A User-Centric Evaluation Framework for Recommender Systems

Pearl Pu
HCI Group
Swiss Federal Institute of Technology (EPFL)
CH-1015, Lausanne, Switzerland
pearl.pu@epfl.ch

Li Chen
Department of Computer Science
Hong Kong Baptist University
224 Waterloo Road, Hong Kong
lichen@comp.hkbu.edu.hk

Rong Hu
HCI Group
Swiss Federal Institute of Technology (EPFL)
CH-1015, Lausanne, Switzerland
rong.hu@epfl.ch

ABSTRACT

This research was motivated by our interest in understanding the criteria for measuring the success of a recommender system from users' point view. Even though existing work has suggested a wide range of criteria, the consistency and validity of the combined criteria have not been tested. In this paper, we describe a unifying evaluation framework, called ***ResQue*** (***Re****commender* ***s****ystems'* ***Q****uality of* ***u****ser* ***e****xperience*), which aimed at measuring the qualities of the recommended items, the system's usability, usefulness, interface and interaction qualities, users' satisfaction with the systems, and the influence of these qualities on users' behavioral intentions, including their intention to purchase the products recommended to them and return to the system. We also show the results of applying psychometric methods to validate the combined criteria using data collected from a large user survey. The outcomes of the validation are able to 1) support the consistency, validity and reliability of the selected criteria; and 2) explain the quality of user experience and the key determinants motivating users to adopt the recommender technology. The final model consists of thirty two questions and fifteen constructs, defining the essential qualities of an effective and satisfying recommender system, as well as providing practitioners and scholars with a cost-effective way to evaluate the success of a recommender system and identify important areas in which to invest development resources.

ACM Classification Keywords

H1.2 [**User/Machine Systems**]: Human factors; H5.2 [**User Interfaces**]: evaluation/methodology, user-centered design.

General Terms

Measurement, Design, Human Factors.

Keywords

Recommender systems, quality of user experience, e-Commerce recommender, post-study questionnaire.

1. INTRODUCTION

A recommender system is a software technology that proactively suggests items of interest to users based on their objective behavior or their explicitly stated preferences. It provides benefits to users and enhances websites' revenue. For users, the benefits include higher efficiency in finding preferential items, more confidence in making a purchase decision, and the potential to discover something new. For the marketer, this technology can significantly enhance users' likelihood to buy the items recommended to them and their overall satisfaction and loyalty, increasing users' likelihood to return to the site and recommend the site to their friends.

Thus far, previous research work on the evaluation of recommender systems has mainly focused on algorithm accuracy [1], especially objective prediction accuracy [9]. More recently, researchers began examining issues related to users' subjective opinions [10, 22] and developing additional criteria to evaluate recommender systems [25]. In particular, they suggest that user satisfaction does not always correlate with high recommender accuracy [14]. Increasingly, researchers are investigating user experience issues in an effort to understand and identify effective preference elicitation methods [13], techniques that motivate users to rate items that they have experienced [2], methods that generate diverse and more satisfying recommendation lists [25], explanation interfaces [23], trust formation with recommenders [4], design guidelines for enhancing a recommender's interface layout [16], and other determinants that influence users' positive perception of recommender systems [22]. As these criteria are emerging, the need is arising for a consolidated definition of what constitutes an effective and satisfying recommender system from the user's perspective.

Our present work aims at developing a unifying framework, called ***ResQue***, based on existing usability-oriented research in this field, as well as principles from well-known usability evaluation models, such as TAM [6] and SUMI [11], and validating the framework using psychometric techniques [15]. As recommender technology is becoming widely accepted, the ability to characterize user experience and users' affective attitudes toward the technology has become extremely important. The conceptualization of ***ResQue***, and its development process, aims at advancing our understanding of the critical user issues of recommender technology and users' motivation in accepting the technology.

The remainder of the paper is organized as follows. Section 2 details the related work on evaluation frameworks from the user point of view in the realm of recommender systems in. Section 3 describes the development of the ***ResQue*** model by defining the constructs and the hypothesized relations. Section 4 and 5 present the refinement and validation process of the model, including detailed information on the experimental setup, data analysis and discussions, as well as a evaluation questionnaire using only

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
RecSys'11, October 23–27, 2011, Chicago, Illinois, USA.
Copyright 2011 ACM 978-1-4503-0683-6/11/10...$10.00.

fifteen questions. Finally, Section 6 outlines our conclusions and future plans.

2. RELATED WORK: EVALUATION FROM THE USER POINT OF VIEW

Xiao and Benbasat [26] employed a surveying method of previous empirical studies of Recommendation Agents (RAs) and developed a conceptual model of twenty-eight validated hypotheses, relating consumers' use of RA in online shopping environments to their perception such as ease of use, control, trust and perceived system effectiveness. Their findings revealed that RA use improved consumers' decision quality. Furthermore, the use of explicit preference elicitation method led to better decision quality, though at the cost of higher decision effort. Similarly, ease of generating new or additional recommendations improved the ease of use of an RA and increased user control resulted in increased users' trust and satisfaction. On the other hand, recommending too many alternatives could lead to poor product choices and reduced consumers' selectivity. Finally, explanations of an RA's inner logic strengthened users' trusting beliefs in the RA's perceived competence and benevolence.

