Easy to Please: Separating User Experience from Choice Satisfaction

James Schaffer
US Army Research Laboratory
Playa Vista, California
james.a.schaffer20.civ@mail.mil

John O'Donovan
University of California Santa Barbara
Santa Barbara, California
jod@cs.ucsb.edu

Tobias Höllerer
University of California Santa Barbara
Santa Barbara, California
holl@cs.ucsb.edu

ABSTRACT

Recommender systems are evaluated based on both their ability to create a satisfying user experience and their ability to help a user make better choices. Despite this, quantitative evidence from previous research in recommender systems indicate very high correlations between user experience attitudes and choice satisfaction. This might imply invalidity in the measurement methodologies of these constructs, whereas they may not be measuring what researchers think they are measuring. To remedy this, we present a new methodology for the measurement of choice satisfaction. Part of our approach is to measure a user's "ease of satisfaction," or that user's natural propensity to be satisfied, which is measured using three different approaches. An (N=526) observational study is conducted wherein users browse a movie catalog. A factor analysis is done to assess the discriminant validity of our proposed choice satisfaction apparatus from user experience. A statistical analysis suggests that accounting for ease-of-satisfaction allows for a model of choice satisfaction that is not only discriminant, but independent, from user experience. This enables researchers to more objectively identify recommender system factors that lead users to good choices.

CCS CONCEPTS

• **Human-centered computing** → *User models*; *HCI design and evaluation methods*; • **Information systems** → *Personalization*;

KEYWORDS

Choice satisfaction, user experience, construct validity, recommender systems, user models, evaluation

ACM Reference Format:
James Schaffer, John O'Donovan, and Tobias Höllerer. 2018. Easy to Please: Separating User Experience from Choice Satisfaction. In *UMAP '18: 26th Conference on User Modeling, Adaptation and Personalization, July 8–11, 2018, Singapore, Singapore.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3209219.3209222

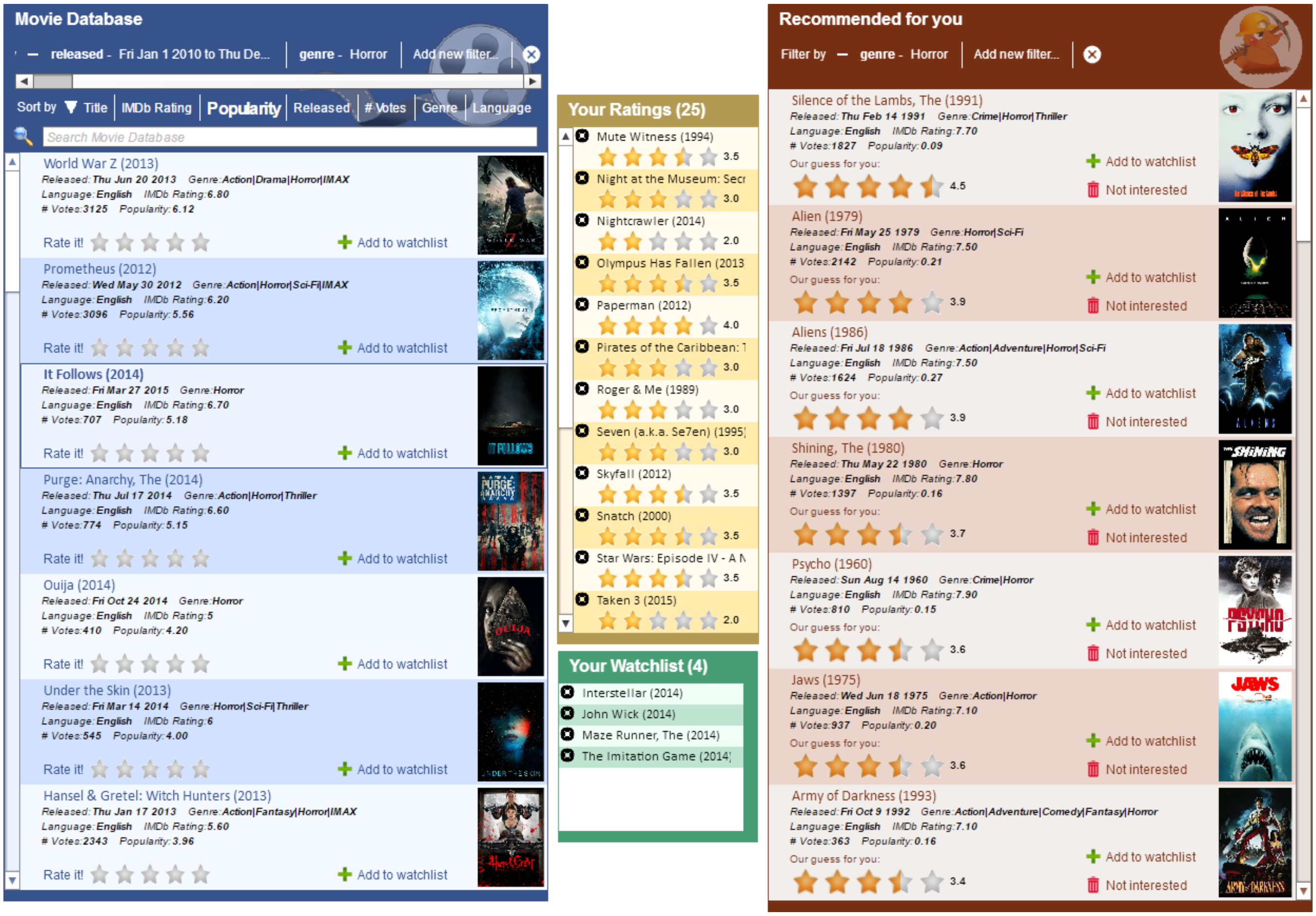
1 INTRODUCTION

One end goal of recommender systems is to assist in the decision-making process by assessing which items are relevant to a user. In order to evaluate a recommender's ability to satisfy this goal, recent research has moved towards quantifying both a user's choice satisfaction (CS) and user experience (UX) [21][28][29]. Based on common definitions of UX (subjective satisfaction with a recommender interface) and CS (a user's subjective satisfaction with a particular choice), these constructs should not be strongly correlated, but yet, this has been demonstrated [3][22]. A contributing issue is that CS and UX are measured in a subjective way due to difficulty in obtaining the ground truth. Factor analysis [35] has been used as a way to assess the validity of these self-reported metrics [22][28]. The UX/CS measurement process consists of exposing a user to recommendations and then obtaining the user's agreement (on a 1-7 Likert scale) with multiple question items that represent CS. Moreover, while some definitions of UX include only a "good-bad" scale (a single feeling) [13], quantification of subjectivity in recommender systems has moved to using multiple "sub-constructs" of attitudes, which are phrased as "user beliefs" [28] or "subjective system aspects" [21]. In this work, we will refer to these sub-constructs as UX attitudes.

