Human Cognition and Decision-making in Recommender Systems

Leave Authors Anonymous
for Submission
City, Country
e-mail address

Leave Authors Anonymous
for Submission
City, Country
e-mail address

Leave Authors Anonymous
for Submission
City, Country
e-mail address

ABSTRACT

Designers of recommender systems must make trade-offs in interface design that have consequences for their users and affect system adoption. Additionally, there is little understanding of how properties such as explanation, control, and error of a recommender system will impact human decision making. We present results of a study (N=526) in which these properties were manipulated to examine impacts on decision satisfaction and adherence to recommendations. Users were allowed to freely interact with a simple browser or a tool that provided automated recommendations. Our analysis produced three novel findings: 1) objective measurement of user understanding of the algorithm, cognitive reflection, and domain knowledge helped to explain an additional 40% of variance (R^2) for decision outcomes over a subjective user experience model, 2) user understanding of the recommendation algorithm was the strongest predictor that recommendations would be accepted, and 3) a statistical model showed 11% reduction in knowledge of the underlying dataset for those who interacted heavily with the recommender system.

ACM Classification Keywords

H.1.2 User/Machine Systems: Human factors; H.5.2 User Interfaces: Evaluation/methodology; H.5.2 User Interfaces: User-centered design

INTRODUCTION

Users frequently make active decisions about which computational tools they use to satisfy their information criteria and they abandon tools that they perceive as untrustworthy or ineffective. Tools that provide sophisticated algorithms are often effective but are simultaneously opaque; meanwhile, simple tools are transparent and predictable but limited in their usefulness – here we refer to this as the simplicity/complexity tradeoff. Designers have responded to this tradeoff by adding features (such as explanations) that improve transparency of complex algorithms or by improving the effectiveness of simple algorithms (such as adding personalization to keyword search). Unfortunately, human attention is limited, which means that using screen space to accommodate explanations also means that cognitive bandwidth must be used as well. Simultaneously, improving the performance of algorithms typically makes the underlying computations more complex, reducing predictability, increasing potential mistrust, and perhaps resulting in user performance degradation [30][37]. Ideally, designers of information retrieval systems would intimately know how complexity and transparency of systems would affect human cognition. However, a comprehensive list of factors that affect decision making in human-agent interaction (HAI) are not yet known.

The simplicity/complexity tradeoff is important for the design of recommender systems. Builders of these systems find that establishing trust in their predictions is a difficult challenge [17], especially when simple tools that perform transparent information filtering are often effective and ubiquitously available. For example, traditional query and filter tools are central on popular sites (Amazon.com, Newegg.com, etc.) that provide browsing interfaces for item catalogs. In some cases, such as in Google search, personalized, profile-based recommendations are integrated with keyword-based search. In other cases, search and recommendation are separate tools that the user must choose between, for example, IMDb presents these features on separate pages. Users frequently make active choices when interacting with these services about which tools to use. Differences between tools can have an impact on user perceptions of underlying data and the accumulation of expertise and domain knowledge, which can, in turn, affect the future choices of those users.

Many information filtering tools are designed to support good decision-making by giving users the ability to quickly summarize large amounts of information. Despite this, recommender systems are typically evaluated in terms of rating accuracy or user experience metrics such as trust and satisfaction. More recent evaluation models include factors such as perceived usefulness, perceived control, perceived transparency, and use intention [26][38]. However, many studies do not allow users to freely choose new items and do not provide alternatives to the recommender system being evaluated, which limits their impact on real-world situations. Users only benefit from recommender systems if new and satisfying selections are made. Therefore, separating user experience from satisfaction with selected items can provide more insight into user behavior.