Knijnenburg et al. [12] provided a framework to explain how objective system aspects (such as its input, process and output) influence users' perception of these aspects, and how this perception eventually influences users' choice satisfaction and their intent to provide feedback. Six structurally related constructs were proposed in their framework: objective recommender system aspects, subjective evaluations, subjective experiences, objective behaviors, situational, and personal characteristics. The characteristics of a number of constructs are similar to ***ResQue***, such as perceived recommendation accuracy, process aspect of the system (interaction adequacy), system effectiveness and choice satisfaction (perceived usefulness of the system), diversity, and subjective evaluation (beliefs and attitudes). Additional coverage includes considerations for users' expertise of the domain, concerns for privacy, and situational and personal characteristics. However, the outcomes were limited to the specific recommender systems (multimedia recommenders) and the algorithms used in the experiments. The constructs, developed in six separate user studies, have not been combined and validated in their entirety.

The models given in [26] and [12] support crucial hypotheses relating users' perception of recommender systems to their choice satisfaction and willingness to provide preference feedback. However, they failed to relate user perception to the likelihood of user adoption of the systems. Moreover, the constructs have not been validated using psychometric methods. Hence, it is unlikely we can use the models as measuring instruments to evaluate the ultimate success of recommender systems.

Ozok et al. [16] explored the usability and user preferences of recommender interfaces in a large-scale user study involving college students. An evaluation framework and a set of guidelines were proposed to evaluate the interface in terms of its structure (e.g., layout and look) and the content (e.g., the availability of the image, price, and quantity). Using this evaluation procedure can point out detailed problems with a recommender's interface. We thus recommend using this model in addition to ***ResQue***.

3. MODEL DEVELOPMENT

An evaluation questionnaire consists of a set of constructs, the participating questions for each construct, and the hypotheses relating the constructs. The starting point is to delimit the domain of the constructs and generate sample questions representing the concepts under consideration for each construct. We carefully surveyed prior work to identify key user experience variables as the main construct, paying special attention to meaningful and carefully conducted user studies (detailed inventory of these user studies will be given in the subsections). The subsequent work involves a significant amount of trial-and-error effort in phrasing the questions until their semantic meanings are clear to users.

During the development of the model, we also compared our constructs with those used in TAM and SUMI, which are two well-known and widely adopted measurement frameworks.

TAM (Technology Acceptance Model) seeks to understand a set of perceived qualities of a system and users' intention to adopt the system as a result of these qualities, thus explaining not only the desirable outcome of a system, but also users' motivation [6]. The original TAM listed three constructs: perceived ease of use of a system, its perceived usefulness and users' intention to use the system. However, TAM was also criticized for its over-simplicity and generality. Venkatesh et al. [24] formulated an updated version of TAM, called the Unified Theory of Acceptance and Use of Technology. In this version, four key constructs (performance expectancy, effort expectancy, social influence, and facilitating conditions) were presented as direct determinants of usage intentions and behaviors.

SUMI (Software Usability Measurement Inventory) is a psychometric evaluation model developed by Kirakowski and Corbett [11] to measure the quality of software from the end-user's point of view. The model consists of 5 constructs (efficiency, affect, helpfulness, control, and learnability) and 50 questions. It is widely used to help designers and developers assess the quality of use of a software product or prototype, and can assist with the detection of usability flaws and the comparison between software products.

Instead of generating a linear model of sample questions, we structured the question items into four layers of higher-level constructs crossing four dimensions: the *perceived system qualities*, *users' beliefs* as a result of these qualities, their *subjective attitudes*, and their *behavioral intentions*. Such a topology can more clearly explicate how users' perception of the physical features of a system influences their beliefs, attitudes, and finally behaviors.

We proposed an initial database of questions for ***ResQue*** in [20]. As the first step of the refinement process, we approached expert researchers in the community to review the model and suggest changes to the items used (known as the Delphi method). Together with a pilot study, this method helped us modify and remove items that were judged to be repetitive, confusing, or both. Finally, we obtained the current model comprising fifteen constructs and forty-three questions. In the paragraphs below, we describe the meaning of the constructs and what they are supposed to measure, as well as a review of existing works that have inspired us to derive these constructs.

3.1 Perceived System Qualities

As the first evaluation layer, *perceived system qualities* refer to users' perception of the objective characteristics (e.g., functional and informational capabilities) of a recommender system. Since they are exogenous variables of the model, we needed evidence from prior work that these variables indeed influence users' beliefs, attitudes and behaviors under controlled experiments over randomized samples. We focus on three essential dimensions: the quality of recommendations, the interaction adequacy and the interface adequacy of the recommender.

