To appear in the journal of User Modeling and User-Adapted Interaction (UMUAI), vol. 22

Explaining the user experience of recommender systems

Bart P. Knijnenburg
bart@usabart.nl[1]
Human-Technology Interaction group, School of Innovation Sciences
Eindhoven University of Technology, The Netherlands
Currently at the Department of Informatics, Donald Bren School of Information and Computer Sciences, University of California, Irvine, USA

Martijn C. Willemsen
m.c.willemsen@tue.nl
Human-Technology Interaction group, School of Innovation Sciences
Eindhoven University of Technology, The Netherlands

Zeno Gantner
gantner@ismll.de
Information Systems and Machine Learning Lab (ISMLL)
University of Hildesheim, Germany

Hakan Soncu
hakan.soncu@microsoft.com
European Microsoft Innovation Center, Germany

Chris Newell
chris.newell@rd.bbc.co.uk
BBC Research & Development, United Kingdom

Abstract

Research on recommender systems typically focuses on the accuracy of prediction algorithms. Because accuracy only partially constitutes the user experience of a recommender system, this paper proposes a framework that takes a user-centric

[1] H.S. and C.N. conducted the Microsoft and BBC field trials, Z.G. implemented recommender algorithms used in the experiments, B.P.K. designed experiments, analyzed data and wrote paper, M.C.W. was co-author and advisor.

approach to recommender system evaluation. The framework links objective system aspects to objective user behavior through a series of perceptual and evaluative constructs (called subjective system aspects and experience, respectively). Furthermore, it incorporates the influence of personal and situational characteristics on the user experience. This paper reviews how current literature maps to the framework and identifies several gaps in existing work. Consequently, the framework is validated with four field trials and two controlled experiments, analyzed using Structural Equation Modeling. The results of these studies show that subjective system aspects and experience variables are invaluable in *explaining why and how the user experience of recommender systems comes about*. In all studies we observe that perceptions of recommendation quality and/or variety are important mediators in predicting the effects of objective system aspects on the three components of user experience: process (e.g. perceived effort, difficulty), system (e.g. perceived system effectiveness) and outcome (e.g. choice satisfaction). Furthermore, we find that these subjective aspects have strong and sometimes interesting behavioral correlates (e.g. reduced browsing indicates higher system effectiveness). They also show several tradeoffs between system aspects and personal and situational characteristics (e.g. the amount of preference feedback users provide is a tradeoff between perceived system usefulness and privacy concerns). These results, as well as the validated framework itself, provide a platform for future research on the user-centric evaluation of recommender systems.

Keywords

Recommender systems, decision support systems, user experience, user-centric evaluation, decision-making, human-computer interaction, user testing, preference elicitation, privacy

1 Introduction

Recommender systems are designed to help the user make better choices from large content catalogs, containing items as distinct as books, movies, laptops, cameras, jokes, and insurance policies (Xiao & Benbasat, 2007; Resnick & Varian, 1997). Before the advent of recommender systems, such content-based systems would offer users the entire catalog (possibly with a generic search/filter feature). Recommender systems, on the other hand, offer each user a personalized subset of items, tailored to the user's preferences. The system derives these user preferences

from implicit or explicit feedback (Pommeranz et al., 2012). Implicit feedback recommenders analyze clicking/purchasing behavior (e.g. amazon.com, see also Hauser et al., 2009). Explicit feedback recommenders let users rate items, (e.g. youtube.com, see also McNee et al., 2002; Cena et al., 2010; Gena et al., 2011), critique items (see also Chen & Pu, 2012; Viappiani et al., 2006), assign weights to item attributes (see also Häubl et al., 2004), or indicate their specific needs (e.g. HP.com 'help me choose', see also (Felix et al., 2001)). Finally, the system calculates recommendations by comparing the user's preferences to the features of the catalog items (content-based recommender systems), or to other users' preferences (collaborative filtering recommenders).

A typical interaction proceeds as follows: First, the user's preferences are elicited. Based on the collected preference data, the system tries to predict how much the user would appreciate each of the available items in the catalog. Finally, the system presents the user those items that have the highest predicted value to the user. In some recommender systems this terminates the interaction, in other systems the users continue to indicate their preferences and receive recommendations continually.

An essential aspect of any recommender system is the algorithm that provides personalized recommendations based on the user's preferences (Burke, 2002). The more accurate the predictions of this algorithm, the more accurately the system can predict the best recommendations for the user. Not surprisingly, a significant part of the research on recommender systems concerns creating and evaluating better prediction algorithms (McNee et al., 2006; Cosley et al., 2003; Ziegler et al., 2005). An excellent overview of available algorithms can be found in (Burke, 2002) and in (Adomavicius & Tuzhilin., 2005); more recent approaches were presented in (Koren et al., 2009; Koren, 2010; Hu et al., 2008). Herlocker et al. (2004) provide a thorough discussion of available evaluation metrics.

The premise of this algorithm research is that better algorithms lead to perceivably better recommendations, which in turn lead to better user experience in terms of choice satisfaction and perceived system effectiveness. However, several researchers have argued that there are other factors that influence the *user experience* (users' subjective evaluation of their interaction with the system), and that these factors have not received the amount of attention they deserve (McNee et al., 2006; Cosley et al., 2003; Murray & Häubl, 2008; Murray & Häubl, 2009; Ozok et al., 2010; Pu et al., 2012; Konstan & Riedl, 2012). System aspects other than accuracy can influence satisfaction and other evaluative measures (e.g. diversification; Ziegler et al., 2005; Willemsen et al., 2011). Furthermore, situational or personal aspects (e.g. product expertise; Kamis & Davern, 2004; Knijnenburg & Willemsen, 2009; Knijnenburg & Willemsen, 2010; Knijnenburg et al., 2011; and privacy concerns; Teltzrow & Kobsa, 2004; Komiak & Benbasat, 2006) can also

influence how people interact with and evaluate the system. Unfortunately, even studies that consider aspects other than accuracy look at a limited set of variables that influence each other (e.g., how satisfaction changes due to a diversification, or how choices become more accurate with the inclusion of a recommender engine) without integrating these variables into a model of overall user experience.
An integrated view on the user experience of recommender systems can be obtained by means of user-centric development (McNee et al., 2006) and evaluation (Pu & Chen, 2010; Pu et al., 2012). The current paper therefore extends and tests our user-centric evaluation framework for recommender systems proposed in Knijnenburg et al. (2010). To understand and improve the user experience of recommender systems, it is necessary to conduct empirical evaluations that consider the entire process of how the user experience comes about. Therefore, our framework describes how objective aspects of the system (e.g. the algorithms used) are subjectively perceived by the user (e.g. if they perceive differences in recommendation quality for these different algorithms), and how these perceptions, together with personal and situational characteristics, result in specific user experience and interaction with the system (e.g. whether a higher perceived recommendation quality leads to a more positive evaluation of the system, a higher satisfaction with the chosen items, and a change in user behavior). Such a framework will provide a deeper understanding of how objective system aspects influence the user experience and behavior through perceived system aspects. It thereby allows for a better understanding of *why and how* certain aspects of the system result in a better user experience and others do not, which helps further user-centric research and development of recommender systems.

2 Components of the framework

The main goal of our framework is to provide a set of structurally related concepts that can be used in empirical studies to describe and measure the user experience of recommender systems. User experience is an ill-defined concept, and lacks well-developed assessment methods and metrics (McNamara & Kirakowski, 2006; Law et al., 2009). In our framework, we distinguish between objective system aspects (e.g. algorithms, user interface features), subjective system aspects (users' perceptions of these objective system aspects) and user experience (users' evaluations of their interaction with the system) and interaction (users' behaviors). We also consider the context of the interaction in terms of personal and situational characteristics. Before we describe the framework itself, we will discuss several theories that served as a basis for our framework.

2.1 Existing theories

2.1.1 Normative and attitudinal models

At the core of many psychological models of human behavior is the Theory of Reasoned Action (TRA) by Fishbein and Ajzen (1975). This theory claims that attitudinal and normative factors influence behavioral intention, which in turn predicts actual behavior. Davis, Bagozzi and Warshaw (1989; see also Davis, 1989) adopted the attitudinal part of this theory in their Technology Acceptance Model (TAM). In the TAM, the attitude towards using a technology is explained by the perceived usefulness and perceived ease of use of the system. Venkatesh, Morris, Davis and Davis (2003) created a similar theory called the Unified Theory of Acceptance and Use of Technology (UTAUT), based on the normative part of TRA, showing how personal and situational characteristics can influence behavioral intention. In the UTAUT, attitudinal concepts are entirely replaced by more experience-related evaluative concepts (performance expectancy, effort expectancy, social influence, and facilitating conditions).

With respect to our framework, these theories make a distinction between behaviors (and behavioral intentions) and the attitudes that cause these behaviors. These attitudes are in turn caused by experiential factors like perceived usefulness and ease of use (TAM), and by personal and situational characteristics (UTAUT).

2.1.2 User experience models

Hassenzahl (2008) defines user experience (UX) as "a momentary, primarily evaluative feeling (good-bad) while interacting with a product or service. Good UX is the consequence of fulfilling the human needs for autonomy, competence, stimulation (self-oriented) through interacting with the product or service (i.e. hedonic quality)." Hassenzahl's (2005) model of user experience describes how certain objective aspects of the system (e.g. its interaction and presentation style) are perceived in terms of pragmatic attributes (i.e. does the system deliver high quality results in an effortless way?) and hedonic attributes (i.e. does it stimulate, is it desirable?). These perceptions in turn cause an experiential evaluation in terms of appeal, pleasure and satisfaction.

With respect to our framework, Hassenzahl's model links objective system aspects to evaluative experiential factors though subjective perceptions. The distinction between perception and evaluation is subtle but important: Perception denotes whether certain objective system aspects register with the user at all, while evaluation denotes whether the perceived aspect has any personal relevance to the user. This may for instance give us an insight in why users may perceive a change in recommendation quality but at the same time do not show a change in experience or behavior.

Furthermore, whereas the TRA-related theories are restricted to pragmatic attributes, Hassenzahl also stresses the importance of hedonic attributes. Experimental evidence shows that hedonic attributes like pleasure and 'flow' (a feeling of automatic and highly focused interaction; Csikszentmihalyi, 1975) indeed also determine the user experience (Koufaris, 2003; Hsu & Lu, 2004; Yu et al., 2005).

2.1.3 User experience models for recommender systems

Hayes et al. (2002) propose a framework for testing the user satisfaction of recommender algorithms in operational systems. Their approach is restricted to behavioral measures of satisfaction, and their focus is primarily on the algorithm. Furthermore, Hayes et al.'s work has limits because they advocate a setup in which several algorithms are tested at the same time for the same user, an approach which provides a significant departure from the normal user experience of a recommender system which generally employs only one algorithm at a time.

Zins and Bauernfield (2005) constructed a model of the user experience of recommender systems based on a survey conducted among users of two travel recommenders and a system for finding digital cameras. Their model shows how personal characteristics influence trust, flow, and browsing behavior, and how these in turn influence system satisfaction. A clear limitation of their model is that it does not explain how objective system aspects may influence the user experience.

McNee et al. (2006) created an analytic model of Human-Recommender Interaction (HRI) for the development of recommender systems. The goals and tasks of the users are analyzed and used to determine the appropriate recommender system dialogue and 'personality'. McNee et al. 's model clearly serves a different purpose than our framework (development as opposed to evaluation). They do however suggest linking subjective HRI metrics to traditional objective performance metrics of algorithm accuracy, and stress the importance of the context (e.g. users' goals and tasks) in which recommendations are made.

Xiao and Benbasat (2007) presented an extensive literature review of the marketing-oriented research on recommender systems. Their overview, too, provides insight into the mechanisms underlying the user experience of recommender systems, albeit from many different studies (each focusing just on one part of the entire experience). Their resulting framework shows how certain characteristics of recommender systems cause changes in users' evaluation and decision-making behaviors, and their adoption of the recommender system. It also includes personal and situational characteristics that moderate these effects. The framework we present below bears a lot of similarity to Xiao and Benbasat's framework, but goes beyond it by including subjective system aspects. Moreover, while their framework is constructed mainly for the purpose of summarizing

existing research, we pose our framework as a starting-point for the evaluation of recommender systems.

Pu and Chen (2010, see also Pu et al., 2012) provide an extensive questionnaire to test several specific experience concepts of recommender systems. Their research model also explicitly considers perceived system qualities as antecedents of user beliefs, attitudes and behavioral intentions, and in that way it is similar to our framework. Our framework, however, takes a more abstract approach by providing a description of the structural relationships between the general, higher level concepts that play a role in user experience, without strictly specifying operational, lower level constructs and the questions that measure them[2]. To answer specific research questions, researchers need to define and operationalize a set of specific, lower level constructs, and Pu and Chen's questionnaires can be used as a starting point for this operationalization; another option is to use the pragmatic procedure for recommender system evaluation that is based on our framework (Knijnenburg et al., 2011a). However, since user experience is highly contingent upon the purpose of the system under evaluation, the specific concepts and specific questionnaire items to measure these concepts may differ from study to study. Moreover, Pu and Chen's model does not include context (personal and situational characteristics), and it does not suggest how objective system aspects influence the various constructs in their framework, making it more difficult to select concepts from their framework for a particular user study.

Ozok et al. (2010) provide a wide range of design guidelines based on a questionnaire of recommender system usability. Their results describe the effects of specific system aspects on the usability of recommender systems. However, they employ a descriptive approach, which relies on the users' stated opinions about recommender systems in general instead of experimental manipulations of a specific system.

2.2 The main structure of the framework explained

Figure 1 shows our framework. Like Hassenzahl (2005) and Xiao and Benbasat (2008), we take *objective system aspects* (OSA) as a starting point for the evaluation. The objective system aspects consist of the algorithms used by the system, the visual and interaction design of the system, the way it presents the recommendations, and additional features such as social networking. Like Hassenzahl (2005), we link these objective system aspects to *subjective system aspects* (SSA), which represent users'

[2] In this respect, our framework resembles Fishbein and Ajzen's (1975) TRA, which also instructs researchers to elicit a different set of personal and normative beliefs to measure per experiment.

perception of the objective system aspects. Also like in Hassenzahl's (2005) model, these subjective system aspects include both pragmatic characteristics (usability and quality) and hedonic characteristics (appeal). The subjective system aspects, measured with questionnaires, are expected to mediate the influence of the objective system aspects on the user experience. The main reason for including subjective system aspects as mediators is that recommender systems provide a personalized experience to the user, and that this personalization may not be equally apparent to all users. The subjective system aspects can show whether the objective aspects are perceived at all.