Researchers should be concerned both about the use of self-reported UX/CS and the validity of fine-grained UX attitudes, for three reasons. First, although self-reported metrics may have predictive power, the content of the questions presented may not reflect what is actually measured. For example, the Dunning-Kruger test of self-reported expertise is more likely to indicate a test subject's *incompetence* rather than competence [8]. For this reason, it is important to rationalize what observable behavior or outcomes correlate with a proposed factor. Since we can infer high correlation between UX/CS from the work of Knijnenburg et al. [22], this may lead us to believe that self-assessed UX and CS may be affected by other situational or personal characteristics. For instance, users could simply be feeling good "in the moment" due to a sleek interface, which may incidentally lead researchers to believe they are making good decisions through inflated CS. Second, discriminant validity [9][16] appears to be an issue, since it has not been assessed in studies of recommender systems UX. This is more troublesome in light of large inter-attitude β coefficients reported in [22] (and to some extent, [28]). Finally, the predictive validity of many UX sub-constructs appears to be weak. This is because predicting attitudes with attitudes is not useful, especially when the effort required to collect these attitudes is equal. If behaviors can be predicted adequately using a single UX attitude, only that attitude need be measured.

UMAP '18, July 8–11, 2018, Singapore, Singapore

ACM ISBN 978-1-4503-5589-6/18/07…$15.00
https://doi.org/10.1145/3209219.3209222



World War Z (2013)
Released: Thu Jun 20 2013 Genre: Action|Drama|Horror|IMAX
Language: English IMDb Rating: 6.80
Votes: 3125 Popularity: 6.12

Prometheus (2012)
Released: Wed May 30 2012 Genre: Action|Horror|Sci-Fi|IMAX
Language: English IMDb Rating: 6.20
Votes: 3096 Popularity: 5.56

It Follows (2014)
Released: Fri Mar 27 2015 Genre: Horror
Language: English IMDb Rating: 6.70
Votes: 707 Popularity: 5.18

Purge: Anarchy, The (2014)
Released: Thu Jul 17 2014 Genre: Action|Horror|Thriller
Language: English IMDb Rating: 6.60
Votes: 774 Popularity: 5.15

Ouija (2014)
Released: Fri Oct 24 2014 Genre: Horror
Language: English IMDb Rating: 5
Votes: 410 Popularity: 4.20

Under the Skin (2013)
Released: Fri Mar 14 2014 Genre: Horror|Sci-Fi|Thriller
Language: English IMDb Rating: 6
Votes: 545 Popularity: 4.00

Hansel & Gretel: Witch Hunters (2013)
Released: Thu Jan 17 2013 Genre: Action|Fantasy|Horror|IMAX
Language: English IMDb Rating: 5.60
Votes: 2343 Popularity: 3.96

Your Ratings (25)

Mute Witness (1994)	3.5
Night at the Museum: Secr	3.0
Nightcrawler (2014)	2.0
Olympus Has Fallen (2013	3.5
Paperman (2012)	4.0
Pirates of the Caribbean: 1	3.0
Roger & Me (1989)	3.0
Seven (a.k.a. Se7en) (1995)	3.0
Skyfall (2012)	3.5
Snatch (2000)	3.5
Star Wars: Episode IV - A N	3.5
Taken 3 (2015)	2.0

Your Watchlist (4)

Interstellar (2014)
John Wick (2014)
Maze Runner, The (2014)
The Imitation Game (2014)

Recommended for you

Silence of the Lambs, The (1991)
Released: Thu Feb 14 1991 Genre: Crime|Horror|Thriller
Language: English IMDb Rating: 7.70
Votes: 1827 Popularity: 0.09
Our guess for you: 4.5

Alien (1979)
Released: Fri May 25 1979 Genre: Horror|Sci-Fi
Language: English IMDb Rating: 7.50
Votes: 2142 Popularity: 0.21
Our guess for you: 3.9

Aliens (1986)
Released: Fri Jul 18 1986 Genre: Action|Adventure|Horror|Sci-Fi
Language: English IMDb Rating: 7.50
Votes: 1624 Popularity: 0.27
Our guess for you: 3.9

Shining, The (1980)
Released: Thu May 22 1980 Genre: Horror
Language: English IMDb Rating: 7.80
Votes: 1397 Popularity: 0.16
Our guess for you: 3.7

Psycho (1960)
Released: Sun Aug 14 1960 Genre: Crime|Horror
Language: English IMDb Rating: 7.90
Votes: 810 Popularity: 0.15
Our guess for you: 3.6

Jaws (1975)
Released: Wed Jun 18 1975 Genre: Action|Horror
Language: English IMDb Rating: 7.10
Votes: 937 Popularity: 0.20
Our guess for you: 3.6

Army of Darkness (1993)
Released: Fri Oct 9 1992 Genre: Action|Adventure|Comedy|Fantasy|Horror
Language: English IMDb Rating: 7.10
Votes: 363 Popularity: 0.16
Our guess for you: 3.4

Figure 1: A screenshot of "Movie Miner," a faux movie discovery interface, which was used in the study.