The subjective quality of the item selections made by a user in a catalog browsing task, or the *decision satisfaction*, is

Figure 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 dataset 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.

likely to be affected by many factors. Some of these factors may transfer easily across task domains, such as a user's personal characteristics, the presence of explanation or control features, how the participant interacts with the system, and how interaction affects the user's cognitive load during the selection task. Other factors might be specific to task/domain parameters but have parallels in other tasks, such as a user's domain knowledge, situation awareness (SA [9]), and domain-specific experience. Finally, some factors are very specific to task/domain parameters and task complexity [36]. In this work, we evaluate users during a movie selection task where both query-driven and recommendation tools are available and we attempt to explain variability in decision making using the most general factors possible. Due to the presence of a simple alternative to the recommender, adherence to recommendations can also be evaluated. We combine decision-making metrics such as domain knowledge, cognitive load, cognitive reflection, and situation awareness (SA) [11] with user experience metrics, which remain popular in recommender system evaluations. Within our use case of movie selection, this leads to the following research questions:

1. Which factors best explain variability in decision making during a movie selection task? How much variability can be explained?
2. What role does user understanding of the recommender algorithm play in a movie selection task?
3. Does user domain knowledge affect decision-making behavior? How does it change as a function of the tool that is chosen?
4. How does transparency, control, and error from a recommender affect the HAI system?

In order to evaluate when a recommender system may lead its users to bad decisions (thus answering questions 1 and 4 above), two steps must be taken in experiment protocol. First, participants must *freely* select (and give feedback about) items of interest in order to separate decision satisfaction from recommender accuracy and user experience. Second, a *simple, non-recommender* alternative to the recommender system being evaluated must be available to participants. This way, properties of recommender systems that increase adherence can be determined through observing participant behavior rather than the relying on reported use intention.

In this experiment, two tools –a simple browser and a complex collaborative filtering algorithm– were developed to evaluate the above hypotheses. Participants (N=526) accessed the system and were tasked with finding 5-7 new movies, which they added to a watch list. Users could freely choose between the filtering or recommendation tools. Figure 1 shows an example of the system during a participant session.

RELATED WORK

Human-agent interaction (HAI) has been a field of study since the early work on the MYCIN system (see Shortliffe et al

[44]). Additionally, explanation and control from automated algorithms has been studied as early as 1975 [45]. For this experiment, we researched factors that have been shown to affect decision making in the context of HAI.

User Experience, Explanation, and Control

Within the recommender systems research community, there is an increasing understanding of the need for user-centered evaluations [32][43]. Recent keynote talks [7] and workshops [34] have helped to highlight the importance of this topic. Of note is the user experience framework created by Pu [38], 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 participant-reported use intention. Knijnenburg et al [26][28] 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.

Similarly, it been recognized that 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 [21]. 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 [1]. Explanations have been noted to increase transparency, scrutability, trust, effectiveness, efficiency, and satisfaction. Similar findings have also been earlier in expert systems research, for instance, the work by Gregor [16]. Additionally, research on textual and visual explanations in recommender systems has been evaluated in wide range of domains (varying from financial advice [12] to movies [48]).

Decision-Making in Recommender Systems

Although early recommender research identified the importance of evaluating decision making [22], the recommender systems community has moved away from accuracy measures towards understanding user experience [32][43]. Moreover, evaluation studies have focused on understanding adoption of recommender systems as a whole (rather than examining adoption of individual recommendations), rely on participant-reported use intention, and often compare candidates with other, similar recommender systems (for example, Jones et al [23]). This evaluation methodology has been successful at relating system parameters such as diversity, serendipity, and accuracy/ranking measures to user experience, but the connection between user experience (or even accuracy) and decision making in recommender systems is still not well understood. Recently, it has been posited that different psychological factors can influence the decision making of recommender system users [6] and there is an increase in pairing recommender systems theory with decision making theory (for instance, see the work on leveraging user biases [14] and information disclosure [27]. When compared with other work, we focus on evaluating the most general factors possible that can explain decision-making variance and their relationship with explanation, control, and recommendation error.

Domain Knowledge

Users exhibit different levels of expertise when filtering a parameter space in a browsing task. We can imagine a hypothetical skilled user could construct queries that would immediately isolate the part of the "movie space" that contains the items they are interested in. Meanwhile, a user that has little expertise in movies would greatly benefit from following a recommender's advice. Refer to Arnold et al [2] for an examination of the differences in novices and experts.

Domain knowledge can be captured in a laboratory setting using testing methodologies. In visual analytics, insight has come to be known as "an individual observation about the data by the participant", or "a unit of discovery" [40]. Thus, domain knowledge in this study was estimated by developing questionnaires based around a collection of insights related to movie metadata distributions.

