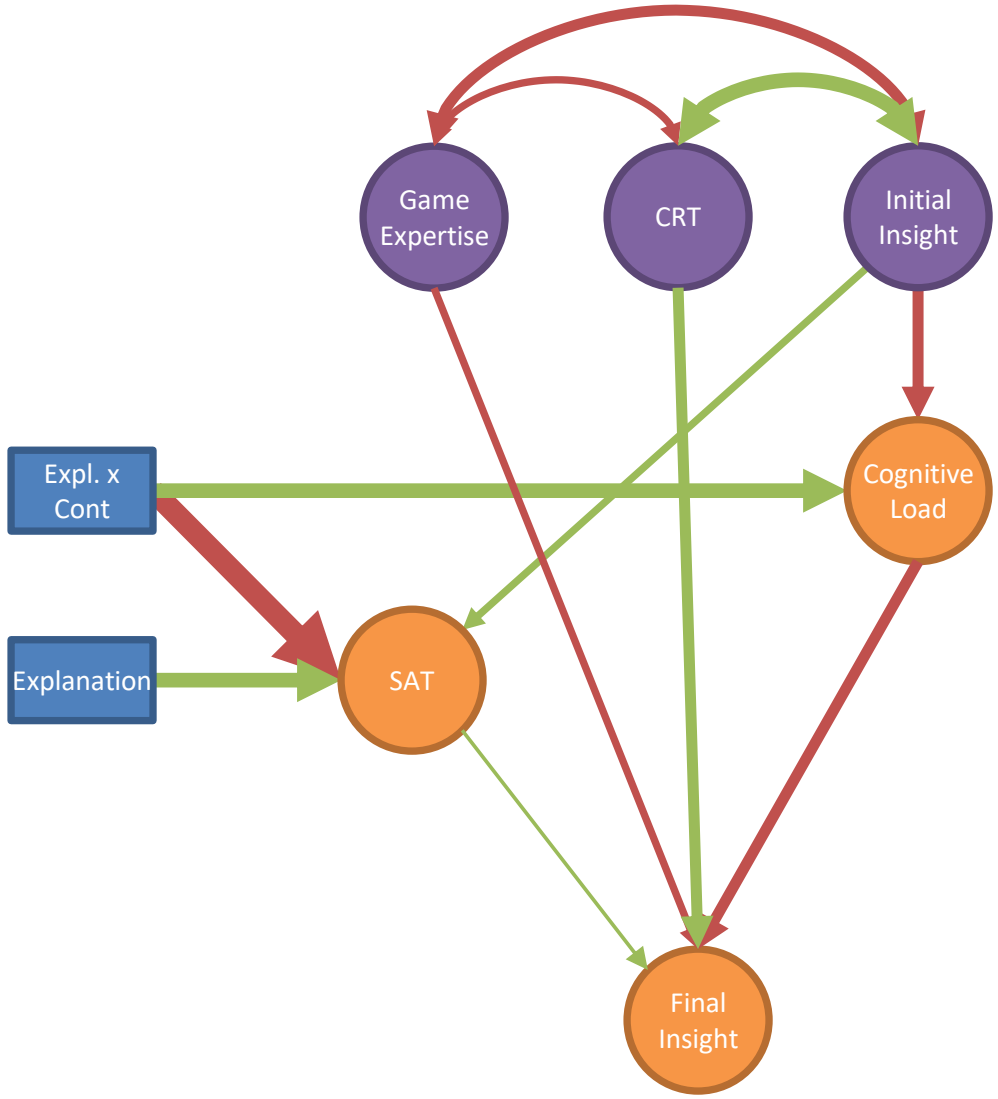


Figure 8.8: A visualization of all factors that directly or indirectly affect final insight. Line thickness indicates magnitude of β or B parameter. Green lines indicate a positive effect. Red lines indicate a negative effect. Note that this is just a visualization of a portion of the model reported in Table 8.7, and does not represent a tested statistical model.

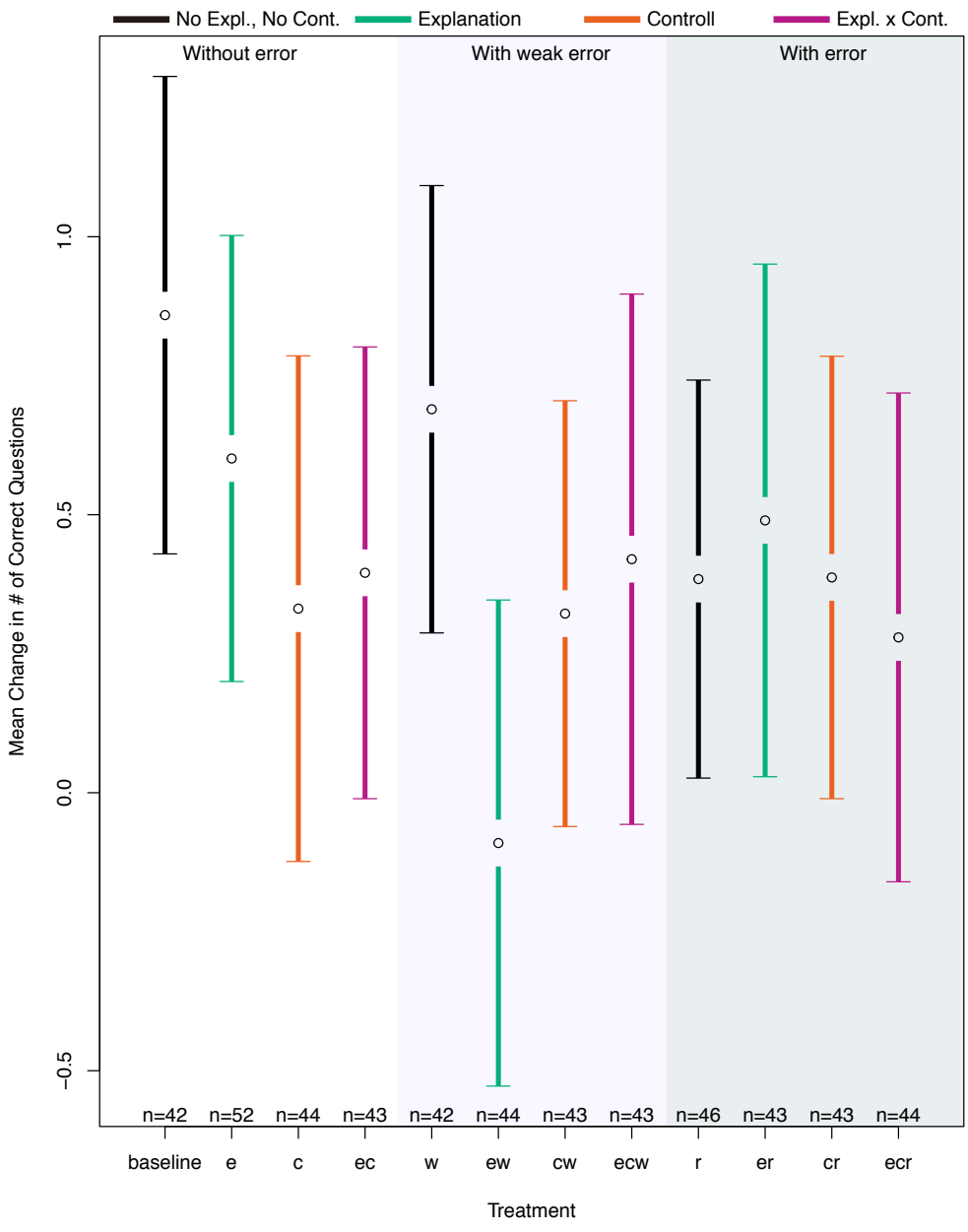


Figure 8.9: Mean change in the number of correct questions between the initial and final insight tests. Error bars are 95% confidence interval.

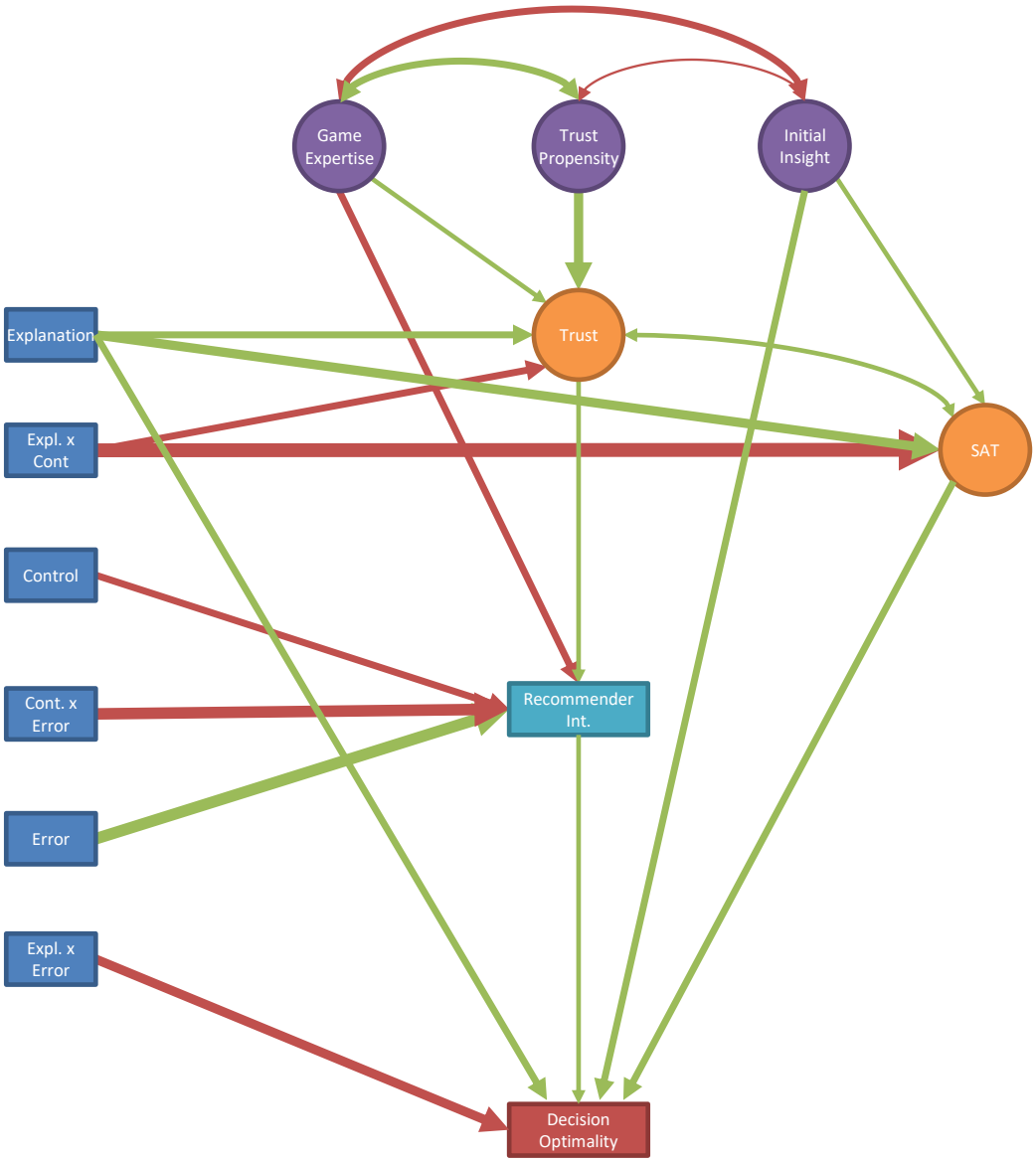


Figure 8.10: A visualization of all factors that directly or indirectly affect decision optimality. Line thickness indicates magnitude of β or B parameter. Green lines indicate a positive effect. Red lines indicate a negative effect. Note that this is just a visualization of a portion of the model reported in Table 8.7, and does not represent a tested statistical model.