This research is principally concerned with the statistical validity of CS measurement in recommender evaluation. A first step toward this is to measure CS in such a way that it is independent, or at least discriminant, from UX. We hypothesized that a person's "ease-of-satisfaction" is a contributor to the large correlations previously observed between UX and CS. Moreover, we suspect that past methodologies for measuring CS may have lead to low discriminant validity with UX attitudes, due to restrictive study designs. To address this, we conducted a user study (N=526) wherein users freely interacted with a recommendation interface to make a movie selection. CS and UX were measured. The proposed method for measuring CS accounts for baseline satisfaction, which is measured on a random selection of movies. Likert-scale questions are used, but feedback is taken for each individually selected movie and then parceled together. We also test whether users can self-assess their own ease-of-satisfaction as an easier-to-measure stand-in for baseline satisfaction. Finally, we examined the correlation between a user's average item rating and baseline satisfaction. In summary, our research questions are:

(1) What is the statistical validity of the recommender systems approach to modeling attitudes and UX?
(2) Can CS be measured in a way that makes it discriminant from UX?
(3) How can a user's personal ease-of-satisfaction be measured and does it predict both UX and CS?

2 RELATED WORK

Here we give a brief overview of methods for measuring CS in the psychology of consumer behavior, how to establish statistical validity, and details of UX and CS in recommender systems evaluation.

2.1 Measurement of CS

CS is studied in the psychology of consumer behavior (recent works include [7][20][37]). Satisfaction can be elicited, for instance, through user feedback on a Likert scale and studies often use longitudinal measures, since a person's initial estimate of their satisfaction with a particular product is suspect. CS can also be thought of as "projected" satisfaction - a feeling about a decision that was made today, but the consequences of which will not accrue until the future, at which point the product would be thoroughly evaluated. A valid concern is whether or not users can even accurately self-report their own CS at the time of choosing – an issue that is further confounded by recent research in the psychology of happiness, which indicates people may overestimate both past and future levels of happiness [38]. Additionally, people have a "sticky happiness baseline" [10], that is, positive or negative events can temporarily disturb a person's reported happiness level but it will always bounce back. This might imply that CS measurements taken before and after the moment of evaluation might be inflated and that baseline happiness might need to be accounted for. An ideal approach might be to measure the baseline state of a particular person before a stimulus is applied and then compare that with a measurement of CS that is taken longitudinally at the time of choice evaluation (in the case of movie selection, this could be just after the movie was viewed).

Establishing the statistical validity of a CS measurement in recommender systems is a difficult research challenge. First, predictive validity would need to be established. Since CS can be thought of as an attitude, the Theory of Reasoned Action [32] would suggest that this attitude would need to accurately predict behavior. For example, users self-reporting that they are completely satisfied with a choice in a movie should later rate that movie highly, or at least reflect that watching the movie was a good use of their time. Second, recommender interfaces can be affective [34] and features such as explanations can inflate a user's satisfaction [33]. For instance, a user could become frustrated with a system during interaction but still end up discovering a high quality item. The CS measurement thus should be statistically discriminant from any measurement of UX and ideally their measurements would be independent.

This research draws inspiration from the above observations and herein we model a user's ease-of-satisfaction to improve the measurement model for CS. A comparison of the effectiveness of longitudinal and immediate measurements of CS is left for future work.

2.2 Recommender Systems Evaluation

In recommender systems, CS and UX first explicitly converge in Knijnenburg et al. [22], which contains descriptions of multiple studies of UX in recommender systems. Here, CS is measured on three occasions in aggregate (and once for a single item) and an introduction to the fine-grained UX attitudes is given, such as perceived recommendation quality, perceived system effectiveness, and perceived recommendation variety. Other attitudes such as understandability and perceived control are given in [21]. The definition of UX in this work seems to stem from Hassenzahl [13], who defined UX as a single "good-bad" feeling about an interface. Knijnenburg et al. instead proposes an evaluation framework which uses multiple attitudinal and perceptual measurements, but one limitation is that the work does not provide statistically-grounded justification for its modeling choices. Discriminant validity is lost between reliable (Cronbach's $\alpha > 0.8$ [6]) constructs when inter-construct correlations reach 0.7. For instance, in [22], high correlations (> 0.7) can be inferred between perceived system effectiveness and CS (study 1), between CS, perceived recommendation quality, perceived system effectiveness, and perceived effort (study 2), between perceived recommendation quality and perceived system effectiveness (study 3), and between perceived recommendation variety, perceived recommendation quality, and perceived system effectiveness (study 4). Moreover, although the UX attitudes appear to correlate with interaction behavior in this work, the attitudes are never used to measure longitudinal CS.

An alternative perspective on UX and CS is given in Pu et al. [28]. Different UX attitudes are proposed, e.g., transparency, confidence/trust, and adequacy to explain use intentions and purchase intention (which are also self-reported). Purchase intention reasonably represents self-reported CS ("I would buy the items recommended, given the opportunity"). An issue with this work is that, while discriminant validity appears slightly better, internal reliability of several proposed constructs would be considered statistically questionable or unacceptable by standards of Cronbach's α [6] (additionally, some constructs are indicated by fewer than 3 items, which makes the construct unidentifiable). This means that correlations of 0.5 or higher might indicate poor discriminant validity (due to attenuation), for instance, between interface adequacy and perceived ease of use, between trust/confidence and purchase intention, and between perceived control and overall satisfaction.

2.3 Other Concepts of UX

Other attempts have been made to quantify and validate UX. We have previously mentioned Hassenzahl's work [13][14], which views UX as a single scale (which would imply a single construct). Two highly cited books on UX [1][25] take a different approach by defining UX and usability to be essentially the same and propose measuring UX through metrics like task completion time. While this might indeed be useful, we agree more with Hassenzahl, Knijnenburg, and Pu that UX is an attitude of a user, not behavioral symptoms of that attitude, or the competence of the user. Next, a general purpose UX apparatus was proposed in [26]. Predictive validity is established in this work through a measure of task time, but this work has shortcomings that are similar to recommender systems UX (internal reliability is reported but discriminant validity is not assessed, high correlations between constructs might be inferred from their similar correlations with the task time metric).