Situation Awareness and Cognitive Reflection

User interaction with recommendation systems may also be significantly impacted by the user's mental model of how a recommender operates. For this, we consider situation awareness (SA) theory [9] and its measurement approach [8]. The theory of SA has already been applied to the problem of agent transparency, see Chen et al [5]. Chen's theory, SA-based Agent Transparency (SAT), builds on the theory of SA. Chen refers to 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. In this work, SAT measures are used but are referred to interchangeably with situation awareness (SA), since global SA is not measured.

Work on attention and cognitive reflection by Daniel Kahneman [25] 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 decision making [13][49][50].

Cognitive Load

The term "cognitive load" originates from education theory [47] and problem solving [46]. It is defined as a "multidimensional construct representing the load that performing a particular task imposes on the learner's cognitive system." Greater cognitive effort by users of systems leads to increased error when performing tasks. For example, when interacting with a recommender system, a cluttered user interface may lead to frustration and thus bad decisions with respect to the items being selected from the catalog, or altogether abandonment of the recommender system. Paas [35] surveys numerous methods of measuring cognitive load during participant tasks, noting that cognitive load can be assessed by measuring mental load and mental effort. Participant self-reported rating scale techniques are reliable and have been successful in the past.

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[1] called *Movie Miner* (Figure 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 dataset. 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" dataset 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 [18][24][33]. Due to recommender speed limitations, the dataset was randomly sampled for 4 million ratings, rather than the full 20 million.

Generating Recommendations

A traditional collaborative filtering approach was chosen for the system. Details for this can be found in Resnick et al [39]. 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 [29]. 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 dataset: Herlocker damping and rating normalization[2].

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

$$sim(u,v) = \frac{\hat{u}\cdot\hat{v}}{||\hat{u}||||\hat{v}||} = \frac{\sum_i \hat{r_{ui}} \hat{r_{vi}}}{\sqrt{\sum_i \hat{r_{ui}}^2}\sqrt{\sum_i \hat{r_{vi}}^2}} \quad (1)$$

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 [20], which penalizes popular items. Finally, the predicted rating, r_{ui}, $i \notin I_u$, was calculated as:

$$r_{ui} = \mu_u + k \sum_{v \in U} sim(u,v) \hat{r_{vi}} \quad (2)$$

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.

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 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.

[1]The reviewers can access a video demo of the system in the supplementary materials

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

EXPERIMENT DESIGN

To answer the research questions given in the introduction, we formed the following hypotheses:

- H_1: domain knowledge affects decision satisfaction
- H_2: recommender-based situation awareness affects decision satisfaction
- H_3: user experience affects decision satisfaction
- H_4: cognitive load affects decision satisfaction
- H_5: interaction behavior (browser or recommender side) affects decision satisfaction
- H_6: interaction with the browser affects insight shift
- H_7: interaction with the recommender affects insight shift
- H_8: explanations from the recommender cause improved insight
- H_9: control over the recommender causes improved decision satisfaction
- H_{10}: explanations from the recommender causes improved decision satisfaction
- H_{11}: recommendation error causes reduced decision satisfaction

Hypotheses 1-5 address the first and second research questions, hypotheses 6-8 address the third research question, and hypotheses 9-11 address the fourth research question. Note that when we state that "X affects Y," we posit a relationship such that the value of X predicts the value of Y. Hypotheses 8-11 posit causality that this experiment can inform due to the between-subjects manipulation.

Hypotheses 1-8 were tested by building a statistical model on participant data that includes measurements for the HAI factors that needed to be tested. 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. Hypotheses 9-11 were tested through a between-subjects manipulation that omitted or included the various features to be tested, which is described below.

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 noise 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 [3][26][42] 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 dataset. The error-free recommender achieved an MAE of 0.144, while the noisy version did considerably worse at 0.181 (nearly a 26% difference).

Noise 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.

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

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
Cog. Reflection	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
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 1. 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.

that originated from the recommender side of the interface. 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 Table 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 [38]) 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 unidimensional "good" or "bad" impression of the recommender) and collapsing the items onto one factor 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 Table 1 had a co-variance higher than 0.5 which indicated good discriminant validity between the factors.