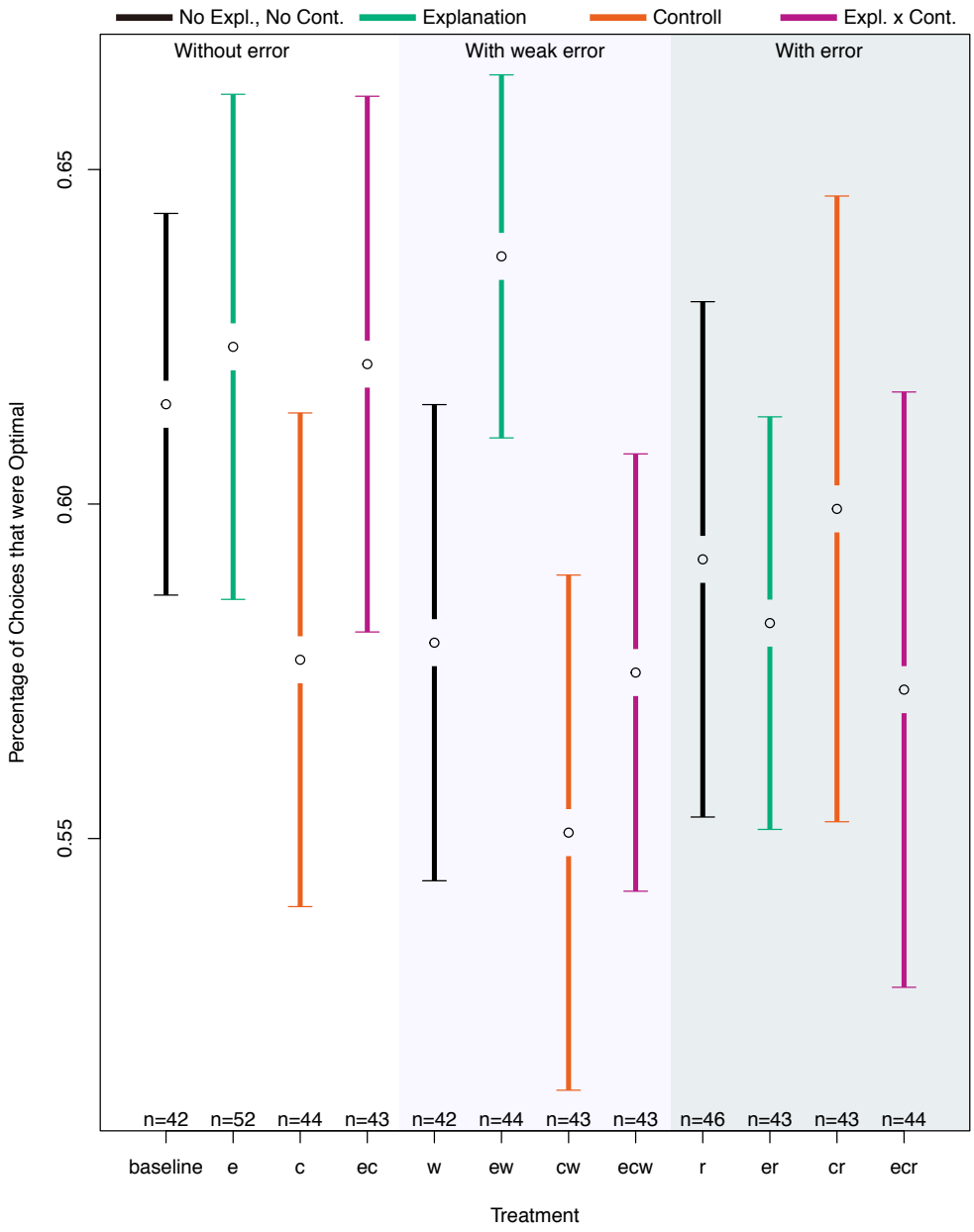


Figure 8.11: Mean percent of optimal choices made in each treatment. Explanations improve performance between "c" and "ec" as well as between "w" and "ew." Error bars are 95% confidence interval.

Chapter 9

Conclusion

In this chapter, the framework presented in Chapter 6 is evaluated using quantitative data from Chapters 7 and 8. Then, a meta-analysis of findings across all experiments is given, which is used to answer the original research questions given in Chapter 1.

9.1 Evaluation of Framework

In this section, we evaluate the measurement framework to identify the factors that bear on adherence and decision making. This analysis can guide future research in the area. This analysis can also be gleaned indirectly from examining regression (β) coefficients in the final fitted SEMs, but we found it more informative to visually compare several different SEMs that were constructed on the participant data:

9.1.1 Evaluating the HAI Measurements for Recommender Systems

The measurement framework for the recommender study includes the metrics listed in Chapter 7, Tables 7.2, 7.1, and 7.3, the basic observed variables (browser interac-

tion, recommender interaction, adherence), and the independent variables (explanation, control, error). Our analysis of the recommender measurement framework also allows us to compare with measurement frameworks that take a more user-experience centric approach (similar to [18] and [19]) when explaining decision making.

- **Black Box**: only independent variables: explanation, control, error, and their interaction effects
- **User Profiling**: independent variables + trust propensity, movie expertise, insight, cognitive reflection
- **Behavior**: independent variables + recommender interaction, browser interaction
- **SAT**: independent variables + SAT
- **User Experience (UXP)**: independent variables + user experience as shown in Table 7.2
- **Subjective System Aspects (SSA)**: independent variables + user experience split into perceived transparency, perceived quality, perceived control, and trust in the recommender
- **All-factor**: all independent and dependent variables included, user experience is as shown in Table 7.2

Each SEM was constructed in an identical way by ordering variables in terms of their causality (e.g., cognitive load cannot be a cause of trust propensity), saturating all regressions, then iteratively trimming non-significant effects. The resulting models were then compared in terms of their R^2 for decision satisfaction (note that due to the ratio of the sample size to number of variables –526:14–, adjusted R^2 and R^2 can only be up to about 1% different, so we report R^2 for simplicity).

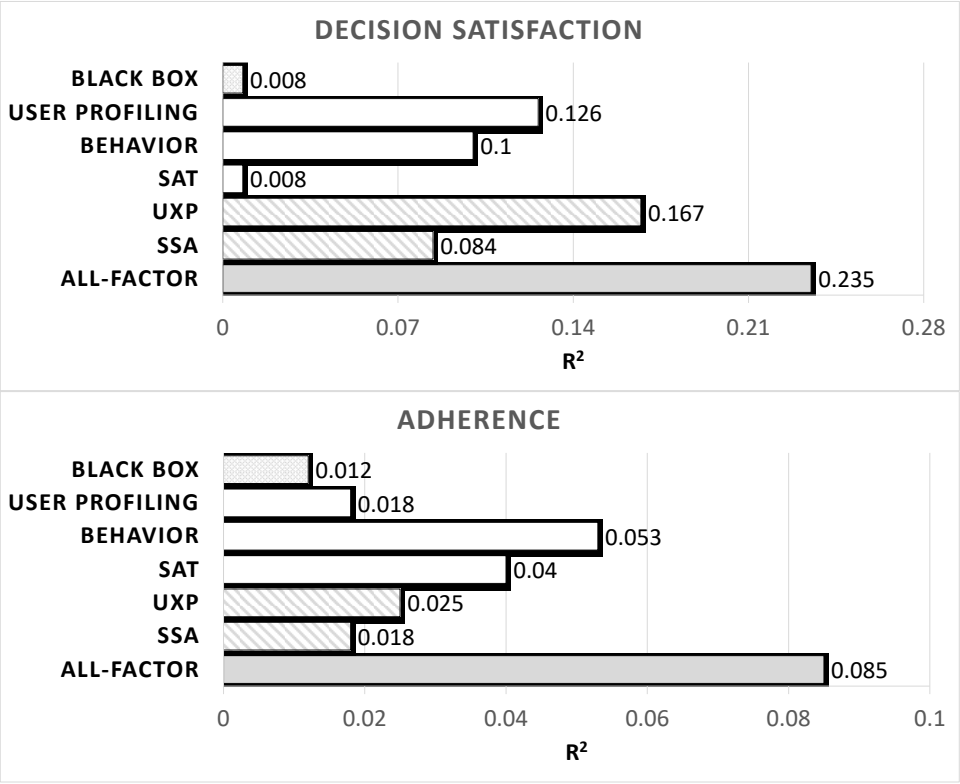


Figure 9.1: An evaluation of the factors that explain variance in decision satisfaction and adherence. R^2 indicates how much each model "determines" each variable, with 1.0 indicating perfect fit and 0.0 indicating no fit/complete noise. The all-factor model explains 40% more decision variance than the user experience model and 60% more adherence variance than the behavior model.

R^2 values of all evaluated HAI models are shown in Figure 9.1. The all-factor model explained the most decision satisfaction and the most adherence. Decision satisfaction was best explained by personal user characteristics, user experience, and our treatment variables. Adherence was linked most strongly to participant behavior and SAT. Note that each component of the all-factor was successful at explaining at least one of the observed outcomes in the study. Finally, note that the black box model almost completely fails at explaining decision satisfaction and does poorly when explaining adherence. Thus, decision satisfaction and adherence could not be explained without the intermediate cognitive, user experience, and behavior metrics.

Second, it can be observed that the SSA model performed significantly worse than the simple UXP model. While the SSA had good item to factor fit and significant regressions, the overall fit of the model was below the threshold for acceptability ($RMSEA = 0.17$).