Like all existing models, our framework carefully distinguishes between attitude and behavior[3], although we use the more specific terms experience and interaction. The *experience* (EXP) signifies users' evaluation of the system. In that way it is closely related to attitudes in TAM (or rather the factors that cause them), with the addition of hedonic aspects (like in Hassenzahl's (2005) model). Experience is also measured with questionnaires, and is conceptually divided into the evaluation of the system (system-EXP), the evaluation of the decision process (process-EXP), and the evaluation of the final decisions made (outcome-EXP). The *interaction* (INT) is the observable behavior of the user. A complex interplay exists between interaction and experience: a positive user experience changes the interaction, but the interaction is also what initially caused the user experience.

Like the models of Xiao and Benbasat (2007) and Venkatesh et al. (2003), our model asserts that experience and interaction typically also depend on *personal* and *situational characteristics* (referred to as PC and SC). Personal characteristics include demographics, trust, domain knowledge, and perceived control (the latter two are prominent in TRA). Situational characteristics are dependent on the context of the interaction; at different points in time, users may have different choice goals, trust and privacy concerns, and familiarity with the system (McNee et al., 2006).

[3] To emphasize this distinction we have slightly altered the labels in the framework since our previous publications.

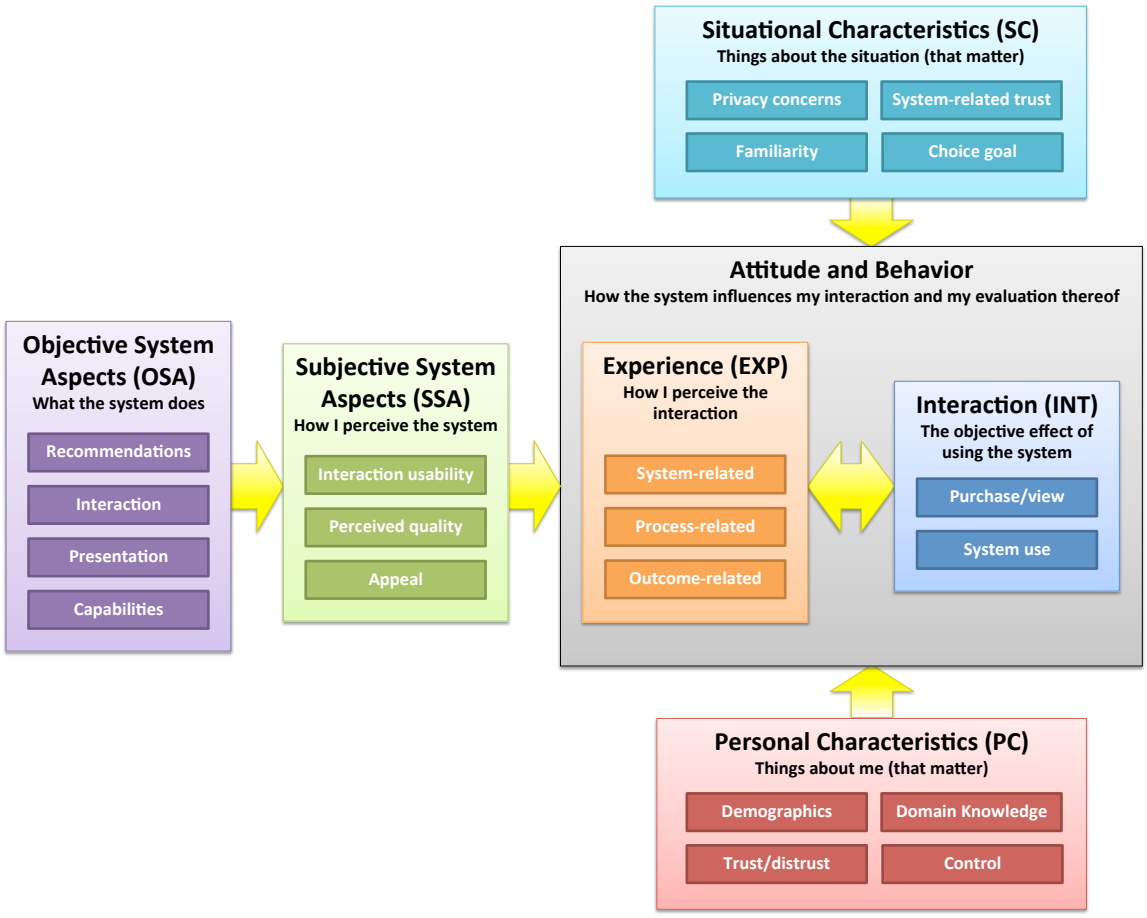

Figure 1: Our framework for the user-centric evaluation of recommender systems.

3 Expected benefits of our framework

Our framework explicitly links the objective interaction (INT) to objective system aspects (OSA) through a series of subjective constructs (SSA and EXP). The framework can be used as a guideline for controlled user experience. Specifically, by manipulating a certain system aspect (OSA) in a controlled experiment (keeping all other aspects the same), one can identify the effect of this aspect on the users' perceptions (SSA), experience (EXP) and behaviors (INT). By careful manipulation of specific objective system aspects, researchers can uncover generic truths about recommender systems. These can then inform the design and development of future versions of the studied aspects.

Moreover, the framework allows one to conduct empirical evaluations in a more integrative fashion than most existing recommender systems research: It allows researchers to consider the interplay between multiple objective system aspects

(e.g. algorithm versus interface), as well as to investigate the trade-offs between several aspects of user experience (e.g. satisfaction versus choice difficulty).
The framework provides insight into the relationships between the general concepts that play a role in the user experience of recommender systems. By tailoring the operationalization of these general concepts to the specific system under evaluation, the framework can be applied to a range of different types of consumer-facing recommender systems, including e-commerce recommenders (recommending products), media recommenders (recommending, for example videos, music or news articles) and social network recommenders (recommending users to befriend or follow). To exemplify this point, in the remainder of this section we use our framework to map a wide array of existing research. Most of this existing research (as well as our own empirical work) specifically considers what in the realm of e-commerce recommenders has been called "experience products", for which the quality is hard to determine before purchase, as opposed to "search products", which have attributes that can be considered before the decision is made (Nelson, 1970). Because it is hard to determine the quality of experience items before the actual decision is made, decision-making processes for such items typically rely more heavily on recommendations and other external sources of information (Bhatnagar & Ghose, 2004; Senecal & Nantel, 2004; Huang et al, 2009; Ochi et al, 2009).
To conclude this section, we indicate specific gaps in current knowledge, and provide an overview of the research opportunities that these gaps present. In the subsequent section, we present the results of several *empirical evaluations* that use parts of the framework for their main hypotheses. After presenting the results of these evaluations one by one, the findings will be integrated under the generic concepts of the framework. This will allow us to validate the framework, and to take a first step towards bridging the uncovered gaps. To effectively integrate our findings, we limited our empirical evaluations to media recommenders.

3.1 The objects of user experience evaluation

User experience as defined in our framework (EXP) is not a one-dimensional concept; it may entail various aspects (broadly ranging from pragmatic to hedonic concepts) and several different *objects of evaluation*. Especially for recommender systems, knowing the object of evaluation is critical in understanding the dynamics of the user experience (Paramythis et al., 2010; Tintarev & Masthoff, 2012): When we say that the user experience of recommender system X is better than that of system Y, are we evaluating the system, the process of using the system to get to a decision, or the chosen item itself? This distinction is important, as different system

aspects may influence different objects of user experience; a visually attractive interface may improve the evaluation of the system (system-EXP), a good preference elicitation method may make decisions easier (process-EXP), and an accurate algorithm may increase the quality of the final decision (outcome-EXP).
Furthermore, the evaluations of the different experience objects may influence each other. For instance, a positive evaluation of the chosen item(s) may "rub off" on the evaluation of the system. To capture the multi-faceted nature of user-experience, our framework therefore considers each of these objects: the system, the process, and the outcome.
In current research, however, this is rarely done; researchers in different domains use different objects of evaluation. Particularly, marketing and decision-making researchers mainly look at the outcome-EXP (the quality of the choice, the users' confidence in making the right choice, and the satisfaction with the chosen item (Hostler et al., 2005; Pedersen, 2000; Vijayasarathy & Jones, 2001; Krishnan et al., 2008; Bechwati & Xia, 2003). They rarely take the system or the choice process as focal points of evaluation.
Human-computer interaction (HCI) researchers have traditionally been more comprehensive in the coverage of all objects of user experience of their research. However, this field has a tendency towards formative evaluations such as Think Aloud testing and Heuristic Evaluation (Van Velsen et al., 2008). The results of such evaluations are limited in generalizability, and therefore not the focus of this paper. The available summative evaluations in the HCI field primarily report on system-EXP variables such as system quality and user loyalty (also operationalized as the intention to return), on process-EXP variables such as cognitive effort and competence in using the system, and on outcome-EXP variables such as decision satisfaction (Pu & Chen, 2007; Chen & Pu, 2009; Pu et al., 2008; Bharati & Chaudhury, 2004; Ochi et al., 2010; Hu & Pu, 2009; Hu & Pu, 2011; Jones et al., 2010; Felfernig et al., 2007).
According to our framework (Figure 1), these user experience effects do not stand alone, but are instead part of a larger chain of effects. Specifically, in the following sections we will argue that the user experience (EXP) is caused by objective system aspects (OSA, via SSA) and personal or situational characteristics (PC or SC). Moreover, the experience variables themselves may be structurally related to one another: Bharati and Chaudhury (2004), for instance, showed that the perceived quality of the system (system-EXP) positively influences the decision satisfaction (outcome-EXP). Although Bharati and Chaudhury investigate the objective system aspects in their study (by manipulating the recommender system under evaluation), they do not include the objective system aspects in the analyzed chain of effects.

3.2 From accuracy to user experience

A large part of existing recommender systems research is focused on creating better prediction algorithms, thereby implicitly assuming that better algorithms will lead to a better user experience. Explicitly testing this assumption would require empirical evaluations with real users on real systems. Several researchers in marketing and decision-making conducted such user-centric evaluations of their recommender systems. For instance, they looked at the reduction in choice effort through a recommender system (Häubl et al., 2004; Pedersen, 2000; Vijayasarathy & Jones, 2001; Diehl et al., 2003; Häubl & Trifts, 2000; Hostler et al., 2005). However, they usually compare a recommender system against a system without recommendation features (or a recommender system against no system at all), rather than looking at the often subtle differences between algorithms. The results of such unbalanced comparisons, in which the "personalized" condition clearly has an advantage over the non-personalized condition, are usually unsurprising (see Van Velsen et al., 2008). However, Chin (2001) argues that this advantage is not always apparent and that a comparison with a non-personalized system may very well be justified. Some researchers compare a recommender system against human recommenders (Krishnan et al., 2008; Senecal & Nantel, 2004), but these studies provide little insight into the effect of algorithm accuracy on the user experience; for that, several algorithms should be pitted against each other.

Surprisingly few studies compare algorithms in live experiments with real users. Researchers who do compare the user experience effects of several algorithms find surprising results. In a comparison of six recommender algorithms, McNee et al. (2002) found that although the "Item-Item CF" algorithm provided the best predictions, users rated it the least helpful. Torres et al. (2004) found that although the "CBF-separated CF" approach had the lowest predictive accuracy among five algorithms, this approach resulted in the highest user satisfaction. In other words, the presumed link between algorithm accuracy (an OSA) and user experience (EXP) is all but evident. Our framework allows researchers of recommender systems to take a step beyond algorithmic accuracy (OSA) towards its effects on user experience (EXP).

3.3 Subjective system aspects as mediators

The presented framework indicates that the apparent missing link between algorithm accuracy and user experience can be found in mediation through perception. The link between algorithm accuracy (an OSA) and user experience (EXP) is often weak (Chin, 2001), and can then only be established by including the mediation through the users' perception of the algorithm accuracy (an SSA). In other

words, the framework hypothesizes that users can perceive algorithm accuracy, and that this perception influences the experience (OSA → SSA → EXP).
In light of these hypotheses, existing research has established that users are, in several instances, able to observe objective differences in recommendation quality (in terms of the framework: OSA → SSA; for examples, see Ziegler et al., 2005; Cosley et al., 2003). It is however not clear how these (typically subtle) differences in perceived recommendation quality affect the user experience (SSA → EXP), because few researchers have tested the effect of their algorithms on the users' perception, behavior *and* experience.
This gap in existing research (i.e. not measuring the SSA as a mediator between OSA and EXP) makes it hard to explain why in some experiments better algorithms do not lead to a better experience. One possible reason might be that users were not able to notice the quality differences (the OSA does not affect the SSA), e.g., the quality differences between two algorithms may have been too small to notice, and thus would not influence the user experience. Another possible explanation (which is not mutually exclusive) might be that users may have observed the quality differences, but may just not have been influenced by these differences in their experience (no link between SSA and EXP), e.g., they may actually like to see good recommendations accompanied with some bad ones, as it makes their final decision easier to justify. Finally, an effect of accuracy on experience may exist, but just be overshadowed by individual differences of perception (this is not unlikely for recommender systems, as their effect is not equally pronounced for each user). In such case one should measure whether the user actually noticed the quality difference or not (SSA is needed as a mediator between OSA and EXP).
The inclusion of SSAs may thus increase the *robustness* of the effects of the OSAs on EXP. Moreover, SSAs provide a more thorough understanding of *how and why* certain features of a recommender system affect the user experience. This does not only hold for algorithm accuracy, but for any manipulated objective system aspects. Chen & Pu (2009) have created a chain of effects from objective algorithm accuracy (OSA) to user-perceived algorithm accuracy (SSA), from objective user effort (OSA) to user-perceived effort (SSA), and from the perceived accuracy and effort (SSA) to intention to purchase and intention to return (INT). They find significant differences between their two tested interfaces for each of these constructs. This path analysis is an important step towards an integrated analysis of recommender system experience, but in our approach we extend it in two directions: First of all, we include the manipulations that cause the objective differences in the model. In the Chen & Pu study they cause significant differences in each construct individually, but an inclusion of the manipulation as a dummy variable into the path model would allow for a mediation analysis of the experimental effects. Secondly, we add constructs explicitly asking the users about their experience (EXP).

3.4 Triangulation of (logged) behavioral data

Although user experience (EXP) is mainly a subjective phenomenon, its effects will likely be reflected in the users' observable behavior (INT). This idea is a fundamental property of all theories based on Fishbein and Ajzen's (1975) Theory of Reasoned Action, although we do not take the direction of the effect to be merely one-way. Whereas attitude causes behavior in TRA, our focus on experience (which is a much more interactive concept than attitude) also considers the inverse effect. For example: users who are more satisfied may increase their usage of the system (EXP →INT), while at the same time increased usage may cause an increase in satisfaction (INT → EXP).