We used a SAGAT-style freeze [10] during the movie selection task to assess recommender-based SA, which showed 9 questions related to ground truth of recommender behavior. Domain knowledge was measured twice - before and after the user finished interacting with Movie Miner. The difference in score between the two tests is referred to as **Insight Shift**. 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 domain knowledge and SA and the variance was fixed to the variance of the sample population. Both the SA and domain knowledge metrics are shown in Table 2.

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 [4] has found that the quality of data collected from MTurk

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?
Domain Know. (and Insight Shift) 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 2. Factors determined by participant responses to domain knowledge and situation awareness questions. Multiple choice answers were given. Domain knowledge was measured at the beginning of the study and insight shift was measured by repeating the domain knowledge questionnaire at the end of the study to see which questions had changed. Situation awareness was measured 8 minutes into the study, during the watchlist phase.

is comparable to what would be collected from laboratory experiments [19]. 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[3]. Items related to trust propensity, movie expertise, and cognitive reflection (also, see Toplak et al [49]) were collected during this phase using the question items shown at the top of Table 1. Questions related to domain knowledge 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 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 domain knowledge 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 detect an insight shift 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 [15].

Evaluating the HAI Model

The measurement framework consists of the eight factors listed in Tables 1 and 2, the basic observed variables (browser interaction, recommender interaction, adherence), and the independent variables (explanation, control, noise). Our analysis of the measurement framework had two goals: first, to compare the measurement framework with a user-experience centric approach (similar to [26] and [38]) when explaining decision making; and second, to identify which factors are most important so that superfluous measurements can be trimmed from future experiments. While this could be gleaned indirectly from examining regression (β) coefficients in the final fitted SEM, we found it more informative to compare several different SEMs that were constructed on the participant data:

- **Black Box**: only independent variables: explanation, control, noise, and their interaction effects
- **User Profiling**: independent variables + trust propensity, movie expertise, domain knowledge, cognitive reflection
- **Behavior**: independent variables + recommender interaction, browser interaction
- **SA**: independent variables + SA

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

Regressand	Regression (←)	β	P(>|z|)
User Experience R^2=0.17	← Trust Propensity	0.35	***
	← Domain Know.	-0.08	*
	← Noise	-0.17	***
Cognitive Load R^2=0.03	← Control	0.12	**
	← Trust Propensity	-0.14	**
SA R^2=0.14	← Trust Propensity	-0.12	**
	← Cog. Reflection	0.25	***
	← Explanation	0.08	0.051
	← Browser Int.	0.17	***
Browser Int. R^2=0.03	← Trust Propensity	-0.15	**
	← Domain Know.	0.09	*
Recommender Int. R^2=0.09	← Movie Expertise	-0.16	***
	← Domain Know.	0.17	***
	← Control	0.19	***
Adherence R^2=0.09	← User Experience	0.10	*
	← Control	0.10	*
	← SA	0.23	***
	← Browser Int.	-0.17	***
Insight Shift R^2=0.05	← SA	0.09	0.052
	← Recommender Int.	-0.11	*
	← User Experience	0.10	*
	← Cog. Reflection	-0.13	*
Decision Satisfaction R^2=0.23	← Trust Propensity	0.2	***
	← User Experience	0.29	***
	← Browser Int.	0.07	0.085
	← Recommender Int.	-0.20	***
	← Explanation	0.12	*
	← Control	0.15	**
	← Expl. x Control	-0.20	**

Table 3. Regressions in the fitted all-factor SEM, which attempts to explain decision satisfaction, insight shift, and adherence to recommendations. Variables on the left are explained by variables on the right. β refers to the regression coefficient (effect size). All latent and observed variables were standardized between 0 and 1. Model fit: $N = 526$ with 99 free parameters = 5 participants per free parameter, $RMSEA = 0.045$ ($CI : [0.041, 0.048]$), $TLI = 0.94, CFI = 0.94$ over null baseline model, $\chi^2(712) = 1454.963$. Significance levels for this table: *** p < .001, ** p < .01, * p < .05. Covariances are shown in Table 4.

- **User Experience (UXP)**: independent variables + user experience as shown in Table 1
- **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 1

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).

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.