3.1.1 Recommendation Quality

The primary task of recommender systems is to suggest items of interest to users. The quality of the suggested items is considered in the literature to be one of the most critical issues determining the success of a system. In our earlier work, we have found strong correlations between the following qualities of the recommended items to users' intention to use the system.

Perceived accuracy is the degree to which users feel the recommendations match their interests and preferences. It is an overall assessment of how well the recommender has understood the users' preferences and tastes. This subjective measure is significantly easier to obtain than the measure of objective accuracy that we used in our earlier work [18]. Our studies show that the two traits are strongly correlated [4]. In other words, if users respond well to this question, it is likely that the underlying algorithm is accurate in predicting users' interest.

Novelty (or discovery) is the extent to which users receive new and interesting recommendations. The core concept of novelty is related to the recommender's ability to educate users and help them discover new items [19]. In [14], a similar concept, called "serendipity", was suggested. Herlocker [9] argued that novelty is different from serendipity, because novelty only covers the concept of "new" while serendipity encompasses not only "new" but also "surprising". However, in conducting the actual user evaluation procedure, the meticulous distinction between these two words will cause confusion for users. Therefore, we suggest novelty and discovery as two similar questions.

Attractiveness refers to whether or not the recommended items are capable of stimulating users' imagination and evoking a positive emotion of interest or desire. Attractiveness is different from accuracy and novelty. An item can be accurate and novel, but not necessarily attractive; a novel item is different from anything a user has ever experienced, whereas an attractive item stimulates the user in a positive manner. This concept is similar to the salience factor in [14].

Diversity measures the diversity level of items in the recommendation list. When recommendations are shown in a list, an item-based accuracy evaluation metric is not appropriate for evaluation tasks. Users would likely become bored and weary if they always get items from the same producers. At this stage, it has been found that a low diversity level might disappoint users and could cause them to leave this recommender [10]. Some research suggests that a recommendation list as a complete entity should be judged for its diversity rather than treating each recommendation as an isolated item [25].

Context compatibility evaluates whether or not the recommendations consider general or personal context requirements. For example, for a movie recommender, the necessary context information may include a user's current mood, different occasions for watching the movie, whether or not other people will be present, and whether the recommendation is timely. A good recommender system should be able to formulate recommendations considering different kinds of contextual factors that will likely take effect.

3.1.2 Interface Adequacy

Interface design issues related to recommenders have also been extensively investigated in [8, 14, 16, 23]. Most of the existing work is concerned with how to optimize the recommender page layout to achieve the maximum visibility of the recommendation, i.e. whether to use image, text, or a combination of the two. A detailed set of design guidelines were investigated and proposed [16]. In our current model, we mainly emphasize users' subjective evaluations of a recommender interface in terms of its **information sufficiency**, the **interface label** and **layout adequacy and clarity**.

3.1.3 Interaction Adequacy

Besides issues related to recommendation quality and interface adequacy, the system's ability to elicit user preferences, allow for user feedback and to explain the reasons why recommendations facilitate purchasing decisions also weighs heavily on users' overall perception of a recommender. Thus, three main interaction mechanisms are usually suggested in various recommenders: **initial preference elicitation**, **preference revision**, and the system's ability to explain its results (i.e., **explanation**).

Behavioral based recommenders do not require users to explicitly indicate their preferences, but collect such information via users' browsing and purchasing history. For rating and preference based recommenders, this process requires a user to rate a set of items or state their preferences on desired items in a graphical user interface [18]. Some conversational recommenders provide explicit mechanisms for users to provide feedback in the form of critiques [4]. The simplest critiques indicate whether the recommended item is good or bad, while the more sophisticated ones show users a set of alternative items that take into account users' desire for these items and the potential superior values they offer, such as better price or more popularity.

3.1.4 Information sufficiency and explicability

Information quality of software is traditionally an important area to assess for technology adoption. In the case of recommender systems, information sufficiency denotes the system's ability to display price, quantity, the image, user reviews, or any other information of an item to help users with making a decision.

More specific to this field is the system's ability to explain why items are suggested to the active user. Herlocker et al. [8], Sinha and Swearingen [22] and Tintarev and Masthoff [23] demonstrated that a good explanation interface could help inspire users' trust and satisfaction by giving them information to personally justify recommendations, increasing user involvement and educating users on the internal logic of the system [8, 23]. In addition, Tintarev and Masthoff [23] defined in detail possible aims of explanation facilities: transparency, scrutability, trust, effectiveness, persuasiveness, efficiency, and satisfaction. Pu and Chen extensively investigated design guidelines for developing explanation-based recommender interfaces [3]. They found that organization interfaces are particularly effective in promoting users' satisfaction of the system, convincing users to buy the recommended items, and bringing them back to the store.

3.2 Beliefs

The second evaluation layer, *beliefs*, refers to a higher level of user perception of a system, which is influenced by perceived qualities. At this layer, users' concerns focus on how effectively and efficiently the system helps them accomplish tasks, such as decision support, as well as the nature of the interactions between the system and the users.