Our background research lead us to conclude that the methodology for measuring UX and CS in recommender systems is still in question. Here, we re-open the issue of how to measure UX and CS in recommender systems research. In a previous study we discovered that users could not provide discriminant answers when asked about understandability, system satisfaction, and perceived persuasion [24] when exposed to recommendation interfaces. Now, we conduct a more expansive user study with a more realistic, open task design and evaluate a new method for measuring CS. We re-evaluate several UX attitudes (understandability, perceived effectiveness, perceived control, and trust/confidence) and directly assess their discriminant validity. The goal for this work is identify reliable **and** discriminant constructs for UX and CS.

3 SYSTEM DESIGN

This section describes the design of the interface in more detail. In designing the system for this study, we kept the following three goals in mind: a) to make the system as familiar to modern web users as possible, b) to make the system as similar to currently deployed recommender systems as possible, and c) to ensure that the study can be completed without forcing the users to accept recommendations from the system, so adherence can be measured. 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 1). The interface was closely modeled after modern movie "browsers" (such as IMDb or Movielens) that typically have recommender functionality. On the left side, the system featured basic search, sort, 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. The Movielens dataset has been widely studied in recommender systems research [27][18][12]. Due to update speed limitations of collaborative filtering, the dataset was randomly sampled for 4 million ratings, rather than the full 20 million.

3.1 Generating Recommendations

A traditional user-user collaborative filtering approach was chosen for the system. Details for this can be found in Resnick et al. [31]. Collaborative filtering was chosen due to the fact that it is well understood in the recommender systems community and it achieves extremely high performance on dense datasets such as MovieLens [23]. The results from this study should generalize reasonably well to other collaborative-filtering based techniques, such as matrix factorization and neighborhood models. We made two minor modifications to the default algorithm based on test results from our benchmark dataset: Herlocker damping and rating normalization[1].

3.2 User Interface Design

The interface provided the following functionality: 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 (CS was measured on 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 CS feedback stage).

3.2.1 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) **SORT**: 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.

3.2.2 Recommendation Side. The recommendation side operated identically to the browser side, except that the list was always sorted by the collaborative filtering prediction and the user could not override this behavior.

4 EXPERIMENT DESIGN

An observational study was conducted where participants interacted with the Movie Miner interface to find a set of movies to watch in the future. Participant behavior was not restricted and the entire setup was designed to match typical online sessions as closely as possible.

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 AMT is comparable to what would be collected from laboratory experiments [15]. Furthermore, since clickstream data can be collected, satisficing is easy to detect.

A list of all measurements taken in the study are given in Table 1. All of these items were taken on a Likert scale, except for when ratings were elicited, where a 5-star rating bar was used. Cronbach's alpha [6], a measurement of internal reliability, is reported. For this number, values above 0.8 are considered good fit, while values below 0.6 are considered unacceptable.

In this section, we first give a brief overview of the participant's experience with the study materials. Then, we discuss the measurements shown in Table 1 in detail. Finally, we detail the analysis strategy and hypotheses.

4.1 Procedure

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]. 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 find and 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

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

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

interface. A twelve minute session in which 5-7 items are selected was deemed sufficient time to select quality items, given that people only browse Netflix for 60-90 seconds to find a single item before giving up [11].

4.2 Measurement Model

We considered two new approaches to quantifying CS: 1) a simple approach that uses ease-of-satisfaction as a statistical control for CS (the EoS control approach), and 2) a more complicated method that builds a change-score model between ease-of-satisfaction and CS (the two-wave approach).

Three ease-of-satisfaction measurements were considered. The first is baseline satisfaction (BS), which was measured shortly after the pre-study by getting participant feedback on movies that were chosen from the database at random. Ten random movies were shown, one at a time, and the responses were averaged together. Next, participants were asked to self-report their own ease-of-satisfaction (SREoS) during the pre-study. Finally, we also considered the user's average item rating as a form of ease-of-satisfaction. It is important to note that the first metric is the only one that can be used for the two-wave approach while remaining statistically valid - this is because the measurement of CS and BS is identical.

CS and UX were measured after the participant had made their selection. For CS, the recommender interface was removed and the questions items were shown for each item chosen by the participant. Note that the question items are phrased in terms of the recommendations, not the interface. This is to help the participant distinguish between the browsing tools and the features of the recommender system. The UX attitudes were collected during the post-study using random-order questionnaires.

To assess predictive validity of CS and UX, we measured three behavioral variables: interaction with the browser tool, interaction with the recommendation tool, and adherence to recommendations. The interaction variables were measured as the total number of times that a user clicked within one of the two tools available, whether it was for inspecting a particular movie, rating a particular movie, or adding a movie to the watchlist. Adherence to recommendations was taken as the ratio of items in the final watchlist that originated from the recommendation side of the interface. This measurement was only possible since we didn't "force" participants to accept recommendations (the browser tool could be used to complete the entire study, if desired). It can be argued that adherence is one of the most important behavioral variables for recommendation research, since it represents behavioral evidence of a user's acceptance of recommendations, rather than a self-reported attitude. Adherence is typically not measured in recommender systems, which we believe is due to a lack of open methodologies, but it has been studied before in expert systems research [2].

4.3 Evaluation Strategy and Hypotheses

Here we describe the methodology for answering the research questions outlined in the introduction.

First, we evaluate the statistical validity of the following user attitudes: understandability, perceived system effectiveness, perceived control, and self-reported trust. Convergent validity of these constructs has already been demonstrated [22][28] and it is reproduced here. Therefore, we focus on discriminant and predictive validity. Discriminant validity is assessed using the Campbell and Fiske test [5], that is, checking that the correlation between constructs is < 0.85 while correcting for attenuation:

$$\frac{r_{xy}}{\sqrt{r_{xx} \cdot r_{yy}}} \tag{1}$$

where r_{xy} is the correlation between construct x and y and r_{xx} is the Cronbach's reliability of construct x. Predictive validity is tested by checking for the significance of regression coefficients when predicting interaction, adherence, and CS. Predictive power is tested through SEM. This leads us to the following hypotheses:

- H_1: Understandability is discriminant from perceived effectiveness, perceived control, and self-reported trust.
- H_2: Perceived effectiveness is discriminant from understandability, perceived control, and self-reported trust.
- H_3: Perceived control is discriminant from understandability, perceived effectiveness, perceived control, and self-reported trust.
- H_4: Self-reported trust is discriminant from understandability, perceived effectiveness, perceived control, and self-reported trust.
- H_5: UX predicts increased or decreased interaction behavior.
- H_6: UX predicts increased or decreased adherence.