Researchers in the field of algorithm accuracy predominantly use behavioral data for their evaluations: they use logged clicks (either item selections or ratings) to train and test their algorithms (Konstan & Riedl, 2012). Researchers in marketing and decision-making also analyze behavioral data, but focus more on decision time, switching behavior after the choice, or total consumption volume (Häubl et al., 2004; Pedersen, 2000; Vijayasarathy & Jones, 2001; Senecal & Nantel, 2004; Hostler et al., 2005; Ho & Tam, 2005; Tam & Ho, 2005; Pathak et al., 2010). A common problem with behavioral data, however, is that they are not always good indicators of users' subjective experience. For instance, Pu and Chen (2006) found that the actual time users spent looking at items in their system did not correlate with users' subjective perceptions, And Spiekermann et al. (2001) found that stated willingness to provide feedback did not correlate with actual feedback behavior.

Another problem with behavioral data is that their interpretation is often troublesome (Van Velsen et al., 2008). For instance, if users stay on a video clip recommendation site for a longer time, does this mean that the efficiency of the system is low (it takes longer for users to find what they want), or that the users enjoy the site more (to the point that they stay longer to watch more clips)? To solve this dilemma, Van Velzen et al. (2008) suggests to "triangulate" the objective behavioral data (INT) with the subjective experience data (EXP) gathered through other methods (e.g. questionnaires).

From a commercial perspective, influencing the users' objective behavior may seem to be the primary objective of recommender systems research, such as getting the user to buy more products (in e-commerce recommenders) or watch more advertisements (in media recommenders). However, experience concepts reflect and influence users' attitudes towards a system, and research shows that positive attitudes are related to increased adoption rates (Fishbein & Ajzen, 1975; Davis et al., 1989; Venkatesh et al., 2003). To get an indication of the longer-term effects of the system, behavioral data should thus be complemented with subjective

experience measurements. In our framework, behavioral data is therefore correlated (triangulated) with subjectively measured experience concepts.

3.5 Personal and situational characteristics in context

The user experience cannot be entirely attributed to the recommender system itself, it may also depend on characteristics of the user (Personal Characteristics, or PC) and the situation in which the user is using the system (Situational Characteristics, or SC) (Chin, 2001). These factors are typically beyond the influence of the recommender system, but do influence the user experience.

Domain knowledge (or 'expertise') is an important PC variable in this respect: Kamis and Davern (2004) show that participants with a higher level of domain knowledge perceive recommender systems as less useful and harder to use than novices. In a music recommender experiment, Hu and Pu (2010) show that expert users perceive recommendations as less accurate, and the system as less helpful. They also state that they would use the system less. Overall, users with a moderate level of expertise rate the system as most effective.

Users' trust in the system may also influence their experience, and vice versa. Komiak and Benbasat (2006) show that good recommendations can increase trust in both the competence and the integrity of a recommender system, and that a higher level of trust eventually leads to an increased intention to adopt the system. Wang and Benbasat (2007) show that trust in recommender systems is furthermore caused by disposition (the user's initial level of trust), calculation (the user's estimation of the costs and benefits for the system to be trustworthy), interaction (the user's expectations about the system, control over the system, and validation of the system results) and knowledge (an inference based on what the user knows about the system). This makes trust both a PC (depending on the user's personality) and an SC variable (depending on the user, the system and the situation).

Very few studies have investigated which personal and situational characteristics exactly motivate and inhibit users to provide preference feedback to the system (see Pommeranz et al., 2012, for a notable exception). This is an important issue, as many recommender systems rely on explicit feedback (e.g. users' ratings) to give good recommendations. Privacy concerns may reduce users' tendency to disclose personal information (Teltzrow & Kobsa, 2004; Chellappa & Sin, 2005; Berendt & Teltzrow, 2005; Ackerman et al., 1999). On the other hand, if it positively influences their user experience (i.e. in terms of better recommendations), users may be more willing to provide feedback (Spiekermann et al., 2001; Brodie et al., 2004; Kobsa & Teltzrow, 2005).

The main shortcoming of existing research on personal and situational characteristics is that these characteristics are often investigated in isolation (again, see Pommeranz et al., 2012, for a notable exception). This makes it hard to evaluate the impact of these characteristics on the user experience relative to other possible factors that influence the user experience (e.g. is a certain PC → EXP more substantive than a certain SSA → EXP?), and to prove the effectiveness of possible remedies for negative influences (e.g. can a certain positive SSA → EXP offset a certain negative PC → EXP?).
In our framework personal and situational characteristics influence the user experience, but we explicitly describe such effects in addition to the effects of manipulated system aspects. This allows for judgments of relative importance, which are investigated thoroughly in our empirical evaluations.

3.6 Integration of user interface research

Both industry practitioners and academic researchers have argued that the interface of a recommender system may have far larger effects on users' experience with the recommender than the recommender's algorithmic performance (McNee et al., 2006; Baudisch & Terveen, 1999; Murray & Häubl, 2008; Xiao & Benbasat, 2007; Ziegler et al., 2005; Ozok et al., 2010). Below we provide a brief overview of user interface aspects (OSA) influencing the user experience (EXP) of recommender systems.

3.6.1 Preference elicitation method

The preference elicitation method is the way in which the recommender system discovers what the user likes and dislikes. In content-based recommender systems, users may indicate their preference by assigning weights to attributes (Häubl et al., 2004; Kramer, 2007), prioritizing user needs (Felix et al., 2001; Stolze & Nart, 2004; Hu & Pu, 2009) or critiquing examples (Pu et al., 2008; Pu & Chen, 2006; Chen & Pu, 2012; Viappiani et al., 2008; Viappiani et al., 2006). The research on these different preference elicitation methods shows that they have a substantive impact on the user experience (Chen & Pu, 2012). Moreover, the optimal preference elicitation method may depend on user characteristics such as domain knowledge (Knijnenburg & Willemsen, 2009; Knijnenburg & Willemsen, 2010, Knijnenburg et al., 2011).
In collaborative filtering recommender systems, the two most common preference elicitation methods are explicit and implicit elicitation. In explicit elicitation, users rate the items with, for example, one to five stars (see Gena et al., 2011, and Pommeranz et al., 2012, for a user-centric exploration of various alternative explicit elicitation methods). In implicit elicitation, preferences are derived from an analysis

of the browsing and selection behavior of users. Research shows that a combination of explicit and implicit elicitation results in a higher recommendation accuracy (Koren et al., 2009), but no research has investigated differences in user experience and behavior between explicit and implicit elicitation[4]. Our framework provides the opportunity to investigate the effects of preference elicitation beyond accuracy.

3.6.2 Size and composition of recommendation sets

Most recommender systems provide an ordered list of recommendations. Whereas a substantial amount of research considers the individual qualities of these recommendations, little research has considered the composition of the list (Hu & Pu, 2011; Chen & Pu, 2012; Ziegler et al., 2005; Cooke et al., 2002). The composition may play an important role in the user experience of a recommender system, because it influences the users' decision-making process through context effects (Simonson & Tversky, 1992; Tam & Ho, 2005).

It is also unclear how many recommendations the system should provide. In conventional choice situations, too few items may restrict the users' freedom of choice, whereas too many items may lead to choice overload (a process-EXP variable, Schwartz, 2004; Iyengar & Lepper, 2000; Scheibehenne et al., 2010). In the context of recommender systems, where all recommended items are highly relevant, this choice overload effect may be even more prominent. When embedded in a user interface, a longer list of recommendations may enjoy the added benefit of attracting more attention (Tam & Ho, 2005).

The order in which recommendations are presented also seems to have an effect on the users' experience and interaction. The consequences of such serial positioning effects are however unclear: Lynch and Ariely (2000) find that sorting by quality reduces price sensitivity in an e-commerce recommender; Diehl et al. (2003), however, find that it increases price sensitivity. Tam & Ho (2005) find that making one recommendation stand out increases user attraction, elaboration, and choice likelihood.

Hu and Pu (2011; see also Chen & Pu, 2012) show that putting a logical structure on the list of recommendations (specifically, categorizing the recommendations to reflect different trade-offs) leads to a higher perceived categorical diversity of the recommendations, a higher satisfaction and decision confidence, and a higher intention to reuse the system and purchase items with it.

---

[4] Algorithms are often designed to handle a specific type of data (e.g. binary, five star rating) and are therefore restricted to a specific preference elicitation method. In practice, they are therefore often treated as one and the same thing. However, to get a more nuanced understanding of the specific effects of algorithms and preference elicitation methods, we make an explicit distinction between these two system aspects.

Ziegler et al. (2005) found that, up to a certain point, sacrificing (actual and perceived) individual recommendation quality in favor of recommendation set diversity can lead to a more positive subjective evaluation of the recommendation set (SSA; see also Bradley & Smyth, 2001). This finding should be compared with other factors influencing the user experience to verify the robustness and extent of this effect. For instance, Willemsen et al. (2011) find that diversification may reduce the choice difficulty (SSA), which further improves the user experience (EXP). Our framework provides a starting point for investigating the impact of the size and composition of recommendation sets on the user experience.
Finally, one can also think about *when* to provide recommendations. Ho & Tam (2005) find that users are most susceptible to be persuaded by recommendations in the earlier stages of the decision process.

3.6.3 Explanations

Another part of the presentation of recommendations is the possibility to explain why certain items are recommended. Herlocker et al. (2000) found that users like explanations in collaborative filtering recommender systems. Studying knowledge-based recommenders, Felfernig et al. (2007) and Cramer et al. (2008, 2008a) show that explanations increase the users' perception of the system's competence (system-EXP) and their trust in the quality of the recommendations (SSA). Tintarev and Masthoff (2012) show that users tend to like personalized explanations (i.e. explanations that highlight facts related to their preferences), but that these may actually be less effective than generic explanations.

3.7 Research opportunities

It is important to realize that our framework is not merely a classification of the important aspects of the user experience of recommender systems. Nor is it a predefined metric for standardized "performance" tests. Instead, it is a generic guideline for in-depth empirical research on the user experience of recommender systems; it conceptually defines a generic chain of effects that helps researchers to *explain why and how* the user experience of recommender systems comes about. This explanation is the main value of the user-centric evaluation of recommender systems (McNee et al., 2006).
The remainder of this paper will describe the empirical evaluations of our own recommender systems: a number of studies that together comprise a preliminary validation of parts of the evaluation framework. In these studies we have tried to repeatedly test a limited set of core variables of our framework. Additionally, the reviewed literature reveals some gaps in existing research in terms of how well it covers the hypothesized relations in our framework. These gaps can be translated

into requirements for our studies; by covering them, our studies provide a more thorough validation of our framework. Specifically, each empirical study will broadly adhere to a subset of the following requirements:

3.7.1 Requirements regarding objective system aspects (manipulations)

1. **Algorithm:** Despite the predominant research interest in algorithms, there is an apparent paucity of knowledge on how algorithm accuracy influences user experience. The main manipulation in most of our studies is the algorithm used for providing recommendations.
2. **Recommendation set composition:** Another important manipulation is the composition of the set of recommendations presented to the user, as this aspect remains largely untreated in existing research.
3. **Preference input data:** Algorithm and interface meet at the point of preference elicitation. In content-based recommenders this topic has been researched extensively. In collaborative-filtering recommenders, however, the topic remains largely untreated, especially when it comes to explicit versus implicit preference elicitation. Several of our empirical evaluations therefore manipulate the type of input (explicit or implicit) used for recommendation.

3.7.2 Requirements regarding subjective system aspects

4. **Perceived aspects as mediators:** Subjective system aspects such as perceived recommendation quality, accuracy and diversity are measured in our studies, as we expect that these perceptions mediate the effect of the objective system aspects on the user experience concepts.

3.7.3 Requirements regarding experience and interaction

5. **User experience evaluation:** System effectiveness (system-EXP), choice satisfaction (outcome-EXP) and usage effort and choice difficulty (process-EXP) will measure the three objects of user experience evaluation where possible. Relations between these EXP variables will be reported.
6. **Providing feedback:** Many recommender systems elicit the users' preferences by analyzing their preference feedback. We therefore extensively analyze the positive and negative antecedents of the users' intention to provide feedback, as well as their actual feedback behavior.
7. **Behavioral data:** To link the attitudinal part of user experience to the users' observable behavior, logging data will be triangulated with subjective concepts.

3.7.4 Requirements regarding personal and situational characteristics

8. **PC and SC:** Important and under-researched personal and situational characteristics such as domain knowledge and privacy concerns are included where possible and useful.

4 Empirical evaluations: validation of the framework

4.1 Background

The framework proposed in this paper was originally developed as an analysis tool for the MyMedia project (Meesters et al., 2008), which is part of the European Commission 7th Framework Programme. The main goal of MyMedia was to improve the state-of-the-art of multi-media recommender systems. A recommender system development framework, the MyMedia Software Framework (Marrow et al., 2009)[5], was created and deployed in real-world applications at four industrial partners. Each partner conducted a field trial with their system, with several aims in mind: technical feasibility, business opportunities, and user experience research. In the current paper we discuss the results of four field trials (FT1-FT4), two conducted using the MyMedia version of ClipClub player, developed by the European Microsoft Innovation Center (EMIC), and two conducted using a web-based TV catch-up service, developed by the British Broadcasting Corporation (BBC). Each trial was designed and evaluated on the basis of the proposed evaluation framework. We amended these trials by conducting two experiments (EX1 and EX2) with a comparatively more controlled but also more artificial quality. The tight control over the users' interaction with the system in these experiments allowed us to consider in more detail the decision-making processes underlying the user experience. The six studies are described in Table 1, and will be discussed in more detail below.

The main goal of the studies is two-fold: To generate new knowledge that fills the gaps in current research, and to validate the proposed evaluation framework. To assist this latter goal, we repeatedly include the recommendation algorithm (an OSA), the perceived recommendation quality (an SSA), and the system's effectiveness (EXP) as core components of our analysis. Furthermore, for each study we discuss the extent to which its results fit the framework.

We realize that the exclusive focus on collaborative filtering recommender systems, all based on the same MyMedia development framework, and all with media content, limits the scope of this validation. Despite this, we believe that the evaluation framework itself has sufficiently generic qualities to apply to recommender systems in general. The framework is based on a broad range of existing research, and does not assume specific operationalizations of the measured

[5] The recommender algorithms used in the studies described here are available in the MyMedia Software Framework, which is available for non-commercial research purposes, http://mymediaproject.codeplex.org, as well as in the open source package MyMediaLite, http://ismll.de/mymedialite.

concepts. Furthermore, the manipulations and main questions in the conducted studies are rather generic, and hence the range of validity of our conclusions can be broadened from multi-media recommender systems to recommender systems in general.