3.2.1 *Perceived Usefulness*

Perceived usefulness of a recommender (called perceived competence in our previous work) is the extent to which a user finds that using a recommender system would improve his/her performance, compared with their experiences without the help of a recommender [3]. This element requests users' opinion as to whether or not this system is useful to them. Since recommenders used in e-commerce environments mainly assist users in finding relevant information to support their purchase decision, we further qualify the usefulness in two aspects: decision support and decision quality.

The objective of decision technologies in general is to overcome the limit of users' bounded rationality and to help them make more satisfying decisions with a minimal amount of effort [36]. Recommender systems specifically help users manage an overwhelming flood of information and make high-quality decisions under limited time and knowledge constraints. **Decision support** thus measures the extent to which users feel assisted by the recommender.

In addition to the efficiency of decision making, the quality of the decision (**decision quality)** also matters. The quality of a system-facilitated decision can be assessed by confidence criterion, which is the level of a user's certainty in believing that he/she has made a correct choice with the assistance of a recommender.

3.2.2 *Perceived Ease of Use*

Perceived ease of use, also known as efficiency in SUMI and perceived cognitive effort in our existing work [4, 10], measures users' ability to quickly and correctly accomplish tasks with ease and without frustration. We also use it to refer to decision efficiency, i.e. the extent to which a recommender system facilitates users to find their preferential items quickly. Although task completion and learning time can be measured objectively, it can be difficult to distinguish the actual task completion time from the measured task time for various reasons. Users can be exploring the website and discovering information unrelated to the assigned task. This is especially true if a system is entertaining and educational, and its interface and content is very appealing. It is also possible that the user perceives that he/she has consumed less time while the measured task completion time is in fact high. Therefore, evaluating perceived ease of use may be more appropriate than using the objective task completion time to measure a system's ease of use.

3.2.3 *Control and Transparency*

User control measures if users felt in control in their interaction with the recommender. The concept of user control includes the system's ability to allow users to revise their preferences, to customize received recommendations, and to request a new set of recommendations. This aspect weighs heavily in the overall user experience of the system. If the system does not provide a mechanism for a user to reject recommendations, a user will be unable to stop the system from continuously recommending items which might cause him/her to be disappointed with the system.

Transparency determines whether or not a system allows users to understand its inner logic, i.e. why a particular item is recommended to them. A recommender system can convey its inner logic to the user via an explanation interface [3, 8, 22, 23]. To date, many researchers have emphasized that transparency has a certain impact on other critical aspects of users' perception. Swearingen and Sinha [22] showed that the more transparent a recommended product is, the more likely users would be to purchase it. In addition, Simonson [21] suggested that perceived accuracy of a recommendation is dependent on whether or not the user sees a correspondence between the preferences expressed in the measurement process and the recommendation presented by the system.

3.3 Attitudes

Attitude is a user's overall feeling towards a recommender, which is most likely to be derived from her/his experience as s/he interacts with a recommender. An attitude is generally believed to be more long-lasting than a belief. Users' attitudes towards a recommender are highly influential to their subsequent behavioral intentions. Many researchers attribute positive attitudes, including users' satisfaction and trust of a recommender, as important factors.

Evaluating **overall satisfaction** determines what users think and feel while using a recommender system. It gives users an opportunity to express their preferences and opinions about a system in a direct way. **Confidence inspiring** refers to the recommender's ability to convince users of the information or products recommended to them. **Trust** indicates whether or not users find the whole system trustworthy. Studies show that consumer trust is positively associated with their intentions to transact, purchase a product, and return to the website [7]. The trust level is determined by the reputation of online systems, as well as the recommender system's ability to formulate good recommendations and provide useful explanation interfaces [3, 8, 13].

3.4 Behavioral Intentions

Behavioral intentions towards a system are related to whether or not the system is able to influence users' decision to use the system and purchase some of the recommended results.

One of the fundamental goals for an e-commerce website is to maximize user loyalty and the lifetime value to stimulate users' future visits and purchases. User loyalty evaluates the system's ability to convince users to reuse the system, or persuade them to introduce the system to their friends in order to increase the number of clients. Accordingly, this dimension consists of the following criteria: user agreement to use the system, user acceptance of the recommended items (resulting in a purchase), user retention and intention to introduce this system to her/his friends. Theory of Planned Behavior [24] states that behavioral intention can be a strong predictor of actual behavior. Although the website's integrity, reputation and price quality will also likely impact user loyalty, the most important factor for a recommender system is to help users effectively find a satisfying product, i.e. the quality of its recommendations [6].

3.5 Hypotheses

To validate our model, we hereby construct a set of hypotheses about how the various constructs relate to each other (Figure 1). We postulate that recommendation quality, interface adequacy, and interaction adequacy would have positive effects on users' beliefs in the recommenders, including users' perceived ease of use, perceived usefulness, and control/transparency. We also

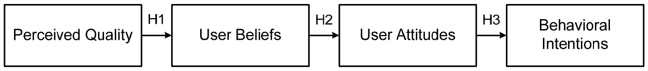

Figure 1. General evaluation framework with hypothesized influence paths.