Covariance (↔)	Beta	P(>|z|)
Trust Propensity ↔ Movie Expertise	0.46	***
Trust Propensity ↔ Cog. Reflection	-0.23	***
Trust Propensity ↔ Domain Know.	-0.11	**
Movie Expertise ↔ Cog. Reflection	-0.16	***
Cog. Reflection ↔ Domain Know.	0.29	***
User Experience ↔ Cognitive Load	-0.54	***

Table 4. Covariances in the fitted model. Regressions are shown in Table 3. User experience and cognitive load were negatively correlated. A covariance was tested between decision satisfaction and adherence but none was found.

Hypothesis	Support?
H_1 DK affects DS	Yes, M
H_2 SA affects DS	No
H_3 UXP affects DS	Yes
H_4 Cognitive load affects DS	No
H_5 interaction behavior affects DS	Yes
H_6 browser interaction affects insight	No
H_7 rec. interaction affects insight	Yes
H_8 explanations cause increased insight	Yes, M
H_9 control causes increased DS	Yes
H_{10} explanation causes improved DS	Yes
H_{11} rec. error causes decreased DS	Yes, M

Table 5. Support for hypotheses. UXP = User Experience, DS = Decision Satisfaction, DK = Domain Knowledge, SA = Situation Awareness, M = supported via mediation.

All-Factor SEM

Regressions in the SEM and fit of the all-factor model are shown in Table 3. This SEM is used to evaluate the hypotheses shown in Table 5. Model co-variances are shown in Table 4. An illustration of the SEM was omitted due to visual complexity. The model was built using R 3.0.3, lavaan 0.5-17.

Support for Hypotheses

The all-factor SEM in Table 3 was used to evaluate the hypotheses shown in Table 5. First, domain knowledge corresponded to slightly decreased user experience, which was a factor in decision satisfaction (H_1 accepted through mediation and H_3 accepted). Second, SA did not affect decision satisfaction in any model test (H_2 rejected). Next, our model did not indicate any significance between cognitive load and decision satisfaction (H_4 rejected). However, an alternative model that does not allow user experience to load onto decision satisfaction shows a significant relationship with cognitive load (this is explored further in the discussion section). Next, interaction with the recommender corresponded to a significant decrease in decision satisfaction and interaction with the browsing tool corresponded to a minor increase in decision satisfaction (H_5 accepted). Browser interaction did not affect insight but interaction with the recommender corresponded to a decrease in insight (H_6 rejected and H_7 accepted). Next, explanations improved insight through a mediation effect by

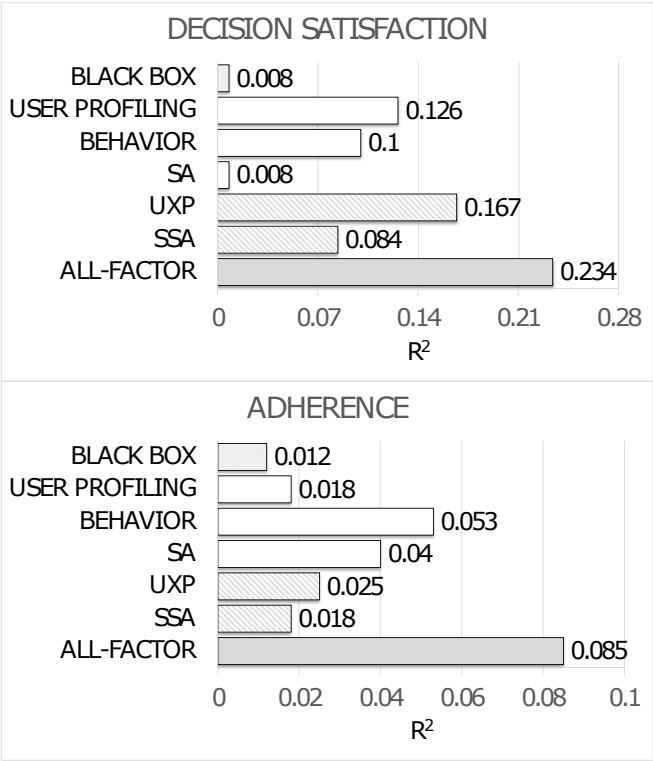

Figure 2. An evaluation of the factors that explain variance in decision satisfaction and adherence. R^2 indicates how much each model "determines" each variable, with 1.0 indicating perfect fit and 0.0 indicating no fit/complete noise. The all-factor model explains 40% more decision variance than the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

SA (H_8 accepted). Both control and explanation improved decision satisfaction (H_9 and H_{10} accepted). Finally, recommendation error negatively affected decision satisfaction by reducing user experience (H_{11} accepted).