FT1 Emic pre-trial	
System	Adjusted Microsoft ClipClub
Content	Continuously updated database of clips targeted at teenagers
Participants	43 EMIC colleagues and partners
Manipulations	Algorithms: • Random recommendations • Vector Space Model (VSM) algorithm based on explicit feedback
Main questions	Does a system that provides personalized recommendations provide a better user experience than a system that provides random recommendations? What factors influence the users' intention to provide feedback?
FT2 EMIC trial	
System	Adjusted Microsoft Clipclub
Content	Continuously updated database of clips targeted at teenagers
Participants	108 externally recruited "young" participants (targeted mean age of 25)
Manipulations	Algorithms: • General most popular items • Bayesian Personalized Ranking Matrix Factorization (BPR-MF) algorithm based on implicit feedback • VSM algorithm based on explicit feedback Scenario: • Users receive no specific information • Users are told that their ratings are collected and that this data is used to provide better recommendations • Users are told that their behavior is monitored and that this data is used to provide better recommendations
Main questions	What is the difference in subjective recommendation quality between the different algorithms? Does a system that provides personalized recommendations lead to a better user experience than a system that recommends the "generally most popular" items? What is the difference between the implicit recommender and the explicit recommender in terms of user experience? What factors influence the users' intention to provide feedback?
FT3 BBC pre-trial	
System	BBC MyMedia player
Content	BBC television programming (up to one week old)
Participants	59 externally recruited British participants, reflecting a balanced representation of the UK television audience
Manipulations	For the rating trial, algorithms: • General most popular items • BPR-MF algorithm based on implicit feedback • MF algorithm based on explicit feedback For the rating trial, time: • Day 1 … Day 9 For the experience trial, time: • Week 1 • Week 2
Main questions	What is the difference in recommendation list quality between the different algorithms? How does the quality of the recommendation lists generated by the different algorithms evolve over time? How does the user experience of the system evolve over time?

Table 1 continues on the next page

FT 4 BBC trial	
System	BBC MyMedia player
Content	BBC television programming (up to one week old)
Participants	58 externally recruited British participants, reflecting a balanced representation of the UK television audience
Manipulations	Algorithms: • General most popular items • BPR-MF algorithm based on implicit feedback • MF algorithm based on explicit feedback
Main questions	What is the difference in subjective recommendation quality between the different algorithms? Does a system that provides personalized recommendations lead to a better user experience than a system that recommends the "generally most popular" items? What is the difference between the implicit recommender and the explicit recommender in terms of user experience?
EX1 Choice overload experiment	
System	Adjusted BBC MyMedia player
Content	Movies (MovieLens 1M dataset)
Participants	174 participants invited from a panel with students or recently graduated students from several Dutch universities
Manipulations	Recommendation set quality and size: • Top-5 (5 best recommendations) • Top-20 (20 best recommendations) • Lin-20 (5 best recommendations, recommendation ranked 99, 199, … 1499)
Main questions	How do the objective quality and size of the recommendation set influence the subjective recommendation set quality and diversity? How do the objective quality and size of the recommendation set influence choice difficulty and choice satisfaction?
EX2 Diversification experiment	
System	Adjusted BBC MyMedia player
Content	Movies (MovieLens 1M dataset)
Participants	137 Amazon Turk workers
Manipulations	Algorithms: • General most popular items • MF algorithm based on explicit feedback • K-Nearest Neighbor (kNN) algorithm based on explicit feedback Recommendation set diversification: • No diversification • Some diversification • Lots of diversification
Main questions	What is the difference in subjective recommendation quality between the different algorithms? Does a system that provides personalized recommendations lead to a better user experience than a system that recommends the "generally most popular" items? What is the difference between the kNN recommender and the MF recommender in terms of user experience? Do users notice our manipulation of the variety of the recommendations? Do users like (objective or subjective) variety in their recommendation sets? If so, does the effect overshadow the effect of algorithm accuracy?

Table 1 (continued): Properties of the studies that were conducted based on the evaluation framework.

4.2 Empirical validation techniques

User experience research can be conducted both qualitatively and quantitatively (Preece et al., 2002; Kaplan & Duchon, 1988). Qualitative research is more exploratory, but is usually less generalizable, and cannot be statistically validated. Quantitative analysis allows for statistical validation, but one has to have clear hypotheses about theoretical constructs and their relations before conducting the study. In many cases it is advisable to apply these techniques in tandem, but for the purpose of this paper we will restrict our analysis to quantitative results (hypothesis for our studies are conveniently provided by the framework). In order to prevent confirmation bias in our validation of the framework, we extensively use data analysis methods like exploratory factor analysis (EFA) and structural equation modeling (SEM), which allow for a more exploratory analysis of quantitative data. Below we discuss the general procedure of our research; appendix A provides a more in-depth description. We recommend researchers who want to use our framework to either follow a similar procedure, or to opt for our pragmatic version described elsewhere (Knijnenburg et al., 2011a).

4.2.1 Measurement

Experience, SSAs, PCs and SCs can be measured using questionnaires. To assure a more robust measurement of all concepts, we typically use a minimum of seven statements (both positively and negatively phrased) that can be answered on a balanced 5- or 7-point scale (from “completely disagree” to “completely agree”) for each unidimensional concept. Exploratory factor analyses can be conducted to test the robustness of the concepts, and to exclude any questions that do not contribute to the measurement of the intended concepts.

4.2.2 Manipulation

Objective system aspects (OSA) can be manipulated, i.e., several versions of the aspect (conditions) can be created, and assigned randomly to each participant. By keeping everything else constant between conditions, one can single out the effect of this manipulated aspect on the user experience (EXP). In our research, one condition always serves as a baseline against which all other conditions are compared[6]. Furthermore, by manipulating several system aspects independently, one can compare the relative impact of these aspects on the user experience (e.g. "Does the effect of the user interface overshadow the impact of different algorithms?").

[6] Similar to the use of dummy variables in standard linear regression

4.2.3 Structure

The incorporation of user perceptions (SSA) increases the robustness and explanatory power of the evaluations, and different objects of user experience (system-EXP, process-EXP, outcome-EXP) can be treated in parallel to gain further insight in the workings of user experience. Logged behavioral data can be triangulated with the user experience concepts to link the subjective experience (EXP) to objective user behavior (INT). This creates a causal chain of effects from manipulated OSAs, via subjectively measured SSAs and EXPs, to objectively measured INTs. Structural equation modeling (SEM; Muthen, 1984) can be used to conduct a mediation analysis on these effects. From a statistical perspective, SEM concurrently tests the robustness of the measured constructs and the relationships between them.

4.2.4 Graphical presentation of SEMs

We graphically present our structural equation models as diagrams containing the constructs (boxes) and the relationships between them (arrows). Rectangular boxes are latent constructs based on the questionnaire items. For the sake of clarity, we do not include the questionnaire items themselves in the diagrams. Elliptical boxes are behavioral metrics, extracted from the systems' data logs. One-headed arrows represent regression coefficients (which have a causal direction); double-headed arrows represent correlation coefficients.

The color of the boxes matches the colors in the framework (Figure 1); each color signifies a specific type of construct (purple: OSA, green: SSA, orange: EXP, blue: INT, red: PC, light blue: SC). The numbers on the arrows represent the regression coefficients (i.e. the strength of the structural relations, also represented by the thickness of the arrows), their standard deviation (between parentheses) and the statistical significance. Non-significant relations are not included in the graphs.

4.3 FT1 EMIC pre-trial

The EMIC pre-trial was conducted to confirm two basic premises for the success of recommender systems in general: the premise that the user experience of a recommender is influenced by the recommendations, and the premise that users will provide adequate preference feedback to train the recommender system.

4.3.1 Setup

The trial was conducted with 43 EMIC colleagues and partners who participated in the trial on a voluntary basis (28 male, average age of 31, SD = 9.45). A detailed treatment of this study can be found in Knijnenburg et al. (2010a). We therefore only briefly discuss the results here.

Participants all used a slightly modified version of the MSN Clipclub system (see Figure 2); the special section with recommendations was highlighted, the social networking features were disabled (as to not interfere with the study), an explanation of the rating facilities was included, and participants were probed to provide at least one rating every five minutes (although this rating request could be denied). Participants were randomly assigned to one of two conditions: 25 participants received random clips as recommendations, and the remaining 18 participants received recommendations provided by a content-based Vector Space Modeling engine (i.e. we manipulated the OSA "personalized vs. random recommendations"). Participants were told that providing ratings would change the recommendations[7].

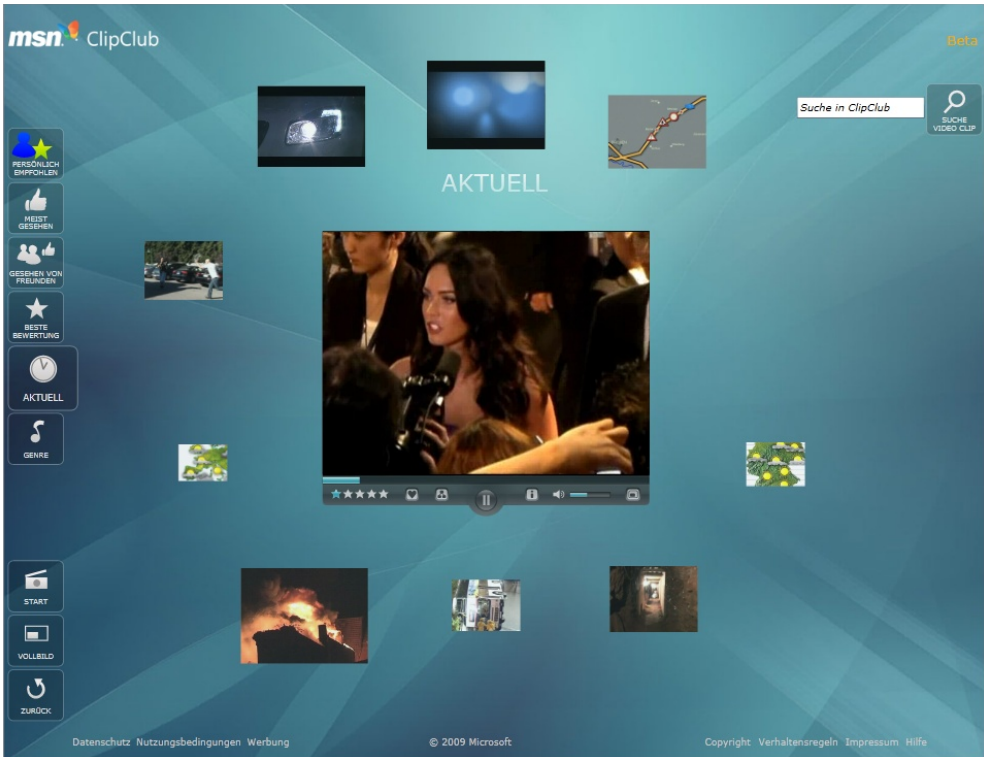

Figure 2: The modified ClipClub prototype.

After half an hour of interaction with the system, several questionnaires[8] were taken to measure the participants' perceived recommendation quality (SSA), choice satisfaction (outcome-EXP), perceived system effectiveness (system-EXP), intention to provide feedback (INT), general trust in technology (PC), and system-specific

[7] Which is true for both conditions, but in the random condition these new recommendations would just be another batch of random items.

[8] A complete overview of the questionnaires used in the studies can be found in appendix B.

privacy concerns (SC). The questions were factor-analyzed to produce metrics for these concepts, and factor scores were included together with the manipulation in a regression path model (Figure 3). Additionally, factor scores were correlated with behavioral metrics obtained from click stream logs.

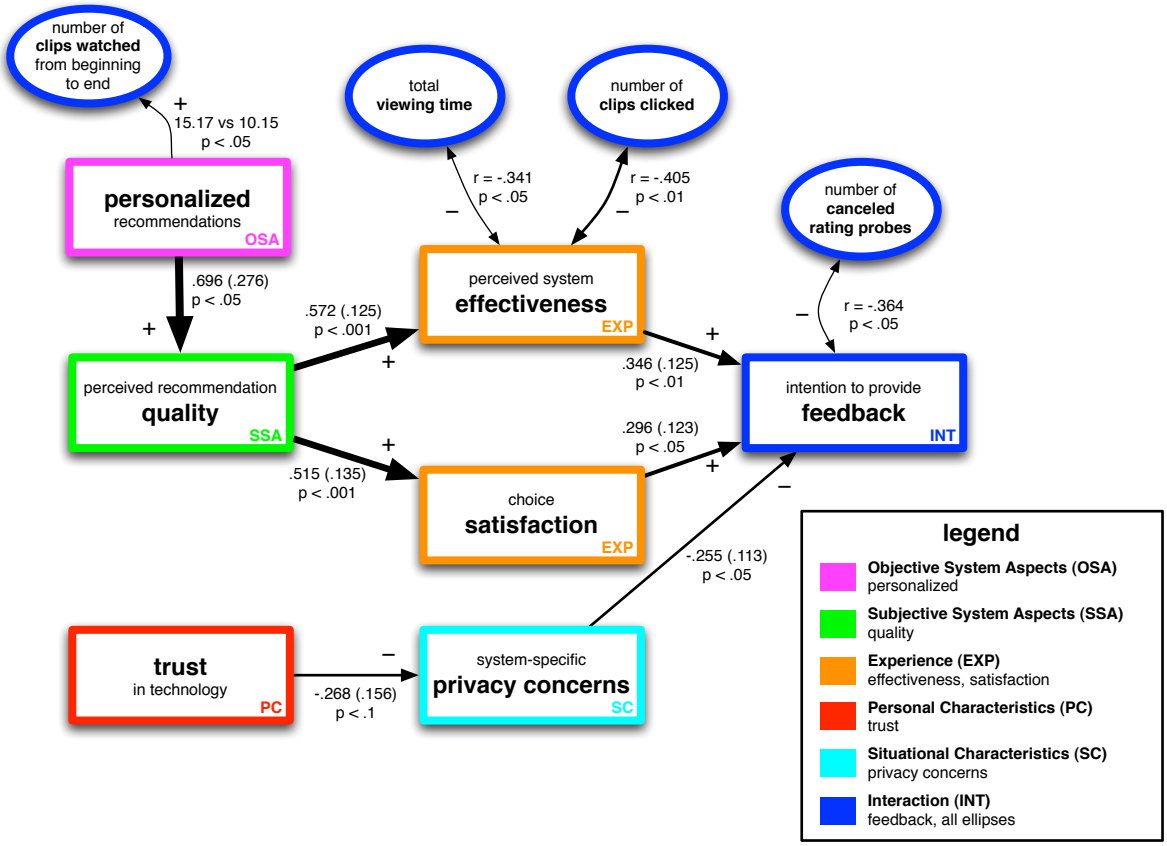

Figure 3: The path model constructed for FT1 - EMIC pre-trial. Please refer to section 4.2.4 "Graphical presentation of SEMs" for interpretation of this figure. Note that the rectangle "personalized recommendations" represents the difference between personalized and random recommendations, random being the baseline model.