Table 2. Test results of internal reliability and convergent validity. Constructs with single items are included for completeness.

Constructs	Items	Internal reliability		Convergent validity		
		Cronbach alpha (0.5)	Item-total correlation (0.4)	Factor loading (0.5)	Composite reliability (0.8)	Variance extracted (0.5)
1. Recommendation Accuracy	1					
The items recommended to me matched my interests.						
2. Recommendation Novelty	2	0.636			0.846	0.733
The items recommended to me are novel.			0.467	0.856		
The recommender system helped me discover new products.			0.467	0.856		
3. Recommendation Diversity	1					
The items recommended to me are diverse.						
4. Interface Adequacy	4	0.701			0.848	0.583
The labels of the recommender interface are clear.			0.501	0.719		
The labels of the recommender interface are adequate.			0.628	0.822		
The layout of the recommender interface is attractive.			0.515	0.723		
The layout of the recommender interface is adequate.			0.584	0.784		
5. Explanation	1					
The recommender explains why the products are recommended to me.						
6. Information Sufficiency	1					
The information provided for the recommended items is sufficient for me to make a purchase/download decision.						
7. Interaction Adequacy	3	0.818			0.915	0.783
The recommender allows me to tell what I like/dislike.			0.697	0.875		
I found it easy to tell the system what I like/dislike.			0.752	0.903		
I found it easy to inform the system if I dislike/like the recommended item.			0.573	0.789		
8. Perceived Ease of Use	2	0.535			0.811	0.682
I became familiar with the recommender system very quickly.			0.400	0.826		
I easily found the recommended items.			0.400	0.826		
9. Control	3	0.855			0.912	0.775
I feel in control of modifying my taste profile.			0.645	0.829		
The recommender allows me to modify my taste profile.			0.762	0.900		
I found it easy to modify my taste profile in the recommender.			0.785	0.911		
10. Transparency	1					
I understood why the items were recommended to me.						
11. Perceived Usefulness	3	0.677			0.823	0.609
The recommender helped me find the ideal item.			0.582	0.843		
Using the recommender to find what I like is easy.			0.512	0.793		
The recommender gave me good suggestions.			0.402	0.699		
12. Overall Satisfaction						
Overall, I am satisfied with the recommender.						
13. Confidence & Trust	4	0.763			0.890	0.669
I am convinced of the items recommended to me.			0.657	0.838		
I am confident I will like the items recommended to me.			0.580	0.791		
The recommender made me more confident about my selection/decision.			0.506	0.713		
The recommender can be trusted.			0.514	0.720		
14. Use Intentions	3	0.754			0.865	0.682
I will use this recommender again.			0.551	0.796		
I will use this recommender frequently.			0.703	0.893		
I will tell my friends about this recommender.			0.556	0.784		
15. Purchase Intention	1					
I would buy the items recommended, given the opportunity.						

Table 1. Profile of participants (the total number is 239).

	Item	Frequency	Percentage (%)
Gender	Male	37	15.48
	Female	202	84.52
Age	Below 20	30	12.55
	21-30	177	74.06
	31-40	28	11.72
	Above 40	4	1.67
Profession	Student	150	62.76
	Researcher/Engineer	26	10.88
	Manager	4	1.67
	Others	59	24.69
Nationality	Asia	87	36.40
	Europe	152	63.60
Evaluated recommender websites	Amazon.com	91	38.08
	Youtube.com	43	17.99
	Douban.com	32	13.39
	Imdb.com	11	4.60
	Others	62	25.94

hypothesize the existence of significant causal effects from users' beliefs to their attitudes. Finally, increased attitudes would eventually lead to users' behavioral intentions, such as users' intention to introduce the system to their friends, intention to return, and intention to purchase the recommended product. In the following sections, we present the model validation.

4. MODEL VALIDATION

Psychometric questionnaires such as the one proposed in this paper require validation of the questions used, data gathering, and statistical analysis before they can be used with confidence. We directed our effort at proofing the reliability and validity of the questions using factor analysis and testing the model's fitness using structural equation methods.

4.1 Experiment Setup

We launched a large-scale survey from July to August 2010 among 239 participants in Europe and Asia.[1] This sample size of users is sufficient for a stable factor estimate, according to the suggestion of using at least 5 participants per item by Nummally [15]. The 239 participants were recruited via mailing lists. More than 70% of them are in the 21-30 age group, with the rest of them distributed in the other three age groups. Their nationalities are very diverse; 87 of them are from Asian countries (e.g., China, Korea and Vietnam), and 152 from European countries (e.g., France, Switzerland, Germany, etc.). All subjects have had previous experience with recommender systems in online environments. The sites our users most frequently evaluated were Amazon (91 subjects), Youtube (43), Douban (32)[2] , and IMDB (11). Detailed descriptions of all participants are shown in Table 1.

Table 3. Inter-construct correlation matrix.