Comparison of HAI Models

R^2 values of all evaluated HAI models are shown in Figure 2. The all-factor model explained the most decision satisfaction and the most adherence. Each component of the all-factor was successful at explaining at least one of the observed outcomes in the study. Specifically, trust propensity, interaction behavior, and user experience feedback were all somewhat successful at explaining decision satisfaction and more successful when combined. Moreover, observed interaction behavior and SA were both somewhat successful at explaining adherence. Finally, note that the black box model almost completely fails at explaining decision satisfaction. 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 ($RMSEA = 0.17$).

DISCUSSION OF RESULTS

Here we discuss four implications of the results described in this work. Based on our findings, we highlight some suggestions for future research and make design suggestions for recommender systems.

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 3 and Figure 2. 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 [24]).

Second, splitting user experience into SSA (similar to the ResQue framework [38] and the model in Knijnenburg et al [26]) decreased model fit, despite each sub-aspect (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[4] 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 discriminative (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 domain knowledge, SA, 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

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

experiment would have been missed if these measures had been omitted.

2. Users that understood how the recommender operated were also more likely to adopt recommendations. SA had the highest positive impact on adherence with a β coefficient of 0.23. 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 SA 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. Increased interaction with the recommender correlated with decreased user domain knowledge and decision satisfaction. Insight shift was measured by taking the difference between the pre- and post- domain knowledge tests. No significant effect was observed when considering browser interaction and, in general, insight did not change significantly between the pre- and post-tests (4.5 vs. 4.3). 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 [31][51]. 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. Note that domain knowledge was maintained when SA was high (i.e. high SA might be considered a "shield" against the recommender's "biasing" effect). Furthermore, explanations increased SA, which in turn helped to maintain domain knowledge. A similar effect was reported in Schaffer et al [41].

Interacting with the recommender also decreased decision satisfaction. Examination of participant behavior showed that 45% of recommender actions were rating actions and the other 55% were filter or "Not Interested" actions. Furthermore, models that attempted to use more nuanced interaction data found that both rating and filter actions on the recommender side were equally responsible for the decrease in decision satisfaction. A possible explanation for this is that users that heavily leveraged filters and provided more ratings were harder to satisfy. An alternative explanation is that users that rated a lot of movies from the recommender's list may have steered their profile towards a very "centralized" location in the collaborative filtering similarity space, which may have decreased recommendation diversity.

4. Explanation, control, and recommendation error steered the decision system towards different outcomes. Explanation played two roles. First, explanation improved SA 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 correlated with increased SA but correlated with 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 SA 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 recommendations 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 SA. 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. Our data 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.

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) objective measurement of user understanding of the algorithm, cognitive reflection, and domain knowledge helped to explain an additional 40% of variance (R^2) for decision satisfaction over a subjective user experience model, 2) recommender SA was the most significant variable affecting adherence to recommendations, 3) interacting with a recommender caused a domain knowledge 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.