9.1.2 Evaluating the HAI Measurements for the Diner's Dilemma

The measurement framework for the Dining Guru study includes the metrics listed in Chapter 8, Tables 8.2, 8.3, 8.5, and 8.4, the basic observed variables (recommender access, adherence), and the independent variables (explanation, control, error).

- **Black Box**: only independent variables: explanation, control, error, and their interaction effects
- **User Profiling**: independent variables + trust propensity, game expertise, insight, cognitive reflection
- **Behavior**: independent variables + Dining Guru access
- **SAT**: independent variables + SAT
- **Trust**: independent variables + trust and perceived control as shown in Table 7.2
- **All-factor**: all independent and dependent variables included, user experience is as shown in Table 7.2

Each SEM was constructed in an identical way by ordering variables in terms of their causality (e.g., cognitive load cannot be a cause of trust propensity), saturating all regressions, then iteratively trimming non-significant effects. The resulting models were then compared in terms of their R^2 for decision optimality (note that due to the ratio of the sample size to number of variables –529:11–, adjusted R^2 and R^2 can only be up to about 1% different, so we report R^2 for simplicity).

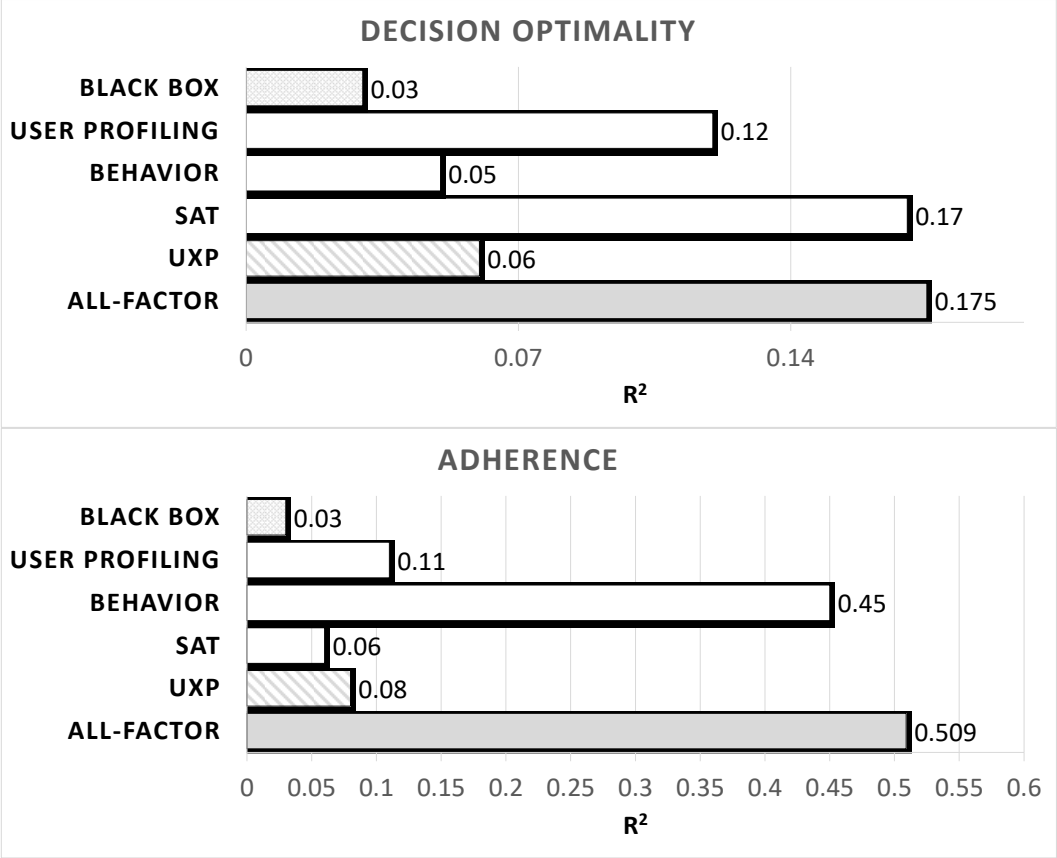


Figure 9.2: An evaluation of the factors that explain variance in decision optimality and adherence. R^2 indicates how much each model "determines" each variable, with 1.0 indicating perfect fit (no scatter) and 0.0 indicating no fit/complete noise.

R^2 values of all evaluated HAI models are shown in Figure 9.2. Decision optimality was best explained by personal characteristics and SAT. Adherence was linked most strongly to participant behavior and trust. Note that, again, each component of the all-factor was successful at explaining at least one of the observed outcomes in the study. Again, the black box model almost does poorly when explaining both decision optimality and adherence, indicating the intermediate cognitive variables are important mediators.

9.1.3 Comparison of Metric Explanatory Power

There are some general similarities in the predictive power of the HAI model for each task. User profiling characteristics were much better at predicting decision quality rather than adherence. In the movie selection study, the profiling metrics indicated how hard it would be to satisfy the user. In the Diner's Dilemma study, user characteristics indicated whether or not the player possessed the knowledge and decision-making capabilities to play the game well. Human interaction behavior also indicated whether or not they adhered to recommendations from the DSS, however, in practice this prediction is not especially useful unless it is known which factors affect human interaction decisions. Finally, in both studies, the effects of explanation, control, and error could not be understood without the mediating variables. This indicates that explanation, control, and error in DSS *influence* human cognitive states and behavior, rather than directly determining them. In both studies, every factor that was measured contributed either to mediating decision quality or adherence. Thus, the authors strongly recommend that these factors be measured in future experiments when the goal is to understand decision making or adherence. Recommendations for doing this are given in the next section.

There are some differences in the predictive power of the HAI models as well. Previously, it has been noted that the distinction between subjective vs. objective measures for success is often overlooked in HCI research [157], however, in this work we can draw some implications due to subjective vs. objective decision requirements in each task. In the Diner's Dilemma task, where decision success was based on objective criteria, SAT was much more successful at predicting decision success than in the movie selection task, where success was subjective. This can be attributed to the fact that players of the Diner's Dilemma game that understood what the Dining Guru was doing, regardless of whether or not they understood the actual game, did well in terms of optimality.

Next, the general success of the user experience metrics to explain decision satisfaction in the movie selection task may be attributed to differences in people's standards of taste. For instance, there could be an un-modeled variable, "ease of satisfaction," that may correlate strongly with user experience and have a significant relationship with decision satisfaction. Furthermore, the relationship between insight and user experience/decision satisfaction in the MS study indicates that domain experts may become harder to satisfy over time. Dis-entangling positive user experiences with systems and subjective satisfaction with selected items remains an open question in recommender systems research.

9.1.4 Adopting Factors into Future Work

In Chapters 7 and 3, item specification for each factor is given for each experiment. In future experiments, it would be possible for test designers to discard items with lower R^2 and only use single items for each factor. Unfortunately, this does not apply to the factors that were assessed with testing questionnaires (insight and SAT). Refer to Chen et al. [54] for additional recommendations on how to assess SAT, and refer to [6] for recommendations on how to assess insight. In this section, we summarize statistical information about the reliable factors that were tested and give recommendations for their future use.

First, the user profiling constructs (cognitive reflection, trust propensity, and reported expertise) were reliable between tasks (see Table 9.1). However, we caution that the participant pool was the same in both cases (Amazon Mechanical Turk). In different participant pools, the measurements may not be reliable. Additionally, it is noteworthy that CRT correlated strongly with initial insight in both tasks, despite the fact that the question content was so different. Finally, the one difference between co-variances between both studies was the lack of correlation between trust propensity and cognitive

User Profile Factor Co-variance	*MS β*	*DG β*
Trust Propensity ↔ Reported Expertise	**0.46 (***)**	**0.30*** **
Trust Propensity ↔ Cognitive Reflection	-0.23 (***)	N.S.
Trust Propensity ↔ Initial Insight	**-0.11 (**)**	**-0.09***
Reported Expertise ↔ Cognitive Reflection	**-0.16 (***)**	**-0.12***
Cognitive Reflection ↔ Initial Insight	**0.29 (***)**	**0.36*** **

Table 9.1: Co-variances between human-profiling factors across both tasks.

reflection in the Dining Guru study. This can be attributed either to the lower reliability of the cognitive reflection factor in the Dining Guru study, or slight differences in the way that trust propensity was measured. In the movie selection task, trust propensity was recommender specific. In the Dining Guru study, trust propensity was related to broader trust in automation.

Trust and cognitive load were reliable in both studies. For cognitive load, including the word "frustration" resulted in the highest R^2 of all items for each study. For trust, the items that contained the actual word of "trust" had the highest R^2. It is worth noting that, in the movie selection study, all questions related to user experience correlated with the question about trust. The empirical data in this study implies that trust is the most general representation of other user experience metrics (perceived effectiveness, perceived control, perceived transparency). Additionally, note the weak correlations between cognitive load, SAT, and trust in both the MS and DG studies, indicating good general discriminant validity. While previous research has found that *global* SA and cognitive load are strongly correlated [158][159], this research suggests that the DSS-specific SAT metric is separable.

Finally, we recommend that fewer items for user experience factors be included in future experiments, due to high inter-correlations in both studies. It has been noted that As suggested before, the most simple data model indicates that users have a "uni-dimensional" view of decision support systems - either "I like this" or "I don't like this."

However, we note that perceived control was discriminant from trust/user experience in the Dining Guru study (co-variance = 0.30). This was likely because the control/no control treatment in that study *completely removed automation from the Dining Guru*. In other contexts where the differences between explanation/no explanation or noise is severe, it may be possible for other subjective systems aspects to be teased out. However, we cannot recommend that participant testing time be taken up with too many subjective metrics due to their high inter-correlation. We recommend using a single item for perceived control, transparency, and effectiveness. An exploratory factor analysis can determine whether or not these items should be combined into a single factor with trust, or separated out.

9.2 Meta Analysis

In this section, the results from all empirical studies are compared and analyzed. We present the level of support for each effect that was discovered. Table 9.2 compares the treatments, user profiling, and dependent measurement models from each study. The movie selection study (MS) and Dining Guru study (DG) have identical measurement models, which facilitates a comparison of effects found in each study. A summary of the differences in task and treatment between the MS and DG study are shown in Table 9.3. Both the Traffic (TR) and Diner's Dilemma (DD) studies used a baseline condition where the decision support system was removed altogether, however, the hypothetical recommendation (HR), MS, and DG studies only varied the presence of DSS features.