	1	2	3	4	5	6	7	8
1. Rec. Novelty	**0.856**							
2. Interface Adequacy	0.092	**0.764**						
3. Interaction Adequacy	-0.030	0.256	**0.885**					
4. Perceived Ease of Use	0.045	0.618	0.194	**0.826**				
5. Control	-0.016	0.160	0.631	-0.008	**0.880**			
6. Perceived Usefulness	0.463	0.489	0.319	0.548	0.284	**0.780**		
7. Confidence & Trust	0.345	0.487	0.195	0.401	0.261	0.741	**0.818**	
8. Use Intentions	0.276	0.429	0.264	0.526	0.043	0.465	0.315	**0.826**

The main user task was to "*find an ideal product to buy or experience from an online site employing recommender technology*" of their own choice, and respond to the ***ResQue*** survey. To refresh a subject's memory of her/his experience with the chosen system, we asked the subjects to write down the name(s) of at least one recommendation that s/he received from the online site just before filling out the evaluation questionnaire. For each question, a 5-point Likert scale from "strongly disagree" (1) to "strongly agree" (5) was used to characterize users' responses.

A total of 10 gifts were given to the winners of an incentive draw: two iPod Touch (259 CHF), two iPod Nano (199 CHF), two iHome iP9 (129 CHF), two iPod Shuffle (99.99 CHF), and two gifts valued at 79.90 CHFs.

4.2 Validity and Reliability of the Model

Users' responses were carefully checked for its quality. For example, we looked for users who would give the same scales for all questions on the same page, on all of the pages, or conflicting answers to similar and reverse questions. Fortunately, we found only one outlier and removed it from further analysis. We then validated the rationality of each construct in our proposed model and their relations by applying factor analysis, structural equation modeling (SEM) and other techniques as described in [15].

We first computed the internal consistency and reliability of the model using Cronbach's alpha and item-to-total correlations. This validation process is intended to reveal internal consistencies of a given construct as well as identifying the clusters of related variables. Since some of the alpha values were under 0.5 (a value viewed as acceptable), we followed some of the common practices to improve reliability levels (such as using Churchill's recommendations [5]). The items with low correlated item-total correlations (< 0.40) were discarded or re-grouped into another construct. After several iterations, we obtained values as indicated in Table 2. They meet the cut-off points of at least 0.5 for Cronbach's alpha and 0.4 for item-total correlation [17].

We examined the convergent validity of the measurement items by factor loading and composite reliability. The results are also shown in Table 2. Factor loadings of all items in each construct exceeded the acceptable level of 0.5 [17]. Composite reliabilities ranged from 0.811 (for perceived ease of use) to 0.915 (for interaction adequacy), and all exceeded the recommended level of 0.8. Therefore, the results demonstrated a convergent validity of the measurement items.

We also assessed discriminant validity via inter-construct correlations (see results in Table 3). Correlations between any two constructs were all less than the square root value of average variances that are shown in the diagonal, which represents a level of appropriate discriminant validity.

These results thus validated that the measures of the constructs as examined in our study are robust in terms of their internal consistency reliability, and that the convergent and discriminant validity of our instruments were also proven satisfactory. The next step was then to perform the structural equation modeling analysis to verify hypotheses.

[1] The survey is available at: http://hci.epfl.ch/resque/survey/index.html.

[2] A popular recommender site in Chinese, recommending movie, music, and books.

4.3 Structural Model

We tested the overall fit of our path model, which evaluates our hypotheses (see Figure 1) on the causal relationships among these evaluation constructs. Figure 2 shows the results of the structural model analysis, including the R2 (coefficients of determination) and path loadings. As all of the R2 estimates are larger than 0.10, they are appropriate and informative to examine the significance of the paths associated with these variables. The model goodness-of-fit indices are χ2 = 404.537 (d.f. = 366), p = 0.081, GFI = 0.904, CFI = 0.984, RMSEA = 0.021, which surpassed the recommended values of these model fit indices.

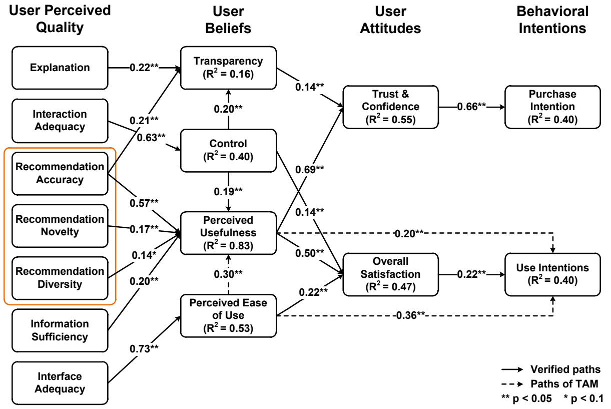

Figure 2. Structural model fit. Path significance: **p < 0.05, *p < 0.1.