Second, we test whether or not the BS and CS metrics are discriminant. Discriminant validity with BS provides additional evidence that the EoS control approach would be valid (it should not be expected that choosing random movies out of the database would satisfy anyone). More importantly, this test is the first step in assessing whether or not the SREoS factor is a valid replacement for BS. The BS construct and the CS construct should also be discriminant from UX for the EoS control approach to be valid. Discriminant validity is again assessed with Equation 1. Thus:

- H_7: CS is discriminant from UX.
- H_8: BS is discriminant from CS.

Third, we test whether a user's ease-of-satisfaction is a factor in UX and CS. That is, we want to know if ease-of-satisfaction is a personal factor that affects *all* attitudes in recommender systems research, even CS. If this were, true, we would expect to see significant regression estimates between the ease-of-satisfaction metric and all other attitudes, including CS. For these tests, we use BS, since it is likely the most objective metric of the user's ease-of-satisfaction. This leads to the following hypotheses:

- H_9: BS predicts increased understandability, p. effectiveness, p. control, and trust.
- H_{10}: BS predicts increased CS.

Next, we can use a Raykov change score model [30] to assess the validity of the two-wave approach. A Raykov change model is essentially a repeated-measures ANOVA for factor analysis. However, researchers may not want to require participants to answer multiple questions when assessing CS. Therefore, we can also test a simple two-wave growth curve model [36] using representative question items from baseline and CS. That is,

- H_{11}: UX is discriminant from changes in CS.

Factor	**Item Description**	R^2	**Est.**
SREoS	[r1] I think I will trust the movie recommendations given in this task.	0.81	1.17
ALPHA : 0.92	[r2] I think I will be satisfied with the movie recommendations given in this task.	0.83	1.18
	[r3] I think the movie recommendations in this task will be accurate.	0.75	1.15
UX/Understand.	[u1] How understandable were the recommendations?	0.538	1.174
ALPHA : 0.61	[u1] Movie Miner succeeded at justifying its recommendations.	0.756	1.482
	[u1] The recommendations seemed to be completely random.	0.427	-1.227
UX/Effectiveness	[e1] I preferred these recommendations over past recommendations.	0.643	1.406
ALPHA : 0.91	[e2] How accurate do you think the recommendations were?	0.781	1.494
	[e3] How satisfied were you with the recommendations?	0.852	1.602
	[e4] To what degree did the recommendations help you find movies for your watchlist?	0.653	1.387
UX/Control	[c1] How much control do you feel you had over which movies were recommended?	0.666	1.293
ALPHA : 0.86	[c2] To what degree do you think you positively improved recommendations?	0.638	1.238
	[c3] I could get Movie Miner to show the recommendations I wanted.	0.706	1.436
UX/Trust	[t1] I trust the recommendations.	0.861	1.573
ALPHA : 0.93	[t2] I feel like I could rely on Movie Miner's recommendations in the future.	0.845	1.640
	[t3] I would advise a friend to use the recommender.	0.723	1.575
UX, *ALPHA* : 0.93	All Understandability, Effectiveness, Control, and Trust items (similar R^2 and Est.)		
Choice Sat. (CS)	[cs1] How excited are you to watch <movie>?	0.78	0.66
ALPHA : 0.93	[cs2] How satisfied were you with your choice in <movie>?	0.89	0.70
	[cs3] How much do you think you will enjoy <movie>?	0.92	0.67
	[cs4] What rating do you think you will end up giving to <movie>?	0.57	0.34
Baseline Sat. (BS)	[bs1] How excited would you be to watch <movie>?	0.91	0.66
ALPHA : 0.97	[bs2] Would you be satisfied with choosing <movie>?	0.964	0.70
	[bs3] How much do you think you would enjoy <movie>?	0.955	0.67
	[bs4] What rating do you think you would end up giving to <movie>?	0.756	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. *ALPHA* indicates the Cronbach's alpha.

To test whether the two-wave approach would be advantageous over the EoS control approach, we can compare the correlations observed between UX and CS in each model.

Fourth, we assess whether or not the SREoS metric can stand in for BS, so it can be used as a stand-in for the EoS control method, which would help researchers save time. This is done again through Equation 1. Additionally, we can also check the model fit metrics, specifically, Bayesian information criterion (BIC) [19]. Next, we want to know whether or not BS can be implicitly teased out from a user's rating profile. This can be done just by testing the R^2 of the user's average profile rating when used as an indicator variable of BS. This leads to:

- H_{12}: SREoS and BS are discriminant.
- H_{13}: Replacing BS with SREoS results in $2 \ln B_{ij} > 10$.
- H_{14}: A user's average profile rating has an R^2 of > 0.5 when used as an indicator variable of BS.

Correlation (↔)	**Corr.**	**Att. Corr.**	**D?**
Understandability ↔ Effectiveness	0.977	1.311	N
Understandability ↔ Control	0.960	1.325	N
Understandability ↔ Trust	0.954	1.267	N
Effectiveness ↔ Control	0.953	1.077	N
Effectiveness ↔ Trust	0.985	1.070	N
Trust ↔ Control	0.953	1.066	N
UX ↔ CS	0.391	0.420	Y
UX ↔ Change in CS	0.060	0.065	Y
BS ↔ CS	0.243	0.256	Y
BS ↔ SREoS	0.223	0.236	Y

Table 2: Correlations and attenuated correlations between UX attitudes. The attenuated correlation cutoff for (D)iscriminant validity is 0.85, resulting in (Y)es/(N)o.

5 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 violating records were removed, resulting in the 526 complete records. Factor analysis was used to test the internal reliability and convergent validity of the factors shown in Table 1, which all resulted in good internal reliability, except for understandability ($\alpha = 0.61$), which was questionable. Grouping the UX attitudes into one factor also resulted in excellent reliability ($\alpha = 0.93$).