4.3.2 Results

The results show that personalized recommendations (as compared to random recommendations) have a higher perceived quality (OSA → SSA), which leads to a higher choice satisfaction (SSA → outcome-EXP) and system effectiveness (SSA → system-EXP). Behavioral data corroborates these hypotheses. Users of the system with personalized recommendations watch a larger number of clips from beginning to end (OSA → INT). Moreover, users who click fewer clips and have a lower total viewing time rate the system as more effective (EXP ⇆ INT), which indicates that a higher system effectiveness is related to reduced browsing activity (see discussion below).

The intention to provide feedback increases with choice satisfaction and system effectiveness (EXP → INT) but decreases when users have a higher system-specific privacy concern (SC → INT), which in turn increases when they have a lower trust in technology (PC → SC)[9]. In terms of behavior, the number of canceled rating probes (popping up after five minutes without rating) is significantly lower in the personalized condition than in the random condition (OSA → INT), and is also negatively correlated with intention to provide feedback (INT → INT).

4.3.3 Discussion of results

The results show that a system with a recommender algorithm that provides personalized recommendations has a better user experience (in terms of both choice satisfaction and system effectiveness) than a system that provides random recommendations. This is not necessarily a surprising result; such a comparison between random and personalized recommendations is a bit unbalanced (Van Velsen et al., 2008).

More interestingly, however, the path model indicates that this effect is indeed mediated by perceived recommendation quality (requirement 4 in section 3.7). Furthermore, there is no residual correlation between choice satisfaction and system effectiveness, and a mediated variant provides a weaker model than the one described. In other words, in this study there was no structural relation between these experience variables (requirement 5).

Users seem to base their intention to provide feedback on a trade-off between having a better user experience and maintaining their privacy (requirement 6 and 8). However, the intention to provide feedback was not correlated with the total number of ratings, indicating that the relation between intention and behavior can be very weak (a well-known psychological phenomenon called the 'intention-behavior gap'; Sheeran, 2002).

User behavior is correlated with the experience variables (requirement 7), but at times also directly with the manipulation of our study. Against our intuition, users who rate the system as more effective have a lower total viewing time and a lower number of watched clips. However, the results also show that at the same time, the number of clips watched from beginning to end is higher in the personalized condition than in the non-personalized condition. The lower total viewing time and number of clicked clips thus reflects a reduction in browsing, not consumption. This makes sense, because recommendations supposed to be an alternative to browsing in this system. The outcomes clearly demonstrate the value of triangulating the behavioral measures with subjective measures. Showing how behavioral measures

[9] An effect of a personal characteristic on a situational characteristic is not explicitly predicted by our framework, but in this case makes perfect sense.

are related to experience variables allows researchers to assign meaning to these measurements, which at times may even counter the researchers' intuition. Moreover, it grounds our subjective theory in observable behavior.

4.4 FT2 EMIC trial

The EMIC extended trial was conducted to reproduce and extend the effects of the pre-trial. Taking a step beyond the unbalanced comparison in the pre-trail between personalized and random recommendations, the extended trial looked at differences between the non-personalized "generally most popular" items (GMP condition), the VSM algorithm using explicit feedback as input (the same VSM condition as used in the pre-trial) and the Bayesian Personalized Ranking Matrix Factorization (BPR-MF; Rendle et al., 2009), a state-of-the-art algorithm using all clicks to predict recommendations ('implicit feedback'; MF-I condition). Furthermore, we specifically controlled what the users were told about the systems' use of their feedback: nothing (none condition); that their rating behavior was being used to provide better recommendations (rating behavior condition); or that all their behavior was being used to provide better recommendations (all behavior condition). We hypothesized that this manipulation would influence users' privacy concerns.

4.4.1 Setup

An external company recruited German participants from a young demographic (targeted mean age of 25). Participants were instructed to use the system as many times as they liked over the 20-day duration of the study, and they were asked to fill out a 47-item user experience questionnaire after each session (which would comprise at least 20 minutes of interaction). The external company paid the participants for their cooperation. Participants were randomly assigned to a scenario (i.e. the OSA "scenario" [no story / rating behavior / all behavior] is a between subjects manipulation), and after each questionnaire they would switch algorithms (i.e. the OSA "algorithm" [GMP / VSM / MF-I] is a within subjects manipulation). Participants were informed of this switch, but were not told which algorithm they would be using. Users used the same system as in FT1 (see Figure 2), but in contrast to the first field trial, there were no explicit rating requests in this trial, and there was also no specific section with recommendations; instead, all categories that the user could navigate to (e.g. sport, gossip, cars, news) showed a personalized subset of the items in that category. The behavior of each participant was logged, allowing for an extensive analysis of the click-stream of each user.
The trial yielded 430 questionnaires from 108 participants. After excluding all questionnaires with fewer than 12 clicks (indicating insignificant usage), 258

questionnaires (60%) remained from 95 remaining participants (88%). These participants had an average age of 27.8 (SD = 4.70). 49 of them were male.
The questions in the questionnaires were first submitted to an exploratory factor analysis (EFA) to determine whether their covariances naturally reproduced the predicted constructs. This resulted in 7 factors:

- Perceived recommendation quality (6 items, e.g. "I liked the items shown by the system", factor $R^2 = .009$[10])
- Effort of using the system (3 items, e.g. "The system is convenient", factor $R^2 = .184$)
- Perceived system effectiveness and fun (10 items, e.g. "I have fun when I'm using the system", factor $R^2 = .694$[11])
- Choice satisfaction (5 items, e.g. "I like the items I've chosen", factor $R^2 = .715$)
- General trust in technology (4 items, e.g. "Technology never works", no incoming arrows)
- System-specific privacy concerns (3 items, e.g. "The system invades my privacy", factor $R^2 = .333$)
- Intention to provide feedback (4 items, e.g. "I like to give feedback on the items I'm watching", factor $R^2 = .264$).

12 items were deleted due to low communalities and/or unwanted cross-loadings. The items were then analyzed using a confirmatory structural equation modeling (SEM) approach with repeated ordinal dependent variables and a weighted least squares estimator, in which the subjective constructs were structurally related to each other, to the conditions (algorithm and scenario), and to several behavioral measures extracted from the usage logs. The final model had a reasonable model fit ($\chi^2(41) = 85.442$, $p < .001$, CFI = .977, TLI = .984, RMSEA = .065)[12]. Figure 4 displays the effects found with this model.

---

[10] R^2 values are taken from the final SEM and not from the EFA. The R^2 for perceived recommendation quality was low because it was predicted by the algorithm condition only. We typically report the best fitting item as an example for the scale, full questionnaires can be found in appendix B.

[11] Fun was intended to be a separate construct, but the exploratory factor analysis could not distinguish this construct from perceived system effectiveness. In other words, in answering our questionnaires participants did not seem to conceptually distinguish these two constructs.

[12] Agreed-upon model fit requirements are described in more detail in appendix A.

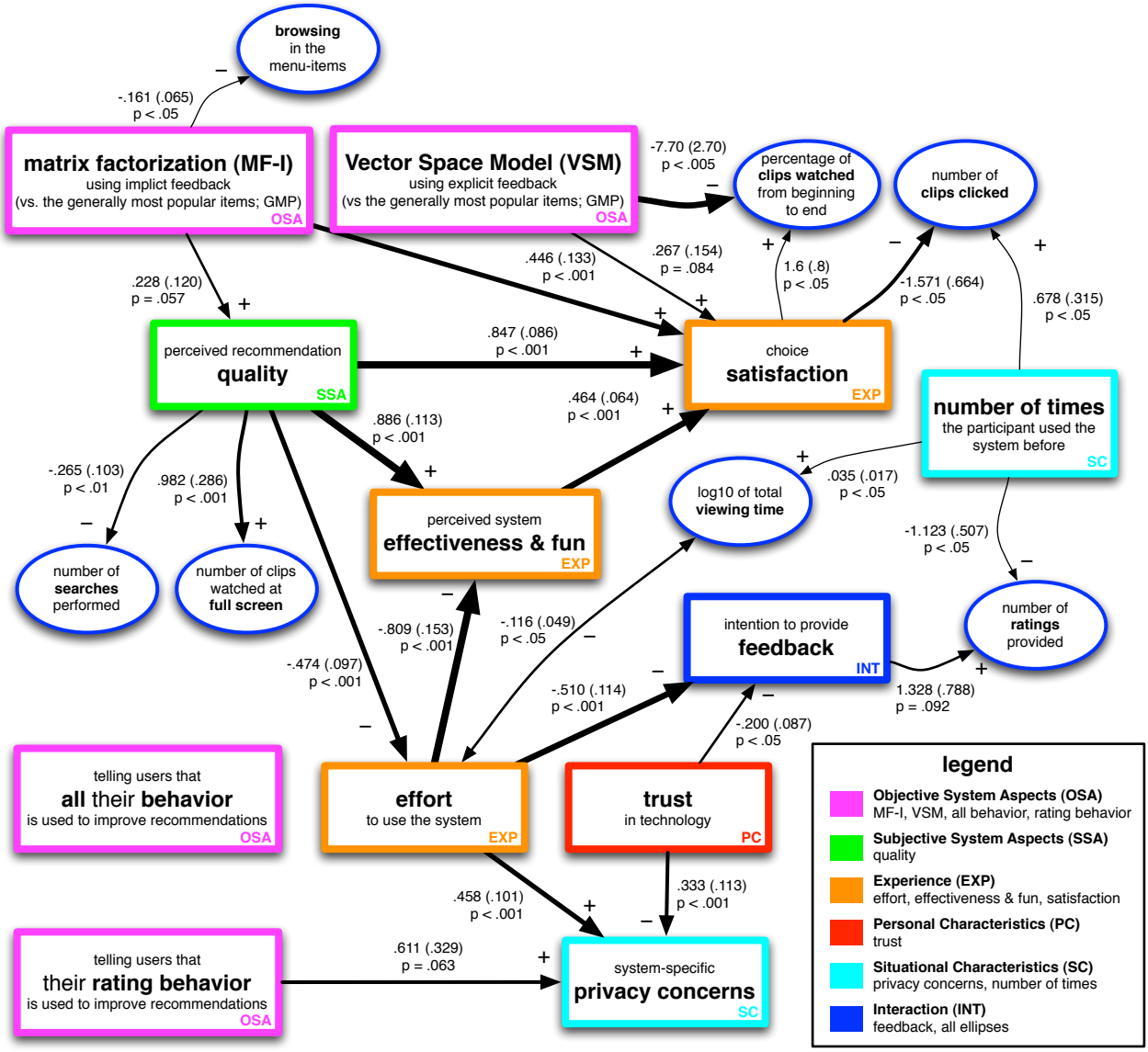

Figure 4: The path model constructed for FT2 - EMIC trial. Please refer to section 4.2.4 "Graphical presentation of SEMs" for interpretation of this figure. Note that the "algorithm" manipulation is represented by the two rectangles "Matrix Factorization (MF-I)" and "Vector Space Model (VSM)", which are tested against the non-personalized baseline "generally most popular (GMP)". Likewise, the "scenario" manipulation is represented by the two rectangles "all behavior" and "rating behavior", which are tested against the baseline "no story".

4.4.2 Results

The model shows that the recommendations from the Matrix Factorization algorithm (MF-I) have a higher perceived quality than the non-personalized "generally most popular" (GMP) items (OSA → SSA). Higher perceived recommendation quality in turn leads to lower effort (SSA → process-EXP), a higher perceived effectiveness and fun (SSA → system-EXP), and a higher choice satisfaction (SSA → outcome-EXP). The effect of the algorithm on choice satisfaction is however only partially mediated; there is also a direct positive effect of the MF-I

algorithm on choice satisfaction (OSA → EXP). The vector space modeling algorithm (VSM) does not even affect perceived quality at all; it only has a direct effect on choice satisfaction (increasing it marginally; $p < .10$). Lower effort leads to more perceived effectiveness and fun (process-EXP → system-EXP; this is in line with Pommeranz et al., 2012), which in turn leads to more satisfactory choices (system-EXP → outcome-EXP).

The effort required to use the system influences the intention to provide feedback: the less effort users have to invest, the more they are willing to provide feedback (process-EXP → INT). Privacy concerns also increase when the system takes more effort to use (process-EXP → SC)[13]. Figure 5 displays the tendency of marginal effects of the different algorithms on recommendation quality, system effectiveness & fun, and choice satisfaction. The graphs show the standardized difference between the algorithms (MF-I and VSM) and the baseline condition (GMP).

Figure 4 shows a direct effect of trust in technology on users' intention to provide feedback (PC → INT). Trust in technology also reduces privacy concerns (PC → SC). Telling users that their rating behavior is used to improve recommendations increases their privacy concerns (OSA → SC)[14]. Notably, telling users that all their behavior is being used to improve recommendations does not have this effect.

In terms of behavioral measures, we find that, like in the pre-trial (FT1), a high choice satisfaction decreases the number of clips clicked (EXP → INT). Users who perceive the recommendations to be of a higher quality also perform fewer searches (SSA → INT), and users browse less between different categories when the MF-I algorithm is used (OSA → INT). Users who are more satisfied with their choices watch more clips from beginning to end (EXP → INT). Interestingly, fewer users do so when the VSM algorithm is used (OSA → INT). Users who perceive the recommendations to be of a higher quality also watch more clips at full screen (SSA → INT). Furthermore, intention to provide feedback increases the number of ratings provided by the user, something we did not find in FT1. Finally, in later sessions the number of clips clicked and the viewing time increase (SC → INT). Participants also provide fewer ratings in later sessions (SC → INT; this is in line with Harper et al., 2005).

---

[13] Our framework does not predict an effect of experience on situational characteristics. However, we understand that privacy concerns may be higher when users have to put more effort into using the system, because this usually means that they also have to provide more information to the system.

[14] Our framework does not predict an effect from an objective system aspect on a situational characteristic to exist. However, this particular OSA was specifically designed to change this particular SC.