REFERENCES

1. Inspection mechanisms for community-based content discovery in microblogs.
2. 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*, pages 79–97, 2006.
3. 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*, pages 35–42. ACM, 2012.
4. 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(1):3–5, 2011.
5. J. Y. Chen, K. Procci, M. Boyce, J. Wright, A. Garcia, and M. Barnes. Situation awareness-based agent transparency. Technical report, DTIC Document, 2014.
6. L. Chen, M. de Gemmis, A. Felfernig, P. Lops, F. Ricci, and G. Semeraro. Human decision making and recommender systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 3(3):17, 2013.
7. E. H. Chi. Blurring of the boundary between interactive search and recommendation. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, pages 2–2. ACM, 2015.
8. M. R. Endsley. Measurement of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37(1):65–84, 1995.
9. M. R. Endsley. Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37(1):32–64, 1995.
10. M. R. Endsley. Direct measurement of situation awareness: Validity and use of sagat. *Situation awareness analysis and measurement*, 10, 2000.
11. M. R. Endsley and D. J. Garland. *Situation awareness analysis and measurement*. CRC Press, 2000.
12. A. Felfernig, E. Teppan, and B. Gula. Knowledge-based recommender technologies for marketing and sales. *Int. J. Patt. Recogn. Artif. Intell.*, 21:333–355, 2007.
13. S. Frederick. Cognitive reflection and decision making. *The Journal of Economic Perspectives*, 19(4):25–42, 2005.
14. J. Freyne, S. Berkovsky, and G. Smith. Rating bias and preference acquisition. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 3(3):19, 2013.
15. C. A. Gomez-Uribe and N. Hunt. The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)*, 6(4):13, 2016.
16. S. Gregor and I. Benbasat. Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS quarterly*, pages 497–530, 1999.
17. 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*, pages 305–308. ACM, 2014.
18. F. M. Harper and J. A. Konstan. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(4):19, 2016.
19. D. J. Hauser and N. Schwarz. Attentive turkers: Mturk participants perform better on online attention checks than do subject pool participants. *Behavior research methods*, pages 1–8, 2015.
20. 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*, pages 230–237. ACM, 1999.
21. J. L. Herlocker, J. A. Konstan, and J. Riedl. Explaining collaborative filtering recommendations. In *ACM conference on Computer supported cooperative work*, pages 241–250, 2000.
22. J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53, 2004.
23. N. Jones and P. Pu. User technology adoption issues in recommender systems. In *Proceedings of the 2007 Networking and Electronic Commerce Research Conference*, number HCI-CONF-2008-001, pages 379–394, 2007.
24. 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(4):4049–4054, 2012.
25. D. Kahneman. *Attention and effort*. Citeseer, 1973.
26. 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*, pages 43–50. ACM, 2012.
27. 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(3):20, 2013.
28. B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, and C. Newell. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22(4-5):441–504, 2012.
29. Y. Koren and R. Bell. Advances in collaborative filtering. In *Recommender systems handbook*, pages 145–186. Springer, 2011.

30. 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(1):50–80, 2004.

31. 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(6):999–1008, 2010.

32. S. M. McNee, J. Riedl, and J. A. Konstan. Being accurate is not enough: How accuracy metrics have hurt recommender systems. In *Extended Abstracts of the 2006 ACM Conference on Human Factors in Computing Systems (CHI 2006)*, 2006.

33. 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*, pages 263–266. ACM, 2003.

34. 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 (intrs). In H. Werthner, M. Zanker, J. Golbeck, and G. Semeraro, editors, *RecSys*, pages 347–348. ACM, 2015.

35. 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(1):63–71, 2003.

36. J. W. Payne. Task complexity and contingent processing in decision making: An information search and protocol analysis. *Organizational behavior and human performance*, 16(2):366–387, 1976.

37. W. Payre, J. Cestac, and P. Delhomme. Fully automated driving impact of trust and practice on manual control recovery. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 58(2):229–241, 2016.

38. 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*, pages 157–164. ACM, 2011.

39. 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*, pages 175–186, 1994.

40. P. Saraiya, C. North, and K. Duca. An insight-based methodology for evaluating bioinformatics visualizations. *Visualization and Computer Graphics, IEEE Transactions on*, 11(4):443–456, 2005.

41. 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*, pages 345–356. ACM, 2015.

42. J. Schaffer, T. Höllerer, and J. O'Donovan. Hypothetical recommendation: A study of interactive profile manipulation behavior for recommender systems. In *FLAIRS Conference*, pages 507–512. Citeseer, 2015.

43. G. Shani and A. Gunawardana. Evaluating recommendation systems. In *Recommender systems handbook*, pages 257–297. Springer, 2011.

44. E. Shortliffe. *Computer-based medical consultations: MYCIN*, volume 2. Elsevier, 2012.

45. 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(4):303–320, 1975.

46. J. Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive science*, 12(2):257–285, 1988.

47. J. Sweller. Cognitive load theory, learning difficulty, and instructional design. *Learning and instruction*, 4(4):295–312, 1994.

48. N. Tintarev and J. Masthoff. Personalizing movie explanations using commercial meta-data. In *Adaptive Hypermedia*, 2008.

49. 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(7):1275–1289, 2011.

50. 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*, pages 1587–1592. Cognitive Science Society Austin, TX, 2013.

51. 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*, page 4. ACM, 2008.