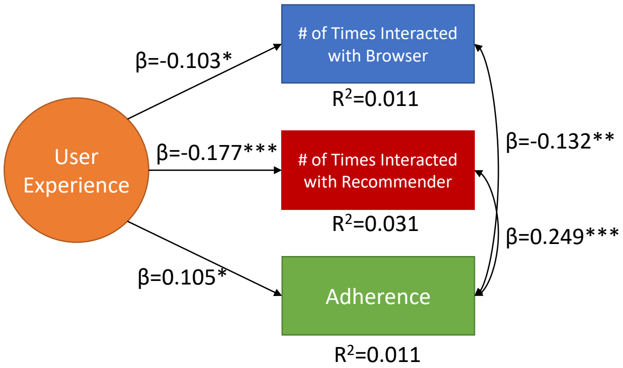


Figure 2: A visual of an SEM that predicts behavior based on the combined UX attitudes. Model fit metrics: $N = 526$ with 102 free parameters, $RMSEA = 0.054$ ($CI : [0.046, 0.062]$), $TLI = 0.972, CFI = 0.977$ over null baseline model, $\chi^2(102) = 257.515$.

5.1 Discriminant Validity

Table 2 indicates discriminant validity between constructs. Discriminant validity was tested by first correcting for attenuation (Equation 1) and then examining correlations between the factors. All UX attitudes were found to not be discriminant from each other, thus we reject H_{1-4}. Next, the UX attitudes were grouped into a single factor to avoid multi-collinearity when assessing predictive validity. An SEM was built (Figure 2)wherein the regression coefficients of browser interaction, recommender interaction, and adherence were determined. Minor, but significant, coefficients were found for each behavior, leading us to accept H_5 and H_6.

Next, we assess the discriminant validity of the BS and CS constructs. Table 2 shows that UX is discriminant from CS, with an attenuated correlation of 0.420. BS is also discriminant from CS, with an attenuated correlation of 0.256. Thus we accept H_7 and H_8.

5.2 Ease-of-Satisfaction: Baseline

Here we assess the predictive validity of the first ease-of-satisfaction metric: BS. An SEM was built (Figure 3) where BS was used as a predictor of the UX attitudes and CS. Regression coefficients of BS for the UX attitudes all achieved a significance of $p < 0.001$. The significance of the BS regression coefficient for CS was found to be $p < 0.01$. Additionally, an alternate SEM was tested where the UX attitudes were combined into a single construct. This again lead to significant coefficients for BS for UX and CS ($p < 0.001$). It should be noted that when controlling for BS, the correlation between CS and UX becomes 0.323, down from 0.391, increasing their discriminant validity. This leads us to accept H_9 and H_{10}.

Next, we assess the validity of using a growth curve model to measure change in CS (the two-wave method). This models the difference between CS and BS, rather than just the magnitude of CS. First, a Raykov change model is built (Figure 4), where BS is used as wave 1 and CS is used as wave 2. Variances of the change variables are not modeled, thus they are fixed (shown by a 0 in the figure). All pairs of items and the latent constructs are correlated. Then, to test UX as a predictor of change in CS, it is correlated with the wave 1 construct and the correlation coefficient is checked for the wave 2 construct. The coefficient was found to be non-signficant ($p = 0.240$, $\beta = 0.060$), so we accept H_{11}. This model indicates that the change in CS over the baseline is independent of the UX attitudes that were reported by participants.

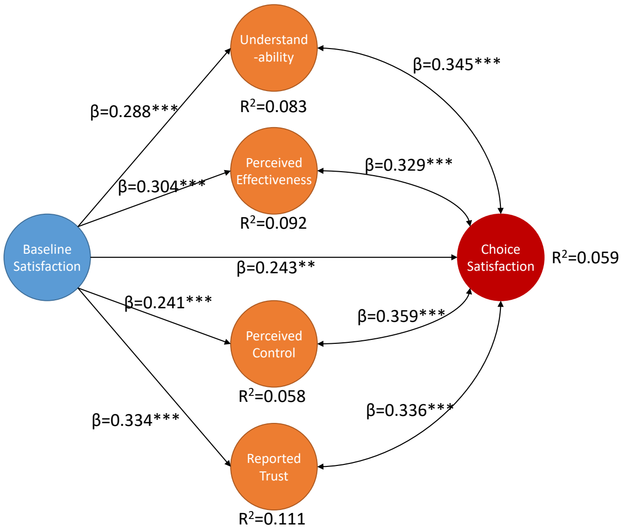


Figure 3: A visual of an SEM that shows the role of BS on UX attitudes. Model fit metrics: $N = 526$ with 180 free parameters, $RMSEA = 0.062$ ($CI : [0.056, 0.068]$), $TLI = 0.971, CFI = 0.965$ over null baseline model, $\chi^2(174) = 527.717$.

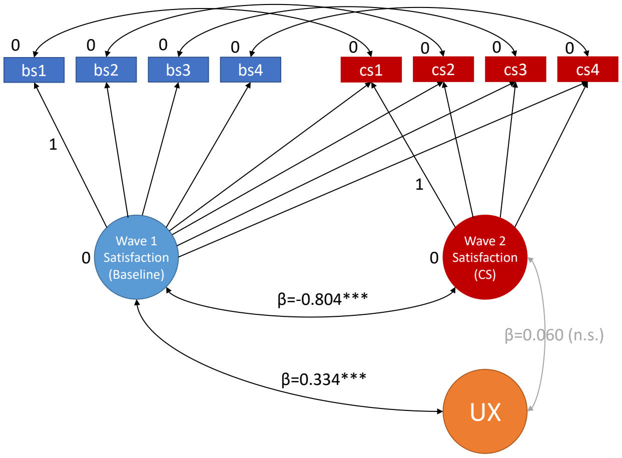


Figure 4: A Raykov change model using two waves: BS (first) and CS (second). UX does not predict a change in satisfaction. Items for satisfaction are shown in this figure to demonstrate how item loadings were configured. Overall fit metrics are typically not reported for growth curve models.