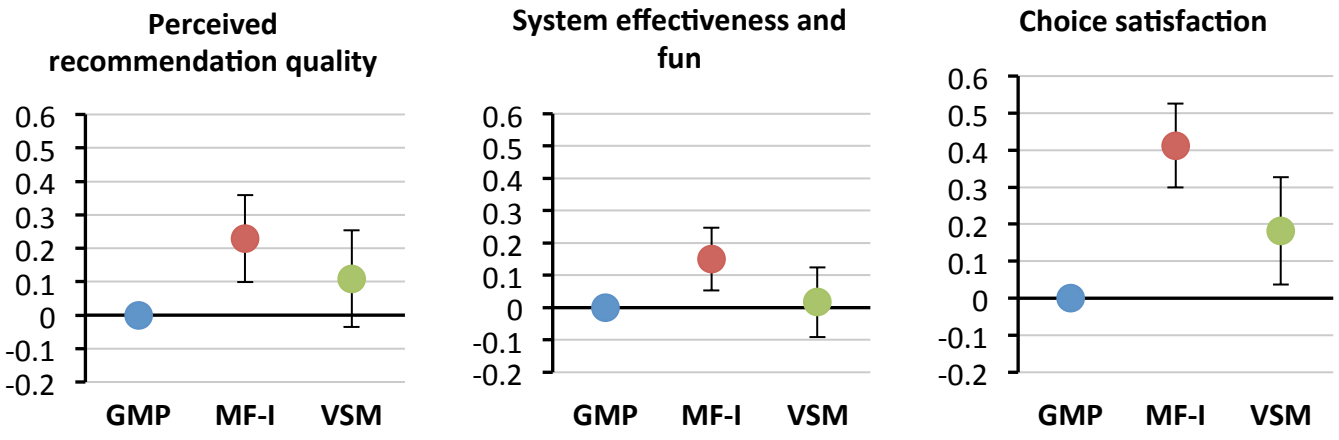

Figure 5: Tendency of marginal (direct and indirect) effects of the "algorithm" condition on perceived recommendation quality, system effectiveness and fun, and choice satisfaction. Error bars indicate +/-1 standard error compared to the value of GMP, which is fixed to zero. Scales of the vertical axes are in sample standard deviations. The Matrix Factorization algorithm (MF) provides higher perceived recommendation quality and a higher choice satisfaction than the non-personalized "most popular" algorithm (GMP); the Vector Space Modeling algorithm (VSM) does not provide a significantly better experience.

4.4.3 Discussion of the results

In general terms, this study confirms the structural relations found in FT1, but there are a number of important differences and additional insights. Specifically, the study allows us to compare two different algorithms with a non-personalized baseline (requirement 1, see section 3.7). The results indicate that the BPR-MF recommendations of the MF-I condition are most successful; these have a marginally higher perceived quality and lead to a higher choice satisfaction (see Figure 5). The VSM recommendations only lead to a marginally higher choice satisfaction. A possible reason for these differences is that the BPR-MF algorithm used implicit feedback, which can be gathered more quickly than the explicit feedback used by the VSM algorithm (requirement 3).

The inclusion of the construct "effort of using the system" changes part of our path model compared to FT1: Whereas in FT1 choice satisfaction and perceived system effectiveness increased users' intention to provide feedback, we now observe that it is actually the effort of using the system that causes this effect (requirement 6). This difference between FT1 and FT2 indicates the importance of measuring all aspects of the user experience (requirement 5) as this allows us to better understand the nature of the difference.

The trial confirms that privacy and trust are important factors influencing the intention to provide feedback (requirement 6), although the effects are now structurally different. Like in FT1, trust in technology reduces privacy concerns, but it now also directly increases the intention to provide feedback (instead of a

mediated effect via privacy concerns). Due to the increased power of the study, FT2 enables us to measure the link between feedback intention and behavior.
We also observe that telling users that their rating behavior is used (compared to telling them nothing, or that all their behavior is used) increases their privacy concerns. This is surprising, because using all behavior (including rating behavior) is by definition more intrusive than using only rating behavior. The wording in the two conditions is almost the same: "In order to increase recommendation quality, we are going to use [your ratings data] / [all your activities]". One possible reason for the heightened privacy concerns in the "rating behavior" scenario is that users in this scenario have active control over the amount of information they provide to the system. They are thus encouraged (or forced) to make an explicit trade-off between the potential usefulness of providing information and the effect this may have on their level of privacy. In the "all behavior" condition, there is nothing the users can do to prevent the system from analyzing their behavior, and it is also difficult to anticipate the privacy effects of exhibited and forborne behavior. Users may therefore be less sensitized with regard to privacy concerns, or even forget that the system is continuously analyzing their behavior.
By triangulating our behavioral data with the subjective constructs (requirement 8), we are also able to revalidate the finding from FT1 that a good experience means: "less browsing, but more time enjoying the content". Finally, the extended period of use allows us to look at the effect of familiarity with the system. Specifically, users are exploring the system to a further extent in later sessions, which may indicate less reliance on the recommendations.

4.5 FT3 BBC pre-trial

The BBC pre-trial used a special-purpose recommender system to present users with recent programming of the BBC (see Figure 6) and strove to compare three different algorithms side by side. As FT2 revealed that usage over time changes (i.e., we observed less reliance on the recommendations over time), the pre-trial was conducted to investigate how recommendation quality would evolve over time.

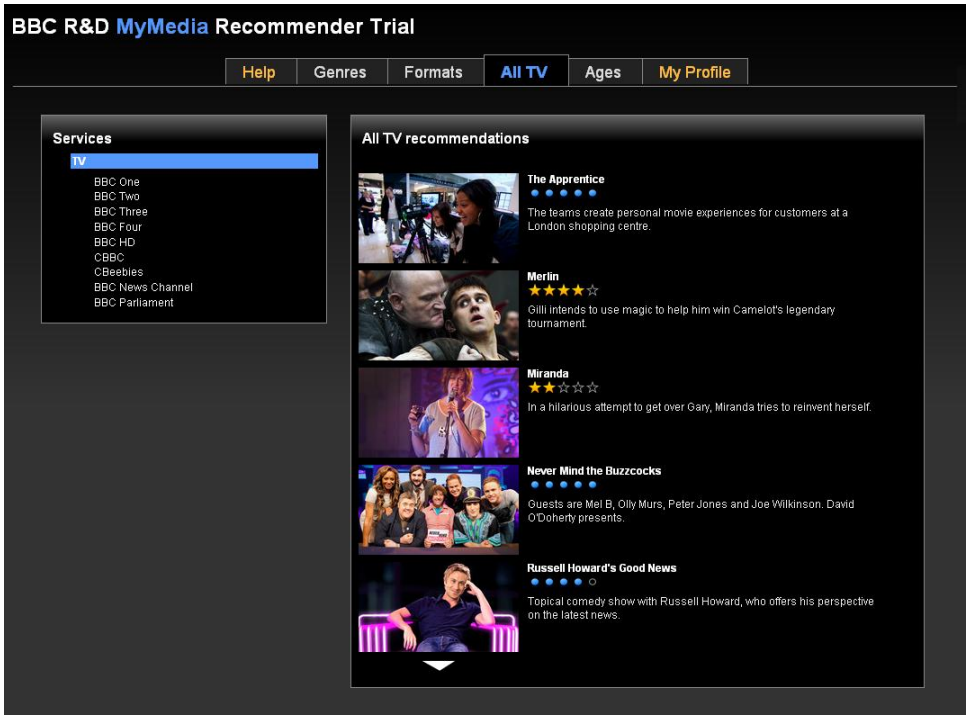

Figure 6: The BBC MyMedia recommender prototype.

4.5.1 Setup

An external market research company recruited 59 British participants to participate in both FT3 and FT4. The participants were selected to reflect a balanced mix of ages (one third between 18 and 35, one third between 36 and 50, and one third between 51 and 66). Within each age group, the selection included equal numbers of males and females. The selected sample also consisted of equal numbers of participants from five regions across the UK. Finally, half of the participants were occasional users of the existing TV/radio catch-up service, the other half were frequent users. Participants were paid a small fee for each daily task, and a bonus for each completed user experience questionnaire. Over a period of two weeks (not counting weekends) participants performed two daily tasks. Specifically, the participants were asked every day to rate a list of recommendations presented by three different algorithms (rating task), and to use the system freely (free use task). The participants had an average age of 41.5 (SD = 12.6). 27 of them were male.
The algorithms used in the rating task were the non-personalized 'generally most popular' items (the GMP condition), and two personalized Matrix Factorization algorithms. One is described in Rendle and Schmidt-Thieme (2008) and relies on user ratings to provide recommendations ('explicit feedback'; the MF-E condition). The other, BPR-MF (Rendle et al. 2009), relies on all clicks in the interface ('implicit feedback'; the MF-I condition). Every day, each participant rated three lists (one

from each of the algorithms) of five recommended programs. The order in which the lists were presented changed every day.
The system in the 'free use' task employed the MF-E algorithm for all participants throughout the entire two-week period. After both weeks, participants filled out a questionnaire asking about their user experience with the 'free use' system.

4.5.2 Results of the rating task

In the rating task users were asked to rate five recommended programs separately, as well as the recommendation list as a whole. As can be seen in Figure 7 below, the average ratings decreased over time (contrast $F(1,15) = 12.7$, $p < .005$, $r = .68$). Either participants gradually became more critical about the recommendations (a psychological phenomenon called habituation), or they were actually confronted with recommendations of a lower quality (e.g. because items previously recommended were not included in the recommendations to prevent repetitiveness).
After an initial training period, at days four to seven (May 20 to 25), we observe significantly higher ratings of the MF-E recommendation lists ($M_{MF\text{-}E} = 3.52$ vs. $M_{GMP} = 2.80$, contrast $F(1,29) = 45.4$, $p < .001$, $r = .78$) as well as the MF-I recommendation lists ($M_{MF\text{-}I} = 3.33$ vs. $M_{GMP} = 2.80$, contrast $F(1,29) = 19.1$, $p < .001$, $r = .63$). In the second week (May 24 to 27), the ratings for the lists provided by MF-E and MF-I started to drop significantly (interaction of MF-E vs. GMP with linear decreasing time: $F(1,32) = 6.90$, $p < .05$, $r = .42$; interaction of MF-I vs. GMP with linear decreasing time: $F(1,32) = 10.9$, $p < .005$, $r = .50$), going down to the level of GMP. In other words, the quality of the MF-I and MF-E recommendations improves over time, but later falls back to the level of the non-personalized GMP recommendations.

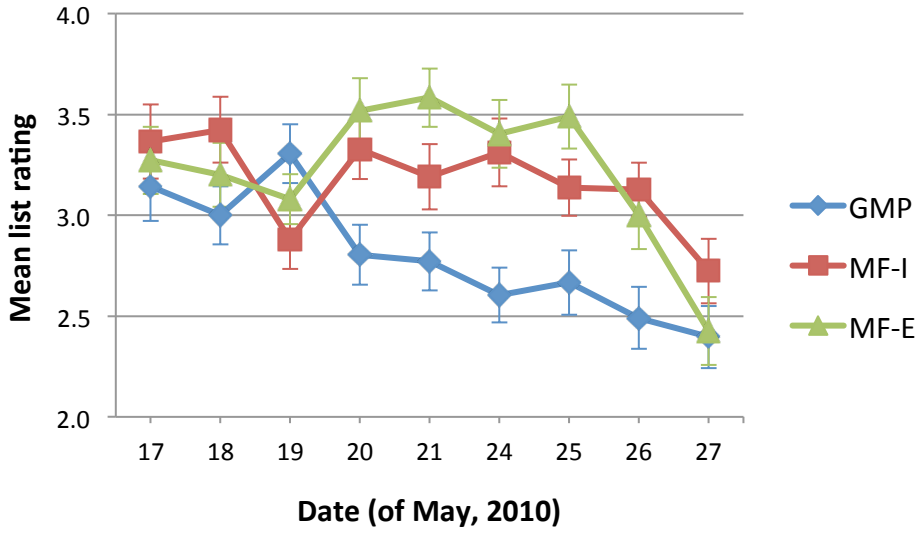

Figure 7: The means of the list ratings, over time, of the recommendation lists for the generally most popular items (GMP) and the Matrix Factorization algorithm recommendations based on explicit (MF-E) and implicit (MF-I) feedback. The error bars present +/-1 Standard Error.

4.5.3 Results of the free use task

To analyze the free use task, we compared the outcomes of the user experience questionnaires collected at the two different time points (i.e. the SC "time" [end of week 1 / end of week 2], can be seen as a within subjects manipulation). All 59 users partook in this part of the study, but 8 users only completed the first week's questionnaire, resulting in 110 data points. The results of the experience questionnaire (17 questions) were first submitted to an exploratory factor analysis (EFA) to see if their covariances naturally reproduced the predicted constructs. This resulted in three factors:

- Perceived recommendation variety (2 questions, "The recommendations contained a lot of variety" and "All the recommended programmes were similar to each other", factor $R^2 = .052$)
- Perceived recommendation quality (7 questions, e.g. "The recommended items were relevant" and "I liked the recommendations provided by the system", factor $R^2 = .148$)
- Perceived system effectiveness (6 questions, e.g. "I would recommend the MyMedia recommender to others" and "The MyMedia recommender is useful", factor $R^2 = .548$).

Two questions were deleted due to low communalities. The items were then analyzed using a confirmatory SEM approach with repeated ordinal dependent variables and an unweighted least squares estimator, in which the subjective constructs were structurally related to each other and to the dichotomous independent variable 'time' (week 1 vs. week 2). The final model had a good model

fit ($\chi^2(22) = 27.258$, p = .20, CFI = .982, TLI = .984, RMSEA = .047). Figure 8 displays the effects found with this model.

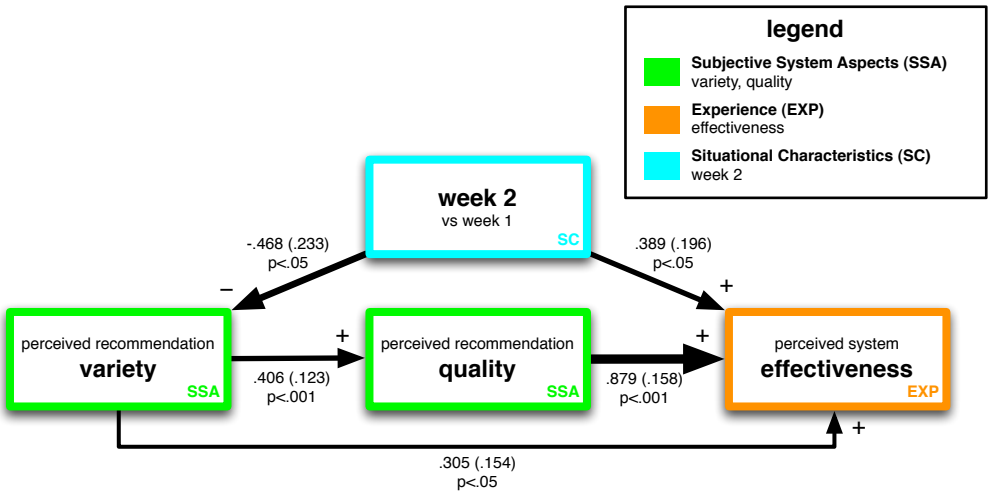

Figure 8: The path model constructed for FT3 - BBC pre-trial. Please refer to section 4.2.4 "Graphical presentation of SEMs" for interpretation of this figure.