We first examined the path relationships between constructs included in user perceived quality and ones in user beliefs. It shows that perceived usefulness is statistically significantly influenced by perceived recommendation accuracy (β = 0.57, p < 0.001), recommendation novelty (β = 0.17, p < 0.05), and moderately related to recommendation diversity (β = 0.14, p < 0.01). Recommendation accuracy significantly leads to transparency (β = 0.21, p < 0.01). Interface adequacy has significantly leading effect on perceived ease of use (β = 0.73, p < 0.001). Interaction adequacy also shows significant influence on control (β = 0.63, p < 0.001). Moreover, information sufficiency has a significant influence on perceived usefulness (β = 0.20, p < 0.01) and explanation has a significant effect on transparency (β = 0.22, p < 0.01).

Furthermore, perceived ease of use is significantly affecting perceived usefulness (β = 0.30, p < 0.001). Perceived ease of use also shows significant impacts on overall satisfaction (β = 0.22, p < 0.05). Perceived usefulness significantly influences users' attitudes, including overall satisfaction (β = 0.50, p < 0.001) and trust and confidence (β = 0.69, p < 0.001). Control significantly affects transparency (β = 0.20, p < 0.01), perceived usefulness (β = 0.19, p < 0.001) and overall satisfaction (β = 0.14, p < 0.05). Transparency significantly influences users' trust and confidence (β = 0.14, p < 0.05).

As for the relationship between user beliefs/attitudes and behavioral intentions, we found that perceived usefulness shows significant influence on use intention (β = 0.20, p < 0.05). Use intention is also significantly affected by satisfaction (β = 0.22, p < 0.05) and perceived ease of use (β = 0.36, p < 0.01). On the other hand, purchase intention is strongly impacted by trust and confidence (β = 0.66, p < 0.001).

5. DISCUSSIONS

The validation process of ***ResQue*** produced the final model with fifteen constructs and thirty-two questions (Table 2), and proved the hypotheses relating the constructs (Figure 2). In a nutshell, the model provides an elaborate explanation on how the perceived qualities of the recommended items (accuracy, novelty and diversity), interface, interaction, and information qualities (labels, layout, and ability to explain and collect feedback) influence users' belief of the transparency, ease of use, usefulness, sense of control of the system, and users' overall satisfaction and trust with the systems, and how these qualities influence and motivate users' behavioral intentions, including their intention to return to the system and purchase the recommended products.

The core of the ***ResQue*** model kept the original constructs of TAM (the dotted lines of Figure 2), which confirms that a useful technology must also be easy to use and easy to understand for the wide adoption to happen. Furthermore, sense of control mediates perceived usefulness much like ease of use. That is, the more users feel in control of making the system recommend the right items to them, the more they would find the system useful. To understand users' motivation for buying items that were recommended to them via e-commerce websites, our research postulated and validated the trust construct, its antecedents and its influence on users' purchasing behavior.

ResQue also revealed more intricate details than the original TAM. Given that the majority of the evaluated recommenders fall into e-commerce and entertainment websites, the behavioral construct was split into *usage* and *purchase* intentions. This distinction helps explain that while overall satisfaction is highly correlated to the simple use of the system, trust/decision confidence is crucial to persuade users to purchase the recommended items. Moreover, even though the recommendation diversity also influences the perceived usefulness, its influence is not as strong as the recommendation novelty and accuracy, at least according to the subjects of our study.

Several constructs and questions were eliminated due to low correlations with other variables used in the model, such as the attractiveness and context compatibility of recommended items. To put them back in the model, more controlled experiments should be performed to show evidence that they indeed influence users' attitudes and behavioral adoptions.

5.1 Short Version of *ResQue*

An important goal of this research was to come up with a fast but reliable way to evaluate a recommender system. If a questionnaire with thirty-two questions is too long, we provide the following method to reduce the scope.

Since questions of a given construct are highly correlated, asking one question is enough for the assessment of that construct. The short version, therefore, consists of one question from each of the fifteen constructs. Additional questions can be added for semantic robustness without exceeding user effort limit. The actual choice depends on the context of the recommender system and the evaluator's objectives.

6. CONCLUSION

This paper presents an overview of recent user experience research in recommender technology. The examination of combined criteria for usability and satisfaction led to the conceptualization of the first balanced measurement framework, ***ResQue***, to assess users' attitudes and acceptance towards a recommender. Most importantly, a Web-based survey confirms that ***ResQue*** provides validity and reliability of its structures, and that the proven paths carry meaning causal relationships among the constructs.

The paper details how we have combined and validated existing criteria into a well-balanced user-centric evaluation framework for recommender systems. This model and its fifteen criteria define the essential qualities of an effective and satisfying recommender system and the key determinants motivating users to adopt the recommender technology. Our work was able to extend beyond prior work, which localized on few assessment criteria, and provided a meaningful explanation of the overall characteristics of user experience. The two types of users' behavioral intentions, i.e., purchase intention and use intentions, have been delineated to determine and explain the recommender's role in e-commerce and entertainment websites. Finally, we provided both long and short versions of a questionnaire to help designers and researchers perform a usability and user acceptance test during any stage of the system implementation. The questionnaires can be applied to assess different types of recommenders, including rating-based, utility-based, and knowledge-based systems, regardless of the backend engines used.