The main difference between the MS and DG studies lies in the nature of the task and the criterion for decision success. Both task spaces have been studied extensively in the decision sciences. The Diner's Dilemma represents a single-attribute binary choice task, which is a special case of multi-attribute binary choice, which is a special case of

Study	**Treatments**	**Profiling**	**Measurements**	**Data Rows**
Traffic (TR)	Explanation, Support from DSS	Social Media Use	Decision optimality, insight, perceived effectiveness, cognitive load	>20,000
Hypothetical Recommendation (HR)	Explanation	Social Media Use	Perceived effectiveness, trust, interaction behavior	>10M
Diners Dilemma (DD)	Support from DSS, Task parameters	Trust Propensity	Insight (and Global SA), Decision Optimality	∼100
Movie Selection (MS)	Explanation, control, error (ECR)	Cognitive reflection, trust propensity, expertise, insight	Trust, cognitive load, insight, interaction, adherence, decision satisfaction	>10M
Dining Guru (DG)	Explanation, control, error (ECR)	Cognitive reflection, trust propensity, expertise, insight	Trust, cognitive load, insight, interaction, adherence, decision optimality	∼150

Table 9.2: A comparison of the measurement and treatment manipulation models of all studies presented in this dissertation.

multiple alternative choice. Moreover, multi-attribute binary choice is one of the final decision stage of multiple alternative choice tasks [160], as decision makers quickly filter large numbers of alternatives at the start of the the decision making process and then choose between a few alternatives. Additionally, LeJarraga et al built a general model of human decision-making in binary choice tasks [161]. The iterated Prisoner's Dilemma has also been studied extensively and its applicability to real-world situations has been well established[1] [162][163][164]. Decision success for each study was based on different parameters, with success in the MS study being subjective and success in the DG study being completely objective. It should also be noted that the treatment manipulations for explanation and control were minimal. This was done due to an understanding that decision makers are sensitive to the environment in which decisions are made [160] and also to increase the relevance of the results (it is easier to implement a text-based explanation that a visual one). Differences between effects in the MS and DG study can thus be attributed to differences in the task parameters and decision criterion, while similarities in effects thus have strong support for their generalization.

9.2.1 User Profiling

A summary of effects linked to personal user characteristics is shown in Table 9.4. Across both studies, trust propensity predicted higher perceptions of the decision support system, whether this was increased trust, user experience, or perceived control. Cognitive reflection also co-varied significantly with initial insight tests regardless of domain (in fact, results from both studies suggest humans can be split into high CRT/high insight and high trust/high "expertise" groups). In the MS study, there was a link between trust propensity, user experience, and low SAT, but this was not seen in the DG study,

[1]https://www.wired.com/2012/10/lance-armstrong-and-the-prisoners-dilemma-of-doping-in-professional-sports/

Study	**MS**	**DG**
Decision Task	Catalog browsing	Binary choice
Number of Decision Iterations	5-7	173
Simple Decision Support	Search/rank/filter interface	Game Summary Interface
Complex Decision Support	Collaborative Filtering	Maximum Likelihood Estimation
Decision Success Criteria	Subjective	Objective
Domain	Movie Metadata knowledge of distributions about metadata space	Game Rules knowledge of mapping from current game state to correct choice
Explanation Manipulation	Text-based explanation of algorithms calculation	Text-based explanation of algorithms calculation
Control Manipulation	Allows additional metadata filters to be applied on ranked recommendation list	Requires specification of input parameters/allows exploration of metadata space
Error Manipulation	Noise added to recommendation score, changing top recommendations	Noise added to expected values of binary choice, changing per-round recommendations

Table 9.3: Comparison of task parameters, decision success criteria, and treatment differences between the movie selection (MS) and Dining Guru (DG) studies.

Effect	*MS*	*DG*	*Other?*
Trust propensity predicts higher perceptions of DSS	**Yes (***)**	**Yes (***)**	
Trust propensity predicts more incorrect beliefs about DSS	Yes (**)	No	
Cognitive reflection predicts more correct beliefs about DSS	Yes (**)	No	
Cognitive reflection predicts more correct (pretask) beliefs about data domains	**Yes (***)**	**Yes (***)**	
Cognitive reflection predicts more correct (post-task) beliefs about data domains	**Yes (***)**	**Yes (***)**	
Self-reported expertise predicts less interaction with a DSS	**Yes (***)**	**Yes (***)**	
Self-reported expertise predicts less adherence to a DSS	No	Yes (***)	
Correct beliefs about data domains predicts more interaction with DSS	Yes (Browser*), Recommender(**)	No	
Reported trust in a DSS predicts correct beliefs about that DSS	No	Yes (**)	

Table 9.4: Support for effects related to user profiling factors. Results that have been reproduced in this dissertation are shown in bold.

which indicated a co-variance between SAT and trust. A link between trust propensity and recommender perceptions was also reported in [18]. The DSS presented in the DG study was more simple than the one presented in the MS study, which may explain why cognitive reflection was significant in MS but not in DG. Finally, high-insight users in the MS study interacted more with the recommender, but high insight was not a predictive factor in interaction with the Dining Guru. This may be explained by differences in DSS facilities - the recommender provided information (the recommendation score) that was not present on the browser side of the interface. The Dining Guru only aided in summarizing what information was already available, thus, more capable players may not have seen an increased need for use.

9.2.2 SAT-Insight Theory

A summary of effects linked to SAT and insight is shown in Table 9.5. Across both studies, SAT was an effective mediator of the effects of explanation and SAT was also linked to higher post-task insight and adherence. Where decision success was objective (DG and DD), higher insight predicted better decision performance. Cognitive load predicted post-task insight in both the DG study and the TR study. Finally, in the DG study, increased SAT predicted increased performance. This final effect, which was not present in the MS study, may have been due to the better alignment between goals of the Dining Guru DSS and the task performance, which was theoretically fixed to the maximum (and even in error treatments, still did better than the mean human decision maker). It should be noted that these factors predicted better post-task insight, but they did not explain any insight *change* that may have occurred during interaction with the DSS (those effects are listed in the final subsection here). These research results suggest that, to effect high post-task insight in situations where DSS must be used, maximal understanding of DSS should be a goal of system designers. Moreover, systems should engage user cognition by reducing the level of automation, or by providing more way to interact (control was a significant cause of cognitive load in both the MS and DG studies). While the models built here controlled for cognitive reflection (which is a good indicator of decision making ability), more research would be needed to determine if cognitive load and SAT are causes of increased insight or are simply being affected by an un-modeled variable.

9.2.3 User Experience/Trust

A summary of effects linked to SAT and insight is shown in Table 9.6. Cognitive load was negatively correlated with user perceptions of the DSS, indicating the potential for

Effect	*MS*	*DG*	*Other?*
Explanation causes an increase in correct beliefs about DSS	**Yes (*)**	**Yes (**)**	
Correct beliefs about DSS predicts correct (posttask) beliefs about data domains	**Yes (**)**	**Yes (.)**	
Correct beliefs about DSS predicts increased adherence	**Yes (*)**	**Yes (***)**	
Correct beliefs about data domains predicts better decision performance	**Yes (***)**	**No**	**DD (***)**
Correct beliefs about DSS predicts better decision performance	No	Yes (***)	
Increased cognitive load predicts increased insight	**No**	**Yes (***)**	**TR (**)**

Table 9.5: Support for effects related to SAT and insight. Results that have been reproduced in this dissertation are shown in bold.

DSS to mentally relieve analysts. Higher perceptions/trust also led to more adherence in both studies, although the predicting factor in DG was perceived control, whereas trust itself only had a mediating relationship with adherence through interaction with the Dining Guru. The effect of perceived control should not be seen as too surprising, as the control treatment caused a significant increase in perceived control and users that took the time to specify control parameters in the DG study would not have done so unless they planned to adhere to the system output. Finally, note that higher system perceptions was linked to higher satisfaction with selected items in the MS study. It may be that increased system satisfaction caused an increase in item satisfaction, however, it may also be that some users are generally easy to satisfy and some are generally hard to satisfy. More research where ease of satisfaction is controlled for would be needed to fully contextualize these results.

9.2.4 Expected Responses to the ECR Profile

A summary of effects linked to SAT and insight is shown in Tables 9.7, 9.8, and 9.9.

Effect	*MS*	*DG*	*Other?*
Cognitive load and user perceptions of DSS are negatively correlated	**Yes (***)**	**Yes (***)**	
Higher user perception of DSS predicts better decision performance	Yes (***)	No	
High user perception of DSS predicts more adherence	**Yes (*)**	**Yes, perceived control only (**)**	

Table 9.6: Support for effects related to user perceptions and trust. Results that have been reproduced in this dissertation are shown in bold.

Effect	*MS*	*DG*	*Other?*
Explanation improves decision performance	**Yes (*)**	**Yes(*)**	
Explanation increases user perception of DSS	**Yes, in a more complex alternative model (*)**	**Yes, both trust (*) and perceived control (***)**	**HR (trust***, perceived acc*)**
Explanation increases user interaction with DSS	**No**	**Yes (*), full mediation by trust (***)**	**HR (.)**
Explanation increases effectiveness of control actions	No	No	HR(*)
Explanations from complex DSS cause users to form incorrect beliefs about the data	**Yes (.), mitigated by error**	**Yes(**)**	

Table 9.7: Support for effects caused by DSS explanation in the studies. Results that have been reproduced in this dissertation are shown in bold.

Effect	*MS*	*DG*	*Other?*
Control increases cognitive load	**Yes (**)**	**Yes, when paired with explanation (**)**	
Control increases adherence	**Yes (*)**	**Yes (***), mediated by perceived control (**)**	
Control increases user perception of DSS	No	Yes, perceived control (***)	
Control increases decision performance	**Yes (**)**	**Yes (***), full mediation by perceived control (***) and adherence (***)**	
Control over a complex DSS causes users to form incorrect beliefs about the data	**Yes (**), mitigated by explanation and error**	**Yes(***), exacerbated by explanation**	

Table 9.8: Support for effects caused by control manipulations in the studies. Results that have been reproduced in this dissertation are shown in bold.

Effect	*MS*	*DG*	*Other?*
Error decreases user perceptions of DSS	Yes (***)	No	
Error decreases decision performance	**Yes (***), full mediation by user perception (***)**	**Yes (*), when paired with explanation**	
Error decreases adherence	No	Yes (**)	
Errors in a complex DSS cause users to form incorrect beliefs about the data	Yes (*), mitigated by explanation and control	No	

Table 9.9: Support for effects caused by error manipulations in the studies. Results that have been reproduced in this dissertation are shown in bold.

In both studies, the presence of explanations caused better decision success. Recall that in both studies, explanation was given under varying levels of DSS error. In the MS study, browser interaction increased significantly when explanations and error were both present. In the DG study, explanation predicted an increase in adherence through mediating variables, but error predicted a large drop in adherence. These results suggest that explanations can potentially help users identify when DSS systems make errors so that alternatives can be used instead, regardless of whether decision success is subjective or objective. The authors thus recommend the use of explanations, as well as multiple alternative information systems, whenever decision success is critical.