5.3 Ease-of-Satisfaction: Stand-ins

Finally, we assess stand-ins for the BS metric. First, self-reported ease-of-satisfaction and BS were found to be discriminant with an attenuated correlation of 0.236 (see Table 2). Thus H_{12} is rejected. Second, the BIC score (a measure of model fit, lower is better) in

Figure 3 was 26156.811. To test H_{13}, we swapped BS with self-reported ease-of-satisfaction (the rest of the model remained the same), which changed the BIC score to 27133.771. To check for significant differences, we evaluate $2 \ln B_{ij} = 20.41 - 20.33 = 0.08 < 10$ [19], so we reject H_{13} (indicating the two models are roughly equivalent). Since it was found that BS and SREoS are discriminant, we tested a model where both measures were used to predict UX attitudes and CS. Although the resulting model had lower BIC, improvements were observed in RMSEA, CFI, and TLI, as well as the R^2 for each UX attitude and CS. Third, the average profile rating of each user (which was collected in this study) was loaded onto BS. The R^2 for the user's average rating was 0.007, so we reject H_{14}.

6 DISCUSSION OF RESULTS

1. What is the statistical validity of the recommender systems approach to modeling attitudes and UX? In this experiment, there was no statistical evidence in favor of the multi-attitude approach described by Pu et al. [28] and Knijnenburg et al. [22]. Inter-construct correlations far exceeded the 0.85 cutoff for discriminant validity. Moreover, consider that the *less* correlation there is between each construct, the more efficient each construct becomes as a measurement (discriminant constructs are more likely to have better combined predictive power) – 0.85 is just a recommended cutoff. The results in this experiment indicate it is unlikely that *good* discriminant validity could be reached using a nuanced attitudinal approach. The reason for this might be that users simply do not, at least without attentional direction, form complex mental models of recommendation interfaces. Next, assessing the predictive validity of UX might be supported by the Theory of Reasoned Action [32], however, previous work has shown little evidence to suggest that modeling UX attitudes leads to the ability to predict adherence to recommendations or good decisions, only other user attitudes and interaction behavior. Here, the regression coefficient when predicting adherence was significant but minor ($\beta = 0.105$), which does not create a particularly strong case for predictive validity. Still, more interaction correlated with a slightly worsened UX, which in turn correlates with a slightly worsened CS. This result reinforces research results in consumer product selection, where it is known that more deliberation results in lower CS [7][37].

Although no evidence was found here that nuanced UX attitudes are discriminant, this does not close the book on them. There are some methodological differences between this work and [22]/[28] that could explain the differences. First, question phrasing was slightly different. Second, not all attitudes measured in the other studies were measured here. Finally, this study used a more open, realistic methodology where users undertook a typical movie selection task. While focusing on this task, users may not have paid much attention to scrutinizing the particular features of the system. This is in contrast with [22]/[28], which were more restrictive and had large differences between experiment treatments.

Our current recommendation is for recommender systems practitioners to allocate only a few questions towards assessing the UX attitudes that were trialed here. This is also supported by our previous study [24]. This suggestion implies a return to the single scale, "good-bad" feeling espoused by Hassenzahl [13]. Substantially more research is needed to determine if UX is a consistent predictor of behavior or if users are like to form detailed mental models of recommender systems.

2. Can CS be modeled in a way that makes it discriminant from UX? Our apparatus for CS, regardless of modeling approach, achieved discriminant validity from UX, with an attenuated correlation of 0.420. Although this is quite auspicious, this result may be dependent on the particular design that we chose for this study (users freely made their own selections). In algorithm experiments where users are asked to rate items which are pre-selected by the recommender, we cannot say whether or not the construct will again be discriminant. Despite this, it would be ideal if the correlation between CS and UX was very nearly zero, since a person's feeling about a particular movie should not be affected by the information tool that was used to discover that movie. Our analysis suggests that if BS is measured, researchers can create a two-wave model of CS that we have demonstrated to be independent of UX. While our method does require substantial feedback for each item (and collection of the baseline), we believe the construct could be represented by a single question "How much do you think you will enjoy <movie>?" due to an R^2 value of 0.92. In this case, a repeated measures ANOVA could be used instead of a growth curve model, which would simplify analysis.

3. How can a user's personal ease-of-satisfaction be measured and does it predict both UX and CS?. Although self-reported ease-of-satisfaction is not a valid replacement for measuring BS, it had the same effect on statistical conclusions in the model given in Figure 3. If the two-wave method (Figure 4) is not possible, self-reported ease-of-satisfaction could be used as a controlling variable so that the portion of CS variance explained by UX is minimized. Unfortunately, a user's average rating cannot stand in for either of these metrics, so ease-of-satisfaction cannot be inferred just by examining a user's rating profile.

One limitation of this work is that, outside of user experiments, recommendation practitioners may find it difficult to set up a scenario where the ease-of-satisfaction method is practical or useful. However, this research also clearly demonstrates that positive feelings felt by users during a user interface evaluation may easily confuse any practitioner's tests of good decision making. Having happy users does not necessarily mean having a useful and effective information system. Still, more research is needed on the validity of CS measurement, specifically, how it correlates with longitudinal satisfaction with selected items.

7 CONCLUSION

In summary, we investigated ways to account for a user's ease-of-satisfaction when measuring choice satisfaction in studies of recommendation. Our statistical analysis lead us to the conclusion that a two-wave approach to choice satisfaction would be ideal for identifying factors that lead users to good decisions, since changes in choice satisfaction appear to be independent from user experience. We have also identified a shortcut to controlling for ease-of-satisfaction by asking users to perform a self-assessment. The methodology proposed here has the potential to help future researchers identify factors in recommender systems that lead users to satisfying choices.

REFERENCES

[1] William Albert and Thomas Tullis. 2013. *Measuring the user experience: collecting, analyzing, and presenting usability metrics.* Newnes.

[2] Vicky Arnold, Nicole Clark, Philip A Collier, Stewart A Leech, and Steve G Sutton. 2006. The differential use and effect of knowledge-based system explanations in novice and expert judgment decisions. *Mis Quarterly* (2006), 79–97.