The model shows that a higher perceived variety of the recommendations causes users to perceive the recommendations as having a higher quality, and that higher quality recommendations in turn cause an increased perceived system effectiveness (SSA → SSA → EXP). There is also a direct effect of perceived variety on perceived system effectiveness (SSA → EXP); the perceived quality does not fully mediate the effect of variety on effectiveness.
Additionally, in this study the variety of the recommendations turned out to be significantly lower in the second week (SC → SSA), which in turn led to a lower perceived quality of the recommendations in the second week. This is in line with the drop in recommendation list ratings as found in the rating task (see Figure 7). The lower perceived recommendation quality, in turn, decreased the perceived effectiveness of the system. However, time also had a positive direct effect on perceived system effectiveness (SC → EXP), which cancelled out the negative indirect effect (the mean system effectiveness did not differ significantly between week 1 and week 2).

4.5.4 Discussion of the results

The drop in perceived recommendation quality in the second week was consistent between the rating and free use task. The SEM model explains that this happened because the set of recommended programs was less varied in the second week. Arguably, perceived variety drops because TV programs repeat after the first week, at which point they resurface among the recommendations. Interestingly, due to a direct positive effect of time on perceived system effectiveness, these effects in the

end did not reduce the user experience. An ad hoc explanation of this effect could be that although users may have felt that recommending new episodes of previously recommended shows is not a sign of high quality recommendations, they may at the same time have appreciated the reminder to watch the new episode. The study thus clearly shows the merit of measuring subjective concepts in addition to user ratings in understanding the underlying causes of rating differences.

4.6 FT4 BBC trial

The BBC trial was primarily conducted to test the effect of explicit versus implicit preference elicitation methods on the user experience of the system.

4.6.1 Setup

The trial used the same system as the 'free use' system in the pre-trial, but now the algorithm that provided the recommendations in this system was changed every three days between GMP, MF-I and MF-E until each algorithm was used once by each user (see section 4.5.1 for more details on these conditions). In each condition users were given the same background on the system, specifically, no information was given on the way recommendations were computed or on what type of information they were based. Participants were randomly assigned to one of three groups, for which the order of the algorithms was manipulated in a Latin square design, so that each algorithm was being used by one third of the users at any time. 58 users (the same as those in the pre-trial) completed this study, resulting in 174 data points. The user experience questionnaire was similar to the pre-trial, but two questions were added to gain a more robust measurement of recommendation variety (now measured by 4 items). The exploratory factor analysis resulted in the same three factors (see section 4.5.3). The behavior of each participant was logged, allowing for an extensive analysis of the click-stream of each user.

The questionnaire items were then analyzed using a confirmatory structural equation modeling (SEM) approach with repeated ordinal dependent variables and a weighted least squares estimator, in which the subjective constructs were structurally related to each other, to the conditions (algorithm/preference input method), and to six behavioral measures extracted from the usage logs. The final model (Figure 9) had a reasonably good fit ($\chi^2(29) = 49.672$, $p = .0098$, CFI = .976, TLI = .982, RMSEA = .065).

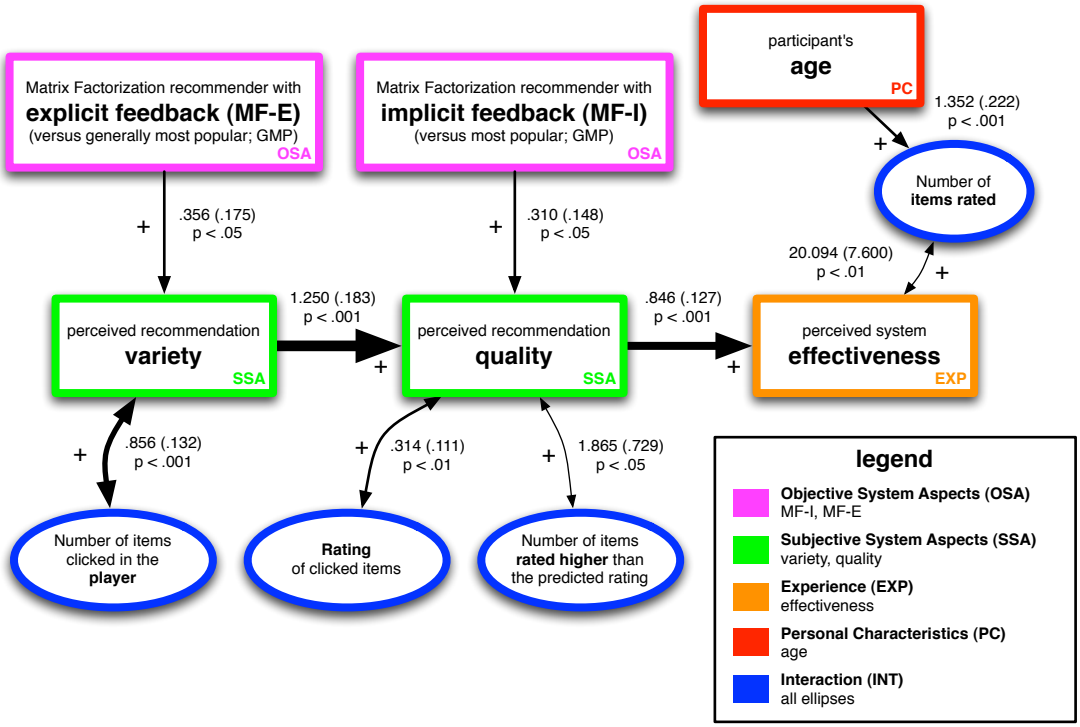

Figure 9: The path model constructed for FT4 - BBC trial. Please refer to section 4.2.4 "Graphical presentation of SEMs" for interpretation of this figure. Note that the "preference elicitation method" manipulation is represented by the two rectangles "explicit feedback (MF-E)" and "implicit feedback (MF-I)", which are tested against the non-personalized baseline "generally most popular (GMP)".

4.6.2 Results

As in FT3 (see Figure 8), we observe that a higher perceived recommendation variety leads to a higher perceived recommendation quality, which in turn leads to a higher perceived system effectiveness (SSA → SSA → system-EXP). Interestingly, the two variants of the matrix factorization algorithm affect different SSAs: compared to GMP, MF-E recommendations (which are based on explicit feedback) have a significantly higher perceived variety, while MF-I recommendations (which are based on implicit feedback) have a significantly higher perceived quality (OSA → SSA). Despite their different trajectories, the net effect of these algorithms is that their increased perceived quality positively affects the user experience (in terms of perceived system effectiveness) compared to the GMP condition (see also Figure 10).

In terms of behavioral measures, perceived variety is correlated with the number of items clicked in the player (as opposed to the catalog; INT → SSA). Recommendation quality is correlated with the rating users give to clicked items, and the number of items that are rated higher than the predicted rating (SSA → INT[15]). Finally, the

[15] SSA → INT is not predicted, but these behaviors are a direct expression of the SSA.

perceived system effectiveness is correlated with the total number of items that participants rate (EXP ⇆ INT). We furthermore note that older participants rate significantly more items (PC → INT).

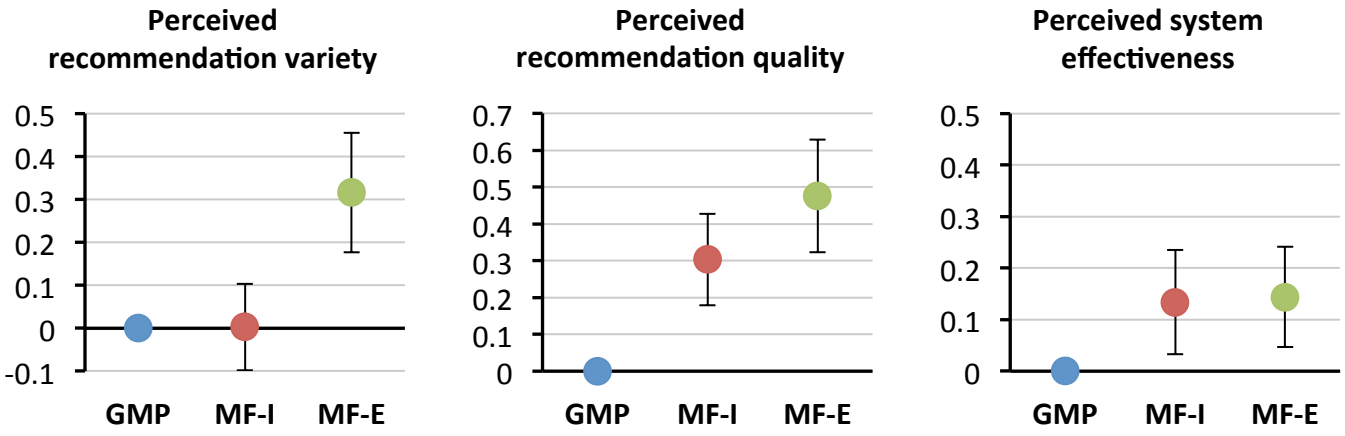

Figure 10: Tendency of marginal (direct and indirect) effects of the "preference elicitation method" condition on perceived recommendation variety, perceived recommendation quality, and system effectiveness. Error bars indicate +/-1 Standard Error. The value of GMP is fixed to zero; scales are in sample standard deviations.

4.6.3 Discussion of the results

This study confirms the basic structural relations between manipulation, subjective system aspects, and experience as also found in FT3 (OSA → SSA → EXP; requirement 4, see section 3.7). In this case, the OSA is the preference elicitation method (requirement 3). Two variants of the matrix factorization algorithm using different types of input (MF-E and MF-I) are tested against a non-personalized baseline (GMP). The fact that both MF-E and MF-I resulted in a better user experience than GMP is not very surprising, but interestingly the model shows that the cause for this effect is different for each algorithm. MF-E (which makes recommendations based on explicit feedback) seems to result in a higher variety in recommendations than GMP, which in turn leads to a higher recommendation quality and a higher perceived effectiveness (OSA → SSA → SSA → EXP); MF-I (which makes recommendations based on implicit feedback) on the other hand seems to have a direct influence on recommendation quality, which also leads to a higher perceived effectiveness (OSA → SSA → EXP).

The lack of an increased diversity for MF-I recommendations could be caused by a "pigeonholing effect": when the MF-I algorithm uses clicks to predict implicit preferences, and gives recommendations based on these preferences, most subsequent clicks will be in line with these predicted preferences, and the algorithm will increasingly home in on a very specific set of recommendations. In MF-E, negative explicit feedback (low ratings) can prevent such pigeonholing. Like Jones et

al. (2009), our results show that explicit control over the recommendations leads to a higher recommendation quality through an increase of perceived variety, but that *beyond* the effect through variety, there is no increase in recommendation quality (e.g. by means of a higher accuracy). The marginal effect on the perceived quality of the recommendations for MF-E and MF-I is about equal, and higher than the quality of the non-personalized recommendations (GMP). The same holds for perceived system effectiveness (see Figure 10).

An interesting way to exploit the difference between MF-I and MF-E would be to combine this result with the result of FT3, which found that diversity decreases over time (see Figure 8). In this respect, a recommender system could start out with the MF-I algorithm, which can provide high-quality recommendations from the start (as it does not need any ratings). As the diversity of the recommendations starts decreasing, the algorithm can put more weight on the users' ratings in determining their preferences, or even switch to the MF-E algorithm altogether. This would give the system a boost in recommendation variety, which should provide a better user experience in the long run.

Several behavioral correlates corroborate our subjective measures. The perceived variety of the recommendations is related to the number of items clicked in the player. These items are related to the item that is currently playing, and are therefore arguably more specialized (and thus over time more varied) than other recommendations. The results further show that users reward good recommendations with higher ratings, and system effectiveness is correlated with the number of ratings the user provides (see also requirement 6). The causal effect here is unclear: participants that rate more items may end up getting a better experience because the algorithms gain more knowledge about the user (INT → EXP). On the other hand, as found in FT1 (see Figure 3), users may intend to rate more items when they notice that this improves their experience (EXP → INT).

4.7 EX1 Choice overload experiment

The four studies described above are all field trials, testing real-life user-behavior in real-world systems. These studies have a high ecological validity, but it is hard to analyze decision processes in detail, because users are unrestricted in their needs, goals, and actions. To get more in-depth knowledge about users' decision processes, we conducted two controlled lab experiments. The tasks in these experiments are fixed, as is the interaction with the system: Users first rate a set of test-items and then make a single decision, each considering the same choice goal. The first of these experiments looked at the effect of the size and quality of the recommendation set on the user experience to study a phenomenon called choice overload. The

experiment is described in detail in (Bollen et al., 2010), so here we will only briefly discuss the results and the merit of using the framework in this experiment.

4.7.1 Setup

The experiment was conducted with 174 Dutch participants (invited from a panel with students or recently graduated students from several Dutch universities) with an average age of 26.8 (SD = 8.6). 89 of them were male. Participants were paid €3 upon successful completion of the experiment. They were asked to use the system (loaded with the 1M MovieLens dataset[16], and using the Matrix Factorization algorithm implementation from the MyMedia Software Framework) to find a good movie to watch. To train the recommender system, they first rated at least 10 movies they knew, and were subsequently presented with a list of recommendations from which to make a choice. Users were randomly assigned to receive either one of three recommendation sets: Top-5 (the best five recommendations, according to our MF algorithm), Top-20 (the best twenty recommendations), and Lin-20 (the best five recommendations, supplemented with the recommendations with rank 99, 199, 299, …, 1499). After making a choice, users completed a set of questionnaires to measure perceived recommendation set variety, recommendation set attractiveness, choice difficulty, satisfaction with the chosen item, and movie expertise. A Structural Equation Model was fitted on these questionnaire constructs and our two dichotomous manipulation variables ("Top-20 vs. Top-5" and "Lin-20 vs. Top-5"); Figure 11 shows the resulting path model; for details on the data analysis, refer to (Bollen et al., 2010).

[16] The MovieLens datasets are freely available at http://grouplens.org. The 1M MovieLens dataset contains one million ratings for 3952 movies and 6040 users.