Next, explanations increased user perceptions of the DSS in the DG study (both trust and perceived control) as well as in the hypothetical recommendation study (trust, perceived accuracy). The all-factor SEM reported in Chapter 7 does not predict increased user perceptions as the result of explanation, but a slightly more complex, worse-fitting alternative model does predict this outcome (however, the perception gain goes away when noise is present, $p = ***$). The importance of explanations in recommender systems may be related only to human decision making ability, whereas system perceptions may be more fruitfully improved simply by making recommendation systems more accurate and relevant. More experiments where recommender error is manipulated along with explanation would be needed to verify this finding.

Explanations increased user interaction with the DSS in both the DG and HR studies, however, an effect was not found in the MS study on either browser interaction or recommender interaction. It should be noted that the increase in user interaction found in the HR study was minor, but that this study also found that the effectiveness of profile deletion actions improved when explanations (dynamic feedback) was given to participants. While our data did not support the finding that explanations did not increase user interaction in the MS study, the presence of additional control features did. Ac-

tually, explanations caused increased interaction in the DG study via trust mediation, but increased user experience with the recommender in the MS study predicted fewer interactions. An explanation for this is that users in the MS study that received poor initial recommendations had lower perceptions of the recommender and also required the specification of additional filters on the recommendations to find titles that matched their decision criteria.

Control features increased both cognitive load, adherence, and decision success across both the MS and DG studies. Control features in the MS study allowed users to customize the recommendation view to their tastes, getting the benefits of both traditional filtering and collaborative filtering. Control features in the DG study allowed users to explore the space of decision outcomes and also the "automation" from the Dining Guru. Users adhered more to the Dining Guru's recommendations when given control and thus decision optimality was improved (again, the Dining Guru performed significantly better than the mean in most treatments). In other words, users took advice from the system when they believed the choice was their idea. Cognitive load is an unfortunate side effect of adding control features, however, increased cognitive load also predicted increased post-task insight. Cognitive load can also be reduced by improving interfaces for control features. Control was not linked with increased user perceptions in any study, although perceived control (unsurprisingly) increased in the Dining Guru study.

Error from the DSS decreased decision performance in both the MS and DG studies. In the MS study, this effect was fully mediated by user perception (with no direct effect found), indicating that users may have simply turned to the browser side of the interface when the recommender failed. In the DG study, the negative effects of error were only found when explanations were also provided, indicating that the explanations allowed participants to see the flaws in the Dining Guru's predictions, and steering them away from adherence despite its optimality relative to the mean of participant performance.

Finally, explanation, control, and error (in the MS study) caused incorrect beliefs to form during the period when the DSS was available. To explain this, recall that 1) in the MS study, users interacted with the browser more and the recommender less in the baseline condition (no ECR), 2) control significantly increased interaction with the MS recommender, which was also linked to the formation of incorrect beliefs (see Chapter 7), 3) users in the explanation and error conditions in MS had a much smaller negative change in insight when compared to the control only condition, 4) explanation and control in the DG study significantly increased cognitive variables that predicted increases in DSS interaction and adherence, 5) EC features, in both studies, increased both the amount of screen space occupied by the DSS or made its behavior less predictable, drawing attention. We thus propose the following explanation: explanation or control features increase the attractiveness of DSS usage, while errors may increase the potential for confusion. In either case, users try to match their mental belief models of both the DSS (SAT) and the data space (insight). Due to the DSS not always representing the data space accurately (MS), relieving cognitive load by doing calculations which may otherwise reinforce insight (DG), or increasing the number of entities to consider during tasks (both studies), users may form incorrect beliefs. Users in baseline conditions, with accurate, featureless DSS, thus have more mental bandwidth for learning and fewer entities to reconcile in their mental model, leaving more cognitive bandwidth for addressing other concerns. Future work will need to focus on determining which of these reasons is the primary cause for decreased insight.

9.3 Summary

This research has investigated how human cognition reacts to the presence and configuration of decision support systems. We have identified general system, user, and

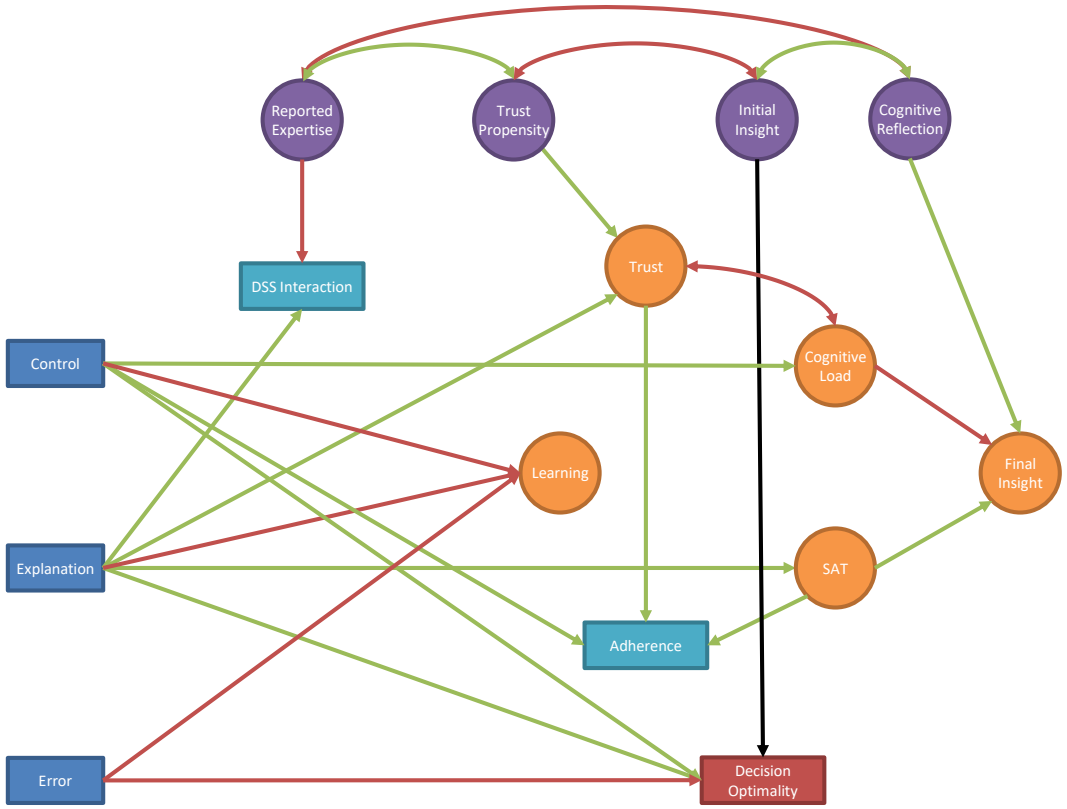


Figure 9.3: A visualization of effects found in this dissertation which have been reproduced. Green lines indicate a positive effect. Red lines indicate a negative effect. Black lines indicate disagreement on the effect direction.

cognitive factors that predict decision behaviors related to interaction with systems, incorporate of system predictions (adherence), and domain decision success. We have presented surprising effects related to user beliefs about systems and data which opens a door to future investigations. The analysis of multiple domains and the use of a common measurement methodology in two experiments has allowed us to better identify effects that should generalize well to other contexts. Furthermore, we have provided a detailed description of that measurement methodology, quantified its success in determining decision behaviors in the chosen domains and given recommendations for adapting measurements for new domains.

In the introduction, we posed the following research questions:

1. Which factors explain variability in decision making (interaction, adherence, success) in the HAI system?

2. How do personal user characteristics and ECR determine decision-making behavior?

3. What is the relationship between correct beliefs about algorithms, their use, and trust?

4. What is the relationship between user beliefs about algorithms and insight in data analysis?

The first research question is answered in part by Figures 9.1 and 9.2. The fitted SEMs from Chapters 7 and 8 answer this question further, as impacts of each factor can be inferred through β or B estimates. Figure 9.3 shows a visualization of the meta analysis given in the previous section. We provide the following answers to research questions 2, 3, and 4:

1. In this work, we have quantitative evidence that suggests that self-reported experts are also likely to be more trusting than the general population of users. These users interact less but adhere to advice more often. True domain experts are more likely to have higher cognitive reflection - these users generally have higher insight but are susceptible to forming incorrect beliefs. Decision support systems could potentially adapt to users based on these user profile metrics.

2. Situation-awareness based agent transparency (SAT) was an effective mediator of system explanation effects when trying to understand adherence to advice. This measurement is thus critical when system explanations are being designed or evaluated. Furthermore, there is not much evidence to indicate that trust and SAT are similar or strongly correlated. This suggests that less trusting users might be convinced to use a system through effecting correct beliefs.

3. SAT was a predictor of post-task insight when controlling for CRT, however, there are likely hidden factors that determine both due to a lack of observed mediation between ECR parameters and changes in insight. The Raykov models implied that drawing attention to the recommendation system through explanation/control caused the formation of incorrect beliefs.

Finally, it is interesting to point out that initial insight had a different effect direction in the DG and MS studies. This was likely due to the difference in decision success between the tasks, with MS being subjective and DG being objective.

While this research has identified a number of HAI factors that transfer across domains and while we have provided expectations for their general relationships in a very limited scope, more research in other decision and task contexts is needed to develop a reliable, general theory about how DSS affect human cognition and decision making behavior. Additional factor modeling, especially task and domain specific factors, will be essential in achieving high levels of prediction about how human-machine systems evolve. This study has also not studied the longitudinal effects of repeated DSS use on cognitive factors, nor how relationships between users and DSS evolve over long periods of time. The insight and SAT metrics used in this task are exploratory and require further validation and study in each task domain where they are applied. Finally, the effects reported here warrant further and more detailed study where more variables are controlled.

In summary, we have discovered that intermediate cognitive variables are crucial for understand the effects of explanation, control, and error in DSS. Furthermore, we discovered that 1) the user profiling metrics: trust propensity, cognitive reflection, reported expertise, and insight increase the ability to predict decision making behaviors in the presence of a DSS, 2) correct user beliefs (SAT) about DSS mediate the effect of system explanation when predicting adherence to recommendations, and 3) while explanations

and control increase trust, user perception, interaction, and adherence with DSS, they also have the potential to cause human analysts to form incorrect beliefs, which can lead to incorrect decisions or affect future decision-making behavior.

Bibliography

[1] N. Carr, *The glass cage: Where automation is taking us*. Random House, 2015.

[2] M. R. Endsley, *Designing for situation awareness: An approach to user-centered design*. CRC Press, 2011.

[3] R. K. Ahuja, K. Mehlhorn, J. Orlin, and R. E. Tarjan, *Faster algorithms for the shortest path problem*, *Journal of the ACM (JACM)* **37** (1990), no. 2 213–223.