[3] Dirk Bollen, Bart P Knijnenburg, Martijn C Willemsen, and Mark Graus. 2010. Understanding choice overload in recommender systems. In *Proceedings of the fourth ACM conference on Recommender systems.* ACM, 63–70.

[4] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2011. Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on psychological science* 6, 1 (2011), 3–5.

[5] Donald T Campbell and Donald W Fiske. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin* 56, 2 (1959), 81.

[6] Lee J Cronbach, Peter Schönemann, and Douglas McKie. 1965. Alpha coefficients for stratified-parallel tests. *Educational and Psychological Measurement* 25, 2 (1965), 291–312.

[7] Ap Dijksterhuis and Zeger Van Olden. 2006. On the benefits of thinking unconsciously: Unconscious thought can increase post-choice satisfaction. *Journal of Experimental Social Psychology* 42, 5 (2006), 627–631.

[8] David Dunning. 2011. The Dunning–Kruger effect: On being ignorant of one's own ignorance. In *Advances in experimental social psychology.* Vol. 44. Elsevier, 247–296.

[9] Claes Fornell and David F Larcker. 1981. Evaluating structural equation models with unobservable variables and measurement error. *Journal of marketing research* (1981), 39–50.

[10] Daniel Gilbert. 2009. *Stumbling on happiness.* Vintage Canada.

[11] Carlos A Gomez-Uribe and Neil Hunt. 2016. The Netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)* 6, 4 (2016), 13.

[12] F Maxwell Harper and Joseph A Konstan. 2016. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 5, 4 (2016), 19.

[13] Marc Hassenzahl. 2008. User experience (UX): towards an experiential perspective on product quality. In *Proceedings of the 20th Conference on l'Interaction Homme-Machine.* ACM, 11–15.

[14] Marc Hassenzahl and Noam Tractinsky. 2006. User experience-a research agenda. *Behaviour & information technology* 25, 2 (2006), 91–97.

[15] David J Hauser and Norbert Schwarz. 2015. Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior research methods* (2015), 1–8.

[16] Jörg Henseler, Christian M Ringle, and Marko Sarstedt. 2015. A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the academy of marketing science* 43, 1 (2015), 115–135.

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

[18] Jason J Jung. 2012. Attribute selection-based recommendation framework for short-head user group: An empirical study by MovieLens and IMDB. *Expert Systems with Applications* 39, 4 (2012), 4049–4054.

[19] Robert E Kass and Adrian E Raftery. 1995. Bayes factors. *Journal of the american statistical association* 90, 430 (1995), 773–795.

[20] Sunghan Kim, M Karl Healey, David Goldstein, Lynn Hasher, and Ursula J Wiprzycka. 2008. Age differences in choice satisfaction: A positivity effect in decision making. *Psychology and aging* 23, 1 (2008), 33.

[21] Bart P Knijnenburg, Svetlin Bostandjiev, John O'Donovan, and Alfred Kobsa. 2012. Inspectability and control in social recommenders. In *Proceedings of the sixth ACM conference on Recommender systems.* ACM, 43–50.

[22] Bart P Knijnenburg, Martijn C Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction* 22, 4-5 (2012), 441–504.

[23] Yehuda Koren and Robert Bell. 2015. Advances in collaborative filtering. In *Recommender systems handbook.* Springer, 77–118.

[24] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2017. User Preferences for Hybrid Explanations. In *Proceedings of the Eleventh ACM Conference on Recommender Systems.* ACM, 84–88.

[25] Mike Kuniavsky. 2003. *Observing the user experience: a practitioner's guide to user research.* Elsevier.

[26] Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and evaluation of a user experience questionnaire. In *Symposium of the Austrian HCI and Usability Engineering Group.* Springer, 63–76.

[27] Bradley N Miller, Istvan Albert, Shyong K Lam, Joseph A Konstan, and John Riedl. 2003. MovieLens unplugged: experiences with an occasionally connected recommender system. In *Proceedings of the 8th international conference on Intelligent user interfaces.* ACM, 263–266.

[28] Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems.* ACM, 157–164.

[29] Pearl Pu, Li Chen, and Rong Hu. 2012. Evaluating recommender systems from the userâĂŹs perspective: survey of the state of the art. *User Modeling and User-Adapted Interaction* 22, 4-5 (2012), 317–355.

[30] Tenko Raykov. 1992. Structural models for studying correlates and predictors of change. *Australian Journal of Psychology* 44, 2 (1992), 101–112.

[31] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. 1994. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In *Proceedings of ACM CSCW'94 Conference on Computer-Supported Cooperative Work.* 175–186. http://www.acm.org/pubs/articles/proceedings/cscw/192844/p175-resnick/p175-resnick.pdf

[32] Michael J Ryan and Edward H Bonfield. 1975. The Fishbein extended model and consumer behavior. *Journal of Consumer Research* 2, 2 (1975), 118–136.

[33] Nava Tintarev and Judith Masthoff. 2007. A survey of explanations in recommender systems. In *Data Engineering Workshop, 2007 IEEE 23rd International Conference on.* IEEE, 801–810.

[34] Marko Tkalcic, Andrej Kosir, and Jurij Tasic. 2011. Affective recommender systems: the role of emotions in recommender systems. In *Proc. The RecSys 2011 Workshop on Human Decision Making in Recommender Systems.* Citeseer, 9–13.

[35] Jodie B Ullman and Peter M Bentler. 2003. *Structural equation modeling.* Wiley Online Library.

[36] Dietrich Von Rosen. 1991. The growth curve model: a review. *Communications in Statistics-Theory and Methods* 20, 9 (1991), 2791–2822.

[37] Timothy D Wilson, Douglas J Lisle, Jonathan W Schooler, Sara D Hodges, Kristen J Klaaren, and Suzanne J LaFleur. 1993. Introspecting about reasons can reduce post-choice satisfaction. *Personality and Social Psychology Bulletin* 19, 3 (1993), 331–339.

[38] Timothy D Wilson, Jay Meyers, and Daniel T Gilbert. 2003. âĂIJHow happy was I, anyway?âĂİ A retrospective impact bias. *Social Cognition* 21, 6 (2003), 421–446.