Our future work includes further analysis and collection of user data to understand cultural influences (European vs. Asian users) as well as the influence of the data domains (entertainment vs. e-commerce) on users' attitudes and behaviors.

7. REFERENCES

[1] Adomavicius, G. and Tuzhilin, A. 2005. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Trans. Knowl*. Data Eng. 17(6), 734-749.

[2] Beenen, G., Ling, K., Wang, X., Chang, K., Frankowski, D., Resnick, P., et al. 2004. Using social psychology to motivate contributions to online communities. *In Proceedings of CSCW '04*. New York: ACM Press.

[3] Chen, L. and Pu, P. 2006. Trust Building with Explanation Interfaces. *In Proceedings of International Conference on Intelligent User Interface,* 93-100.

[4] Chen, L. and Pu, P. 2009. Interaction Design Guidelines on Critiquing-based Recommender Systems. *User Modeling and User-Adapted Interaction Journal (UMUAI)*, Springer Netherlands, Volume 19, Issue3, 167-206.

[5] Churchill, G.A. A paradigm for developing better measures of marketing constructs, *Journal of Marketing Research 16 (1)*, 1979, pp. 64-73.

[6] Davis, F.D. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quart.* 13 319-339.

[7] Grabner-Kräuter, S. and Kaluscha, E.A. 2003. Empirical research in on-line trust: a review and critical assessment *Int. J. Hum.-Comput. Stud. (IJMMS) 58(6),* 783-812.

[8] Herlocker, J.L., Konstan, J.A., and Riedl, J. 2000. Explaining collaborative filtering recommendations. *CSCW'00*, 241-250.

[9] Herlocker, J.L., Konstan, J.A., Terveen, L.G., and Riedl, J. 2004. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst*. 22(1), 5-53.

[10] Jones, N., and Pu, P. 2007. User Technology Adoption Issues in Recommender Systems. *In Proceedings of Networking and Electronic Commerce Research Conference*, 379-394.

[11] Kirakowski, J. 1993. SUMI: the Software Usability Measurement Inventory. *British Journal of Educational Technology***,** 24 (3) 210-214.

[12] Knijnenburg, B. P., Willemsen, M. C., Gantner, Z. and Soncu, H. 2011. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction Journal (UMUAI)*, Special Issue on User Interfaces for Recommender Systems. (upcoming)

[13] McNee, S.M., Lam, S.K., Konstan, J.A., Riedal, J. 2003. Interfaces for eliciting new user preferences in recommender systems. *User Modeling 2003*, 178-187.

[14] McNee, S.M., Riedl, J., and Konstan, J.A. 2006. Being accurate is not enough: How accuracy metrics have hurt recommender systems. *CHI Extended Abstracts*, 1097-1101.

[15] Nunnally, J. C. 1978. Psychometric Theory.

[16] Ozok, A.A, Fan, Q., Norcio, A.F. 2010. Design guidelines for effective recommender system interfaces based on a usability criteria conceptual model: results from a college student population. *Behaviour & Information Technology*, Volume 29, Issue 1, 57 - 83.

[17] Peterson, R.A. 1994. A meta-analysis of Cronbach's coefficient alpha. *Journal of Consumer Research*, Vol. 21 September, pp. 381-91.

[18] Pu, P., Chen, L., and Kumar, P. 2008. Evaluating Product Search and Recommender Systems for E-Commerce Environments. *Electronic Commerce Research Journal*, 8(1-2), 1-27.

[19] Pu, P., Zhou, M., and Castagnos, S. 2009. Critiquing Recommenders for Public Taste Products. *In proceedings of RecSys'09*. New York, ACM Press, 249-252.

[20] Pu, P. and Chen, L. 2010. A User-Centric Evaluation Framework of Recommender Systems. *In the 3rd ACM Conference on Recommender Systems*, Workshop on User-Centric Evaluation of Recommender Systems and Their Interfaces, Barcelona, Spain, Sept. 26-30.

[21] Simonson, I. 2005. Determinants of customers' responses to customized offers: Conceptual framework and research propositions. *Journal of Marketing*, 69, 32–45.

[22] Swearingen, K. and Sinha, R. 2002. Interaction design for recommender systems. *In Interactive Systems (DIS2002).*

[23] Tintarev, N. and Masthoff, J. 2007. Survey of explanations in recommender systems. *ICDE Workshops 2007*, 801-810.

[24] Venkatesh,V., Morris, M.G., Davis, G.B. and Davis, F.D. 2003. User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 2003, 27, 3, 425-478.

[25] Ziegler, C.N., McNee, S.M., Konstan, J.A., and Lausen, G. 2005. Improving Recommendation Lists through Topic Diversification. In *Proc. of WWW 2005*, ACM Press (2005), 22-32.

[26] Xiao, B. and Benbasat, I. 2007. Ecommerce Product Recommendation Agents: Use, Characteristics, and Impact. *Mis Quarterly* 31(1), 137-209.