[4] J. S. Breese, D. Heckerman, and C. Kadie, *Empirical analysis of predictive algorithms for collaborative filtering*, in *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pp. 43–52, Morgan Kaufmann Publishers Inc., 1998.

[5] D. A. Norman, *Cognitive engineering*, *User centered system design: New perspectives on human-computer interaction* **3161** (1986).

[6] C. North, *Toward measuring visualization insight*, *Computer Graphics and Applications, IEEE* **26** (2006), no. 3 6–9.

[7] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis, D. Keim, *et. al.*, *Knowledge generation model for visual analytics*, *Visualization and Computer Graphics, IEEE Transactions on* **20** (2014), no. 12 1604–1613.

[8] R. H. Sprague Jr, *A framework for the development of decision support systems*, *MIS quarterly* (1980) 1–26.

[9] G. Bellinger, D. Castro, and A. Mills, *Data, information, knowledge, and wisdom*, .

[10] S. Albright, W. Winston, and C. Zappe, *Data analysis and decision making*. Cengage Learning, 2010.

[11] M. R. Endsley, *Toward a theory of situation awareness in dynamic systems*, *Human Factors: The Journal of the Human Factors and Ergonomics Society* **37** (1995), no. 1 32–64.

[12] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, *Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs*, *Journal of Cognitive Engineering and Decision Making* **2** (2008), no. 2 140–160.

[13] M. R. Endsley, *Measurement of situation awareness in dynamic systems*, *Human Factors: The Journal of the Human Factors and Ergonomics Society* **37** (1995), no. 1 65–84.

[14] J. O'Donovan and B. Smyth, *Trust in recommender systems*, in *Proceedings of the 10th international conference on Intelligent user interfaces*, pp. 167–174, ACM, 2005.

[15] J. L. Harman, J. O'Donovan, T. Abdelzaher, and C. Gonzalez, *Dynamics of human trust in recommender systems*, in *Proceedings of the 8th ACM Conference on Recommender systems*, pp. 305–308, ACM, 2014.

[16] R. R. Hoffman, M. Johnson, J. M. Bradshaw, and A. Underbrink, *Trust in automation*, *IEEE Intelligent Systems* **28** (2013), no. 1 84–88.

[17] J. D. Lee and K. A. See, *Trust in automation: Designing for appropriate reliance*, *Human Factors: The Journal of the Human Factors and Ergonomics Society* **46** (2004), no. 1 50–80.

[18] B. P. Knijnenburg, S. Bostandjiev, J. O'Donovan, and A. Kobsa, *Inspectability and control in social recommenders*, in *Proceedings of the sixth ACM conference on Recommender systems*, pp. 43–50, ACM, 2012.

[19] P. Pu, L. Chen, and R. Hu, *A user-centric evaluation framework for recommender systems*, in *Proceedings of the fifth ACM conference on Recommender systems*, pp. 157–164, ACM, 2011.

[20] M. J. Eppler and J. Mengis, *The concept of information overload: A review of literature from organization science, accounting, marketing, mis, and related disciplines*, *The information society* **20** (2004), no. 5 325–344.

[21] V. Arnold, N. Clark, P. A. Collier, S. A. Leech, and S. G. Sutton, *The differential use and effect of knowledge-based system explanations in novice and expert judgment decisions*, *Mis Quarterly* (2006) 79–97.

[22] M. Schneider-Hufschmidt, U. Malinowski, and T. Kuhme, *Adaptive user interfaces: Principles and practice*. Elsevier Science Inc., 1993.

[23] B. Shneiderman, *Promoting universal usability with multi-layer interface design*, in *ACM SIGCAPH Computers and the Physically Handicapped*, no. 73-74, pp. 1–8, ACM, 2003.

[24] J. S. Yi, Y.-a. Kang, J. T. Stasko, and J. A. Jacko, *Understanding and characterizing insights: how do people gain insights using information visualization?*, in *Proceedings of the 2008 Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization*, p. 4, ACM, 2008.

[25] R. Chang, C. Ziemkiewicz, T. M. Green, and W. Ribarsky, *Defining insight for visual analytics*, *Computer Graphics and Applications, IEEE* **29** (2009), no. 2 14–17.

[26] E. Shortliffe, *Computer-based medical consultations: MYCIN*, vol. 2. Elsevier, 2012.

[27] P. Resnick and H. R. Varian, *Recommender systems*, *Communications of the ACM* **40** (1997), no. 3 56–58.

[28] J. Y. Chen and M. J. Barnes, *Supervisory control of multiple robots effects of imperfect automation and individual differences*, *Human Factors: The Journal of the Human Factors and Ergonomics Society* **54** (2012), no. 2 157–174.

[29] L. Greenemeier, *Robot pack mule to carry loads for gis on the move*, *Scientific American* (2010).

[30] K. Wu, D. Gauthier, and M. D. Levine, *Live cell image segmentation*, *Biomedical Engineering, IEEE Transactions on* **42** (1995), no. 1 1–12.

[31] P. Hancock and S. Scallen, *The future of function allocation*, *Ergonomics in Design: The Quarterly of Human Factors Applications* **4** (1996), no. 4 24–29.

[32] V. D. Hopkin, *Air traffic control.*, .

[33] J. Gratch, J. Rickel, E. André, J. Cassell, E. Petajan, and N. Badler, *Creating interactive virtual humans: Some assembly required*, tech. rep., DTIC Document, 2002.

[34] S. Gregor and I. Benbasat, *Explanations from intelligent systems: Theoretical foundations and implications for practice*, *MIS quarterly* (1999) 497–530.

[35] I. Altintas, C. Berkley, E. Jaeger, M. Jones, B. Ludascher, and S. Mock, *Kepler: an extensible system for design and execution of scientific workflows*, in *Scientific and Statistical Database Management, 2004. Proceedings. 16th International Conference on*, pp. 423–424, IEEE, 2004.

[36] A. Kobsa, *Adaptive interfaces*, 2004.

[37] B. Shneiderman, *Human responsibility for autonomous agents*, *Intelligent Systems, IEEE* **22** (2007), no. 2 60–61.

[38] B. A. Myers, A. J. Ko, and M. M. Burnett, *Invited research overview: end-user programming*, in *CHI'06 extended abstracts on Human factors in computing systems*, pp. 75–80, ACM, 2006.

[39] A. J. Ko, B. Myers, H. H. Aung, *et. al.*, *Six learning barriers in end-user programming systems*, in *Visual Languages and Human Centric Computing, 2004 IEEE Symposium on*, pp. 199–206, IEEE, 2004.

[40] M. Chen, D. Ebert, H. Hagen, R. S. Laramee, R. Van Liere, K.-L. Ma, W. Ribarsky, G. Scheuermann, and D. Silver, *Data, information, and knowledge in visualization*, *Computer Graphics and Applications, IEEE* **29** (2009), no. 1 12–19.

[41] A. Aamodt and M. Nygård, *Different roles and mutual dependencies of data, information, and knowledgean ai perspective on their integration*, *Data & Knowledge Engineering* **16** (1995), no. 3 191–222.

[42] P. Saraiya, C. North, and K. Duca, *An insight-based methodology for evaluating bioinformatics visualizations*, *Visualization and Computer Graphics, IEEE Transactions on* **11** (2005), no. 4 443–456.

[43] W. A. Pike, J. Stasko, R. Chang, and T. A. O'connell, *The science of interaction*, *Information Visualization* **8** (2009), no. 4 263–274.

[44] C. Stolte, D. Tang, and P. Hanrahan, *Polaris: A system for query, analysis, and visualization of multidimensional relational databases*, *Visualization and Computer Graphics, IEEE Transactions on* **8** (2002), no. 1 52–65.

[45] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon, *Visual analytics: Definition, process, and challenges*. Springer, 2008.

[46] D. A. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann, *Mastering the information age-solving problems with visual analytics*. Florian Mansmann, 2010.

[47] C. Plaisant, J.-D. Fekete, and G. Grinstein, *Promoting insight-based evaluation of visualizations: From contest to benchmark repository*, *Visualization and Computer Graphics, IEEE Transactions on* **14** (2008), no. 1 120–134.

[48] M. C. F. De Oliveira and H. Levkowitz, *From visual data exploration to visual data mining: a survey*, *Visualization and Computer Graphics, IEEE Transactions on* **9** (2003), no. 3 378–394.

[49] D. Keim, F. Mansmann, J. Schneidewind, H. Ziegler, *et. al.*, *Challenges in visual data analysis*, in *Information Visualization, 2006. IV 2006. Tenth International Conference on*, pp. 9–16, IEEE, 2006.

[50] T. M. Green, W. Ribarsky, and B. Fisher, *Building and applying a human cognition model for visual analytics*, *Information visualization* **8** (2009), no. 1 1–13.

[51] J. B. Ullman and P. M. Bentler, *Structural equation modeling*. Wiley Online Library, 2003.

[52] M. R. Endsley and D. J. Garland, *Situation awareness analysis and measurement*. CRC Press, 2000.

[53] M. R. Endsley, *Situation awareness global assessment technique (sagat)*, in *Aerospace and Electronics Conference, 1988. NAECON 1988., Proceedings of the IEEE 1988 National*, pp. 789–795, IEEE, 1988.

[54] J. Y. Chen, K. Procci, M. Boyce, J. Wright, A. Garcia, and M. Barnes, *Situation awareness-based agent transparency*, tech. rep., DTIC Document, 2014.

[55] J. Y. Chen, M. J. Barnes, and M. Harper-Sciarini, *Supervisory control of multiple robots: Human-performance issues and user-interface design*, *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* **41** (2011), no. 4 435–454.

[56] M. R. Endsley, *Designing for situation awareness: An approach to user-centered design*. Taylor & Francis US, 2003.

[57] B. Ludäscher, I. Altintas, C. Berkley, D. Higgins, E. Jaeger, M. Jones, E. A. Lee, J. Tao, and Y. Zhao, *Scientific workflow management and the kepler system*, *Concurrency and Computation: Practice and Experience* **18** (2006), no. 10 1039–1065.

[58] J. Sweller, *Cognitive load theory, learning difficulty, and instructional design*, *Learning and instruction* **4** (1994), no. 4 295–312.

[59] J. Sweller, *Cognitive load during problem solving: Effects on learning*, *Cognitive science* **12** (1988), no. 2 257–285.

[60] F. Paas, J. E. Tuovinen, H. Tabbers, and P. W. Van Gerven, *Cognitive load measurement as a means to advance cognitive load theory*, *Educational psychologist* **38** (2003), no. 1 63–71.

[61] J. A. Colquitt, B. A. Scott, and J. A. LePine, *Trust, trustworthiness, and trust propensity: a meta-analytic test of their unique relationships with risk taking and job performance.*, *Journal of applied psychology* **92** (2007), no. 4 909.