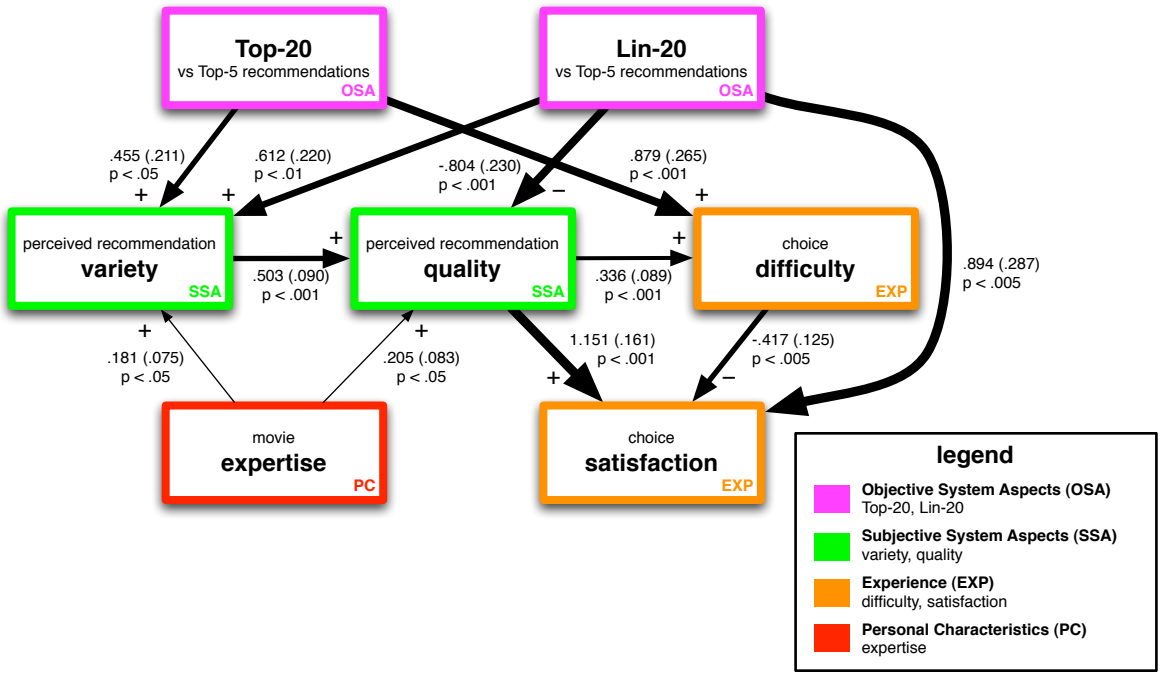

Figure 11: The path model constructed for EX1 - Choice overload experiment. Please refer to section 4.2.4 "Graphical presentation of SEMs" for interpretation of this figure. Note that the manipulation of recommendation set size and quality is represented by the two rectangles "Top-20" (a larger set than Top-5) and "Lin-20" (a larger set than Top-5, but with lower ranked additional items).

4.7.2 Results

The model shows that satisfaction with the chosen item (outcome-EXP) is the result of two opposing forces: a positive effect of the quality of the recommendations (SSA → outcome-EXP) and a negative effect of choice difficulty (process-EXP → outcome-EXP). Furthermore, high-quality recommendations increase choice difficulty (SSA → process-EXP), and recommendation quality itself depends strongly on the perceived recommendation variety (SSA → SSA; consistent with FT3 and FT4).

Relative to Top-5, the Top-20 condition increases the perceived variety of the recommendation set (OSA → SSA) and the difficulty of the choice (OSA → process-EXP): Top20 is more varied and more difficult to choose from. The Lin-20 set is also more varied than Top-5 (OSA → SSA), and negatively affects the recommendation set attractiveness (OSA → SSA). Lin-20 furthermore has a positive residual effect (i.e. controlling for recommendation set quality and choice difficulty) on satisfaction (OSA → outcome-EXP), relative to the Top-5 condition. Finally, in contrast to the findings of Kamis and Davern (2004) and Hu and Pu (2010), increased movie expertise in our study positively affects recommendation set attractiveness and perceived variety (PC → SSA).

4.7.3 Discussion of results

The marginal effect of our manipulation of the recommendation set composition on choice satisfaction is zero (requirement 2, see section 3.7); there is no significant marginal difference between the Top-5, Top-20 and Lin-20 sets (see Figure 12). Thanks to our mediating SSAs and EXPs (requirement 4 and 5), the model is able to show exactly why this is the case. The Top-20 set is more varied (SSA) than Top-5 (and thereby more attractive), but it is also more difficult (process-EXP) to make a choice from this set, and these effects level out to eventually show no difference in choice satisfaction (outcome-EXP). The Lin-20 set is less attractive than the Top-5 (SSA), but there is a positive residual effect on choice satisfaction (outcome-EXP). Arguably, this can be interpreted as a context effect: participants in this condition contrasted their choice against the inferior items that are in the tail of the Lin-20 distribution, making the choice more satisfying because it was easier to justify. In the end, however, these effects too cancel out, so the resulting net effect on choice satisfaction is approximately zero.

Finally, the results show that domain knowledge (PC) influences the perception of recommendation set variety and quality. Although requirement 8 specifies that PC should be related to EXP or INT, we have now at several occasions shown a relationship between PC/SC and SSA. Apparently, situational and personal characteristics can influence users' perceptions as well (e.g. "an expert's eye"), something not recognized in the earlier version of our framework (Knijnenburg et al., 2010).

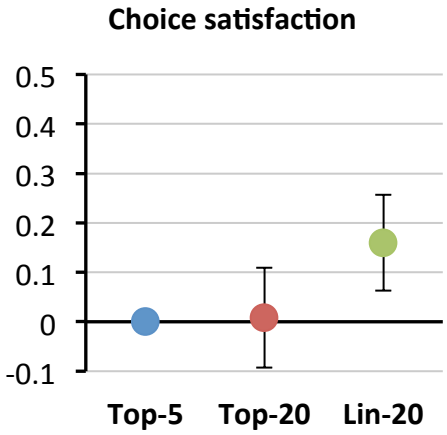

Figure 12: Tendency of marginal (direct and indirect) effects of the "recommendation set composition" condition on choice satisfaction. Error bars indicate +/-1 Standard Error. The value of Top-5 is fixed to zero; scales are in sample standard deviations.

4.8 EX2 Diversification experiment

As a final experiment, we set out to investigate the relative effect of algorithm accuracy and recommendation set diversity on user experience. Ziegler et al. (2005) already showed that diversifying the output of a recommender algorithm, although detrimental for the accuracy of the recommended list, results in an increase in overall satisfaction (OSA→EXP). However, their experiment used single-question measures and a restricted set of algorithms (only item-based and user-based k-Nearest Neighbors). Furthermore, as we have shown in our previous experiments, perceived variety and perceived quality of the recommendations (SSAs) are important mediators between objective aspects (OSA) and user experience (EXP). To better understand the intricate relation between accuracy, diversity and satisfaction, we therefore conducted an experiment in which we manipulated algorithm and recommendation set variety (OSA) and measured perceived accuracy and diversity (SSA) as well as several user experience variables (process-EXP, system-EXP and outcome-EXP). The experiment compares k-Nearest Neighbors (kNN) with a non-personalized "generally most popular" (GMP) algorithm (a baseline) and a state-of-the-art Matrix Factorization algorithm (MF). Like EX1, it uses the MovieLens 1M dataset.

4.8.1 Setup

The experiment was conducted online, using Amazon's Mechanical Turk service to gather 137 participants (78 male, mean age of 30.2, SD = 10.0), mainly from the US and India. Participants were paid US $4 upon successful completion of the experiment. They were asked to rate at least ten items from the MovieLens 1M set, after which they would receive a list of ten recommendations. The composition of the list was manipulated in 3x3 conditions, independently varying the algorithm (GMP, kNN, MF) and the diversification (none, little, lot) between subjects. Participants were asked to choose one movie. Finally, they filled out the user experience questionnaires.

Diversification was based on movie genre, and implemented using the greedy heuristic of Ziegler et al. (2005). Specifically, the heuristic starts with the item with the highest predicted rating, calculates a diversity score for each of the remaining items in the top 100 (based on how many movies in the current recommendation list were of the same genre), creates a compound score S weighing the predicted rating R and the diversity score D according to the formula $S = aD + (1-a)R$, and then chooses from the top 100 the item with the highest compound score. This process is repeated until the list contains ten items. For none, little and lot, the values of 'a' were 0, 0.5 and 0.9 respectively. Diversity scores were calculated using cosine similarity between the genres of the movies.

Participants answered 40 questionnaire items on a 7-point scale. The answers were factor analyzed and produced 6 conceptual factors:

- Perceived recommendation set diversity (5 items, e.g. "Several movies in the list of recommended movies were very different from each other", factor R^2 = .087)
- Perceived recommendation set accuracy[17] (6 items, e.g. "The recommended movies fitted my preferences", factor R^2 = .256)
- Perceived choice difficulty (4 items, e.g. "Selecting the best movie was easy/difficult", factor R^2 = .312)
- Perceived system effectiveness (7 items, e.g. "The recommender system gives me valuable recommendations", factor R^2 = .793)
- Choice satisfaction (6 items, e.g. "My chosen movie could become one of my favorites", factor R^2 = .647)
- Expertise (3 items, e.g. "Compared to my peers I watch a lot of movies", no incoming arrows)

8 items were deleted due to low communalities or excessive cross-loadings. The discovered constructs were then causally related with each other and with the 3x3 conditions in a Structural Equation Model. The resulting model (Figure 13) has a reasonably good fit ($\chi^2(68) = 132.19$, $p < .001$, CFI = .954, TLI = .976, RMSEA = .083). The interactions between diversification and algorithm were included in the model, but not in the graph. This means that the effect of diversification in the graph holds only for the "generally most popular" algorithm condition, and that the effects of kNN and MF hold only for the non-diversified condition.

[17] In the previous studies we used the concept "perceived recommendation quality" instead of "perceived recommendation accuracy". The concepts of "recommendation quality" and "recommendation accuracy" are slightly different: perceived accuracy merely looks at how well the recommendations fit ones preferences, while recommendation quality allows for other possible sources of quality.

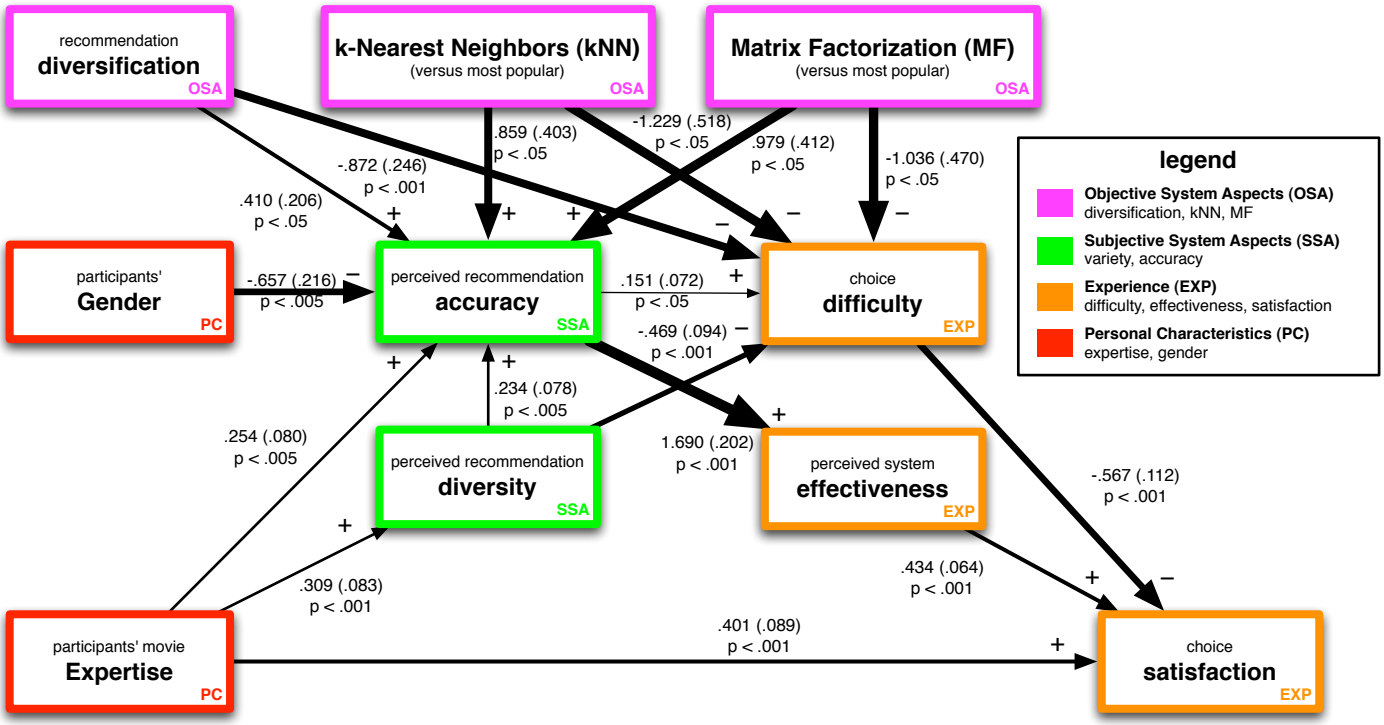

Figure 13: The path model constructed for EX2 - Diversification experiment. Please refer to section 4.2.4 "Graphical presentation of SEMs" for interpretation of this figure. The manipulation "diversification" gives values 0, 1, and 2 to levels "none", "little" and "lot" respectively. The manipulation of algorithm is represented by the two rectangles "k-Nearest Neighbors" and "Matrix Factorization", each tested against the "generally most popular" recommendations. Figure 14 gives a more insightful presentation of the effects of our conditions.

4.8.2 Results

The model shows that non-diversified recommendations provided by the kNN and MF algorithms are perceived as more accurate than the non-diversified "generally most popular" (GMP) recommendations (OSA → SSA). Diversified GMP recommendations are also perceived as more accurate than non-diversified GMP recommendations (OSA → SSA), even though their predicted rating has been traded off with diversity (i.e., the predicted rating of the diversified item set is lower). Accurate recommendations are generally more difficult to choose from (SSA → process-EXP), however, there is also a direct negative effect of kNN, MF and diversification on choice difficulty (OSA → process-EXP). Looking at the marginal effects in Figure 14 ("Choice difficulty") we observe that it is less difficult to choose from the recommendations produced by the high diversification settings and the kNN and MF algorithms (or in other words: it is mainly the non-diversified generally most popular recommendations that are difficult to choose from).
Surprisingly, the lack of a direct effect of diversification on perceived diversity suggests that diversified recommendations do *not* seem to be perceived as more diverse. Figure 14 ("Perceived recommendation diversity") gives a possible explanation for this lack of effect. The relation between manipulated diversity and

perceived diversity is not consistent between the three algorithms. Most notably, two observations differ strongly from our initial expectations: the kNN algorithm with low diversification provides surprisingly diverse recommendations, and the most diversified "generally most popular" recommendations are not perceived to be as diverse as they should be.

In general, more diverse recommendations (in this case *perceived* diversity) are also perceived as more accurate (SSA → SSA) and less difficult to choose from (SSA → process-EXP; see also Willemsen et al., 2011). Users who rate the recommendations as accurate also perceive the system to be more effective (SSA → system-EXP). The more effective the system and the easier the choice, the more satisfied participants are with their choice (process-EXP → outcome-EXP and system-EXP → outcome-EXP, respectively).

Finally, we observe that males find the recommendations generally less accurate (PC → SSA), and that in contrast to findings by Kamis and Davern (2004) and Hu and Pu (2010) expertise increases the perceived accuracy and diversity of the recommendations (PC → SSA) and also increases the choice satisfaction (PC → outcome-EXP).