[62] H. Gill, K. Boies, J. E. Finegan, and J. McNally, *Antecedents of trust: Establishing a boundary condition for the relation between propensity to trust and intention to trust*, *Journal of business and psychology* **19** (2005), no. 3 287–302.

[63] M. K. Lee and E. Turban, *A trust model for consumer internet shopping*, *International Journal of electronic commerce* **6** (2001), no. 1 75–91.

[64] Y.-H. Chen and S. Barnes, *Initial trust and online buyer behaviour*, *Industrial management & data systems* **107** (2007), no. 1 21–36.

[65] B. P. Knijnenburg and A. Kobsa, *Making decisions about privacy: information disclosure in context-aware recommender systems*, *ACM Transactions on Interactive Intelligent Systems (TiiS)* **3** (2013), no. 3 20.

[66] M. Koufaris and W. Hampton-Sosa, *Customer trust online: examining the role of the experience with the web-site*, *Department of Statistics and Computer Information Systems Working Paper Series, Zicklin School of Business, Baruch College, New York* (2002).

[67] D. Kahneman, *Attention and effort*. Citeseer, 1973.

[68] S. Frederick, *Cognitive reflection and decision making*, *The Journal of Economic Perspectives* **19** (2005), no. 4 25–42.

[69] M. E. Toplak, R. F. West, and K. E. Stanovich, *The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks*, *Memory & Cognition* **39** (2011), no. 7 1275–1289.

[70] M. Welsh, N. Burns, and P. Delfabbro, *The cognitive reflection test: how much more than numerical ability*, in *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, pp. 1587–1592, Cognitive Science Society Austin, TX, 2013.

[71] K. Merritt, D. Smith, and J. Renzo, *An investigation of self-reported computer literacy: Is it reliable*, *Issues in Information Systems* **6** (2005), no. 1 289–295.

[72] D. Dunning, *5 the dunning-kruger effect: On being ignorant of one's own ignorance*, *Advances in experimental social psychology* **44** (2011) 247.

[73] J. Kruger and D. Dunning, *Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments.*, *Journal of personality and social psychology* **77** (1999), no. 6 1121.

[74] J. Kruger, *Lake wobegon be gone! the" below-average effect" and the egocentric nature of comparative ability judgments.*, *Journal of personality and social psychology* **77** (1999), no. 2 221.

[75] E. H. Shortliffe, R. Davis, S. G. Axline, B. G. Buchanan, C. C. Green, and S. N. Cohen, *Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the mycin system*, *Computers and biomedical research* **8** (1975), no. 4 303–320.

[76] E. H. Chi, *Blurring of the boundary between interactive search and recommendation*, in *Proceedings of the 20th International Conference on Intelligent User Interfaces*, pp. 2–2, ACM, 2015.

[77] J. O'Donovan, N. Tintarev, A. Felfernig, P. Brusilovsky, G. Semeraro, and P. Lops, *Joint workshop on interfaces and human decision making for recommender systems.*, in *RecSys* (H. Werthner, M. Zanker, J. Golbeck, and G. Semeraro, eds.), pp. 347–348, ACM, 2015.

[78] J. L. Herlocker, J. A. Konstan, and J. Riedl, *Explaining collaborative filtering recommendations*, in *Proceedings of ACM CSCW'00 Conference on Computer-Supported Cooperative Work*, pp. 241–250, 2000.

[79] *Inspection mechanisms for community-based content discovery in microblogs*, .

[80] M. Bilgic and R. J. Mooney, *Explaining recommendations: Satisfaction vs. promotion*, in *Beyond Personalization Workshop, IUI*, vol. 5, 2005.

[81] N. Tintarev and J. Masthoff, *A survey of explanations in recommender systems*, in *Data Engineering Workshop, 2007 IEEE 23rd International Conference on*, pp. 801–810, IEEE, 2007.

[82] N. Tintarev, J. O'Donovan, P. Brusilovsky, A. Felfernig, G. Semeraro, and P. Lops, *Recsys' 14 joint workshop on interfaces and human decision making for recommender systems*, in *Proceedings of the 8th ACM Conference on Recommender systems*, pp. 383–384, ACM, 2014.

[83] R. Sinha and K. Swearingen, *The role of transparency in recommender systems*, in *CHI'02 extended abstracts on Human factors in computing systems*, pp. 830–831, ACM, 2002.

[84] Y. L. Simmhan, B. Plale, and D. Gannon, *A survey of data provenance in e-science*, *ACM Sigmod Record* **34** (2005), no. 3 31–36.

[85] Y. Gil, E. Deelman, M. Ellisman, T. Fahringer, G. Fox, D. Gannon, C. Goble, M. Livny, L. Moreau, and J. Myers, *Examining the challenges of scientific workflows*, *Ieee computer* **40** (2007), no. 12 26–34.

[86] T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, K. Glover, M. R. Pocock, A. Wipat, *et. al.*, *Taverna: a tool for the composition and enactment of bioinformatics workflows*, *Bioinformatics* **20** (2004), no. 17 3045–3054.

[87] T. McPhillips, S. Bowers, D. Zinn, and B. Ludäscher, *Scientific workflow design for mere mortals*, *Future Generation Computer Systems* **25** (2009), no. 5 541–551.

[88] T. L. Robbins, *Social loafing on cognitive tasks: An examination of the sucker effect*, *Journal of Business and Psychology* **9** (1995), no. 3 337–342.

[89] J. Andreoni and J. H. Miller, *Rational cooperation in the finitely repeated prisoner's dilemma: Experimental evidence*, *The economic journal* (1993) 570–585.

[90] U. Gneezy, E. Haruvy, and H. Yafe, *The inefficiency of splitting the bill**, *The Economic Journal* **114** (2004), no. 495 265–280.

[91] D. M. Kreps, P. Milgrom, J. Roberts, and R. Wilson, *Rational cooperation in the finitely-repeated prisoners' dilemma.*, tech. rep., DTIC Document, 1982.

[92] V. Liberman, S. M. Samuels, and L. Ross, *The name of the game: Predictive power of reputations versus situational labels in determining prisoners dilemma game moves*, *Personality and social psychology bulletin* **30** (2004), no. 9 1175–1185.

[93] Y. Teng, R. Jones, L. Marusich, J. O'Donovan, C. Gonzalez, and T. Hollerer, *Trust and situation awareness in a 3-player diner's dilemma game*, in *Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), 2013 IEEE International Multi-Disciplinary Conference on*, pp. 9–15, IEEE, 2013.

[94] E. Onal, J. Schaffer, J. O'Donovan, L. Marusich, M. S. Yu, C. Gonzalez, and T. Hollerer, *Decision-making in abstract trust games: A user interface perspective*, in *Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), 2014 IEEE International Inter-Disciplinary Conference on*, pp. 21–27, IEEE, 2014.

[95] C. Gonzalez, N. Ben-Asher, J. Martin, and V. Dutt, *Emergence of cooperation with increased information: Explaining the process with instance-based learning models*, *Unpublished manuscript under review* (2013).

[96] J. M. Martin, I. Juvina, C. Lebiere, and C. Gonzalez, *The effects of individual and context on aggression in repeated social interaction*, in *Engineering Psychology and Cognitive Ergonomics*, pp. 442–451. Springer, 2011.

[97] J. M. Weber and J. K. Murnighan, *Suckers or saviors? consistent contributors in social dilemmas.*, *Journal of personality and social psychology* **95** (2008), no. 6 1340.

[98] E. Onal, J. ODonovan, L. Marusich, S. Y. Michael, J. Schaffer, C. Gonzalez, and T. Höllerer, *Trust and consequences: A visual perspective*, in *Virtual, Augmented and Mixed Reality. Designing and Developing Virtual and Augmented Environments*, pp. 203–214. Springer, 2014.

[99] D. M. Rousseau, S. B. Sitkin, R. S. Burt, and C. Camerer, *Not so different after all: A cross-discipline view of trust*, *Academy of management review* **23** (1998), no. 3 393–404.

[100] A. Rapoport and A. M. Chammah, *Prisoner's dilemma: A study in conflict and cooperation*, vol. 165. University of Michigan press, 1965.

[101] R. M. Dawes and R. H. Thaler, *Anomalies: cooperation*, *The Journal of Economic Perspectives* (1988) 187–197.

[102] R. M. Dawes, *Social dilemmas*, *Annual review of psychology* **31** (1980), no. 1 169–193.

[103] E. Fehr and K. M. Schmidt, *A theory of fairness, competition, and cooperation*, *Quarterly journal of Economics* (1999) 817–868.

[104] M. Rabin, *Incorporating fairness into game theory and economics*, *The American economic review* (1993) 1281–1302.

[105] E. Fehr and S. Gächter, *Altruistic punishment in humans*, *Nature* **415** (2002), no. 6868 137–140.

[106] R. H. Frank, T. Gilovich, and D. T. Regan, *Does studying economics inhibit cooperation?*, *The Journal of Economic Perspectives* (1993) 159–171.

[107] J. Dana, D. M. Cain, and R. M. Dawes, *What you dont know wont hurt me: Costly (but quiet) exit in dictator games*, *Organizational Behavior and Human Decision Processes* **100** (2006), no. 2 193–201.

[108] D. Kahneman, J. L. Knetsch, and R. H. Thaler, *Fairness and the assumptions of economics*, *Journal of business* (1986) S285–S300.

[109] J. Berg, J. Dickhaut, and K. McCabe, *Trust, reciprocity, and social history*, *Games and economic behavior* **10** (1995), no. 1 122–142.

[110] D. N. Hogg, K. FOLLES, F. Strand-Volden, and B. Torralba, *Development of a situation awareness measure to evaluate advanced alarm systems in nuclear power plant control rooms*, *Ergonomics* **38** (1995), no. 11 2394–2413.

[111] L. J. Gugerty, *Situation awareness during driving: Explicit and implicit knowledge in dynamic spatial memory.*, *Journal of Experimental Psychology: Applied* **3** (1997), no. 1 42.

[112] D. Cosley, S. K. Lam, I. Albert, J. A. Konstan, and J. Riedl, *Is seeing believing?: How recommender system interfaces affect users' opinions*, in *Proceedings of the Conference on Human Factors in Computing Systems*, pp. 585–592, ACM Press, 2003.

[113] L. D. Saner, C. A. Bolstad, C. Gonzalez, and H. M. Cuevas, *Measuring and predicting shared situation awareness in teams*, *Journal of cognitive engineering and decision making* **3** (2009), no. 3 280–308.

[114] M. Buhrmester, T. Kwang, and S. D. Gosling, *Amazon's mechanical turk a new source of inexpensive, yet high-quality, data?*, *Perspectives on psychological science* **6** (2011), no. 1 3–5.

[115] D. J. Hauser and N. Schwarz, *Attentive turkers: Mturk participants perform better on online attention checks than do subject pool participants*, *Behavior research methods* (2015) 1–8.

[116] R. E. Schumacker and R. G. Lomax, *A beginner's guide to structural equation modeling*. Psychology Press, 2004.

[117] J. Schaffer, T. Höllerer, and J. O'Donovan, *Hypothetical recommendation: A study of interactive profile manipulation behavior for recommender systems.*, in *FLAIRS Conference*, pp. 507–512, Citeseer, 2015.

[118] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, *Grouplens: An open architecture for collaborative filtering of netnews*, in *Proceedings of ACM CSCW'94 Conference on Computer-Supported Cooperative Work*, pp. 175–186, 1994.

[119] B. M. Sarwar, J. A. Konstan, A. Borchers, J. Herlocker, B. Miller, and J. Riedl, *Using filtering agents to improve prediction quality in the grouplens research collaborative filtering system*, in *Proceedings of the 1998 ACM conference on Computer supported cooperative work*, pp. 345–354, ACM, 1998.

[120] J. L. Herlocker, J. A. Konstan, L. G. Terveen, John, and T. Riedl, *Evaluating collaborative filtering recommender systems*, *ACM Transactions on Information Systems* **22** (2004) 5–53.

[121] S. Bostandjiev, J. O'Donovan, and T. Höllerer, *Tasteweights: a visual interactive hybrid recommender system*, in *RecSys* (P. Cunningham, N. J. Hurley, I. Guy, and S. S. Anand, eds.), pp. 35–42, ACM, 2012.

[122] K. McCarthy, J. Reilly, L. McGinty, and B. Smyth, *Experiments in dynamic critiquing*, in *IUI '05: Proceedings of the 10th international conference on Intelligent user interfaces*, (New York, NY, USA), pp. 175–182, ACM Press, 2005.

[123] R. Rafter, K. Bradley, and B. Smyth, *Passive profiling and collaborative recommendation*, in *Proceedings of the 10th Irish Conference on Artificial Intelligence and Cognitive Science, Cork, Ireland*, Artificial Intelligence Association of Ireland (AAAI Press), 1999.

[124] C. Boutilier, R. S. Zemel, and B. Marlin, *Active collaborative filtering*, in *In Proceedings of the Nineteenth Annual Conference on Uncertainty in Artificial Intelligence*, pp. 98–106, 2003.

[125] Y. Koren, R. Bell, and C. Volinsky, *Matrix factorization techniques for recommender systems*, *Computer* **42** (Aug., 2009) 30–37.

[126] R. Sinha and K. Swearingen, *The role of transparency in recommender systems*, in *CHI '02 extended abstracts on Human factors in computing systems*, pp. 830–831, ACM Press, 2002.

[127] J. O'Donovan, B. Smyth, B. Gretarsson, S. Bostandjiev, and T. Höllerer, *Peerchooser: visual interactive recommendation*, in *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, (New York, NY, USA), pp. 1085–1088, ACM, 2008.

[128] B. P. Knijnenburg, S. Bostandjiev, J. O'Donovan, and A. Kobsa, *Inspectability and control in social recommenders*, in *RecSys* (P. Cunningham, N. J. Hurley, I. Guy, and S. S. Anand, eds.), pp. 43–50, ACM, 2012.

[129] K. Verbert, D. Parra, P. Brusilovsky, and E. Duval, *Visualizing recommendations to support exploration, transparency and controllability*, in *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, IUI '13, (New York, NY, USA), pp. 351–362, ACM, 2013.

[130] M. Buhrmester, T. Kwang, and S. D. Gosling, *Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data?*, *Perspectives on Psychological Science* **6** (2011), no. 1 3–5, [http://pps.sagepub.com/content/6/1/3.full.pdf+html].

[131] G. Paolacci, J. Chandler, and P. G. Ipeirotis, *Running experiments on amazon mechanical turk*, *Judgment and Decision Making* **5** (2010) 411–419.

[132] B. Mobasher, R. Burke, R. Bhaumik, and C. Williams, *Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness*, *ACM Trans. Inter. Tech.* **7** (2007), no. 4 23.

[133] J. Schaffer, P. Giridhar, D. Jones, T. Höllerer, T. Abdelzaher, and J. O'Donovan, *Getting the message?: A study of explanation interfaces for microblog data analysis*, in *Proceedings of the 20th International Conference on Intelligent User Interfaces*, pp. 345–356, ACM, 2015.

[134] K. McCarthy, J. Reilly, L. McGinty, and B. Smyth, *Experiments in dynamic critiquing*, in *Proceedings of the 10th International Conference on Intelligent User Interfaces*, IUI '05, (New York, NY, USA), pp. 175–182, ACM, 2005.

[135] J. Zhang and P. Pu, *A comparative study of compound critique generation in conversational recommender systems*, in *In Proceedings of the 4th International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH2006*, pp. 234–243, Springer, 2006.

[136] J. L. Herlocker, J. A. Konstan, and J. Riedl, *Explaining collaborative filtering recommendations*, in *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, CSCW '00, (New York, NY, USA), pp. 241–250, ACM, 2000.

[137] N. Tintarev, J. O'Donovan, P. Brusilovsky, A. Felfernig, G. Semeraro, and P. Lops, *Recsys'14 joint workshop on interfaces and human decision making for recommender systems*, in *Eighth ACM Conference on Recommender Systems, RecSys '14, Foster City, Silicon Valley, CA, USA - October 06 - 10, 2014*, pp. 383–384, 2014.

[138] N. Tintarev and J. Masthoff, *Effective explanations of recommendations: user-centered design*, in *Proceedings of the 2007 ACM Conference on Recommender Systems, RecSys 2007, Minneapolis, MN, USA, October 19-20, 2007*, pp. 153–156, 2007.

[139] P. Giridhar, M. T. A. Amin, T. F. Abdelzaher, L. M. Kaplan, J. George, and R. K. Ganti, *Clarisense: Clarifying sensor anomalies using social network feeds*, in *2014 IEEE International Conference on Pervasive Computing and Communication Workshops, PerCom 2014 Workshops, Budapest, Hungary, March 24-28, 2014*, pp. 395–400, 2014.

[140] E. M. Daly, F. Lecue, and V. Bicer, *Westland row why so slow?: fusing social media and linked data sources for understanding real-time traffic conditions*, in *Proceedings of the 2013 international conference on Intelligent user interfaces*, pp. 203–212, ACM, 2013.

[141] P. Giridhar, M. T. Amin, T. Abdelzaher, L. Kaplan, J. George, and R. Ganti, *Clarisense: Clarifying sensor anomalies using social network feeds*, in *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on*, pp. 395–400, IEEE, 2014.

[142] J. S. Yi, Y. ah Kang, J. T. Stasko, and J. A. Jacko, *Toward a deeper understanding of the role of interaction in information visualization*, *Visualization and Computer Graphics, IEEE Transactions on* **13** (2007), no. 6 1224–1231.

[143] K. S. Thomson and D. M. Oppenheimer, *Investigating an alternate form of the cognitive reflection test*, *Judgment and Decision Making* **11** (2016), no. 1 99.

[144] M. R. Endsley, *Direct measurement of situation awareness: Validity and use of sagat*, *Situation awareness analysis and measurement* **10** (2000).

[145] F. M. Harper and J. A. Konstan, *The movielens datasets: History and context*, *ACM Transactions on Interactive Intelligent Systems (TiiS)* **5** (2016), no. 4 19.

[146] J. J. Jung, *Attribute selection-based recommendation framework for short-head user group: An empirical study by movielens and imdb*, *Expert Systems with Applications* **39** (2012), no. 4 4049–4054.

[147] B. N. Miller, I. Albert, S. K. Lam, J. A. Konstan, and J. Riedl, *Movielens unplugged: experiences with an occasionally connected recommender system*, in *Proceedings of the 8th international conference on Intelligent user interfaces*, pp. 263–266, ACM, 2003.

[148] Y. Koren and R. Bell, *Advances in collaborative filtering*, in *Recommender systems handbook*, pp. 145–186. Springer, 2011.

[149] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl, *An algorithmic framework for performing collaborative filtering*, in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 230–237, ACM, 1999.

[150] S. Bostandjiev, J. O'Donovan, and T. Höllerer, *Tasteweights: a visual interactive hybrid recommender system*, in *Proceedings of the sixth ACM conference on Recommender systems*, pp. 35–42, ACM, 2012.

[151] C. A. Gomez-Uribe and N. Hunt, *The netflix recommender system: Algorithms, business value, and innovation*, *ACM Transactions on Management Information Systems (TMIS)* **6** (2016), no. 4 13.

[152] T. Raykov, *Structural models for studying correlates and predictors of change*, *Australian Journal of Psychology* **44** (1992), no. 2 101–112.

[153] Y. Benjamini and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing*, *Journal of the royal statistical society. Series B (Methodological)* (1995) 289–300.

[154] R. A. Cribbie, *Multiplicity control in structural equation modeling*, *Structural Equation Modeling* **14** (2007), no. 1 98–112.

[155] Z. Liu and J. Stasko, *Mental models, visual reasoning and interaction in information visualization: A top-down perspective*, *IEEE transactions on visualization and computer graphics* **16** (2010), no. 6 999–1008.

[156] N. Tintarev and J. Masthoff, *Designing and evaluating explanations for recommender systems*, in *Recommender Systems Handbook*, pp. 479–510. Springer, 2011.

[157] A. D. Andre and C. D. Wickens, *When users want what's not best for them*, *Ergonomics in Design: The Quarterly of Human Factors Applications* **3** (1995), no. 4 10–14.

[158] A. Fernandes and P. Ø. Braarud, *Exploring measures of workload, situation awareness, and task performance in the main control room*, *Procedia Manufacturing* **3** (2015) 1281–1288.

[159] K. C. Hendy, *Situation awareness and workload: Birds of a feather?* DTIC Document, 1995.

[160] J. W. Payne, J. R. Bettman, and E. J. Johnson, *The adaptive decision maker*. Cambridge University Press, 1993.

[161] T. Lejarraga, V. Dutt, and C. Gonzalez, *Instance-based learning: A general model of repeated binary choice*, *Journal of Behavioral Decision Making* **25** (2012), no. 2 143–153.

[162] D. W. Stephens, C. M. McLinn, and J. R. Stevens, *Discounting and reciprocity in an iterated prisoner's dilemma*, *Science* **298** (2002), no. 5601 2216–2218.

[163] G. Ainslie, *Breakdown of will*. Cambridge University Press, 2001.

[164] H. R. Varian, T. C. Bergstrom, and J. E. West, *Intermediate microeconomics*, vol. 4. Norton New York, 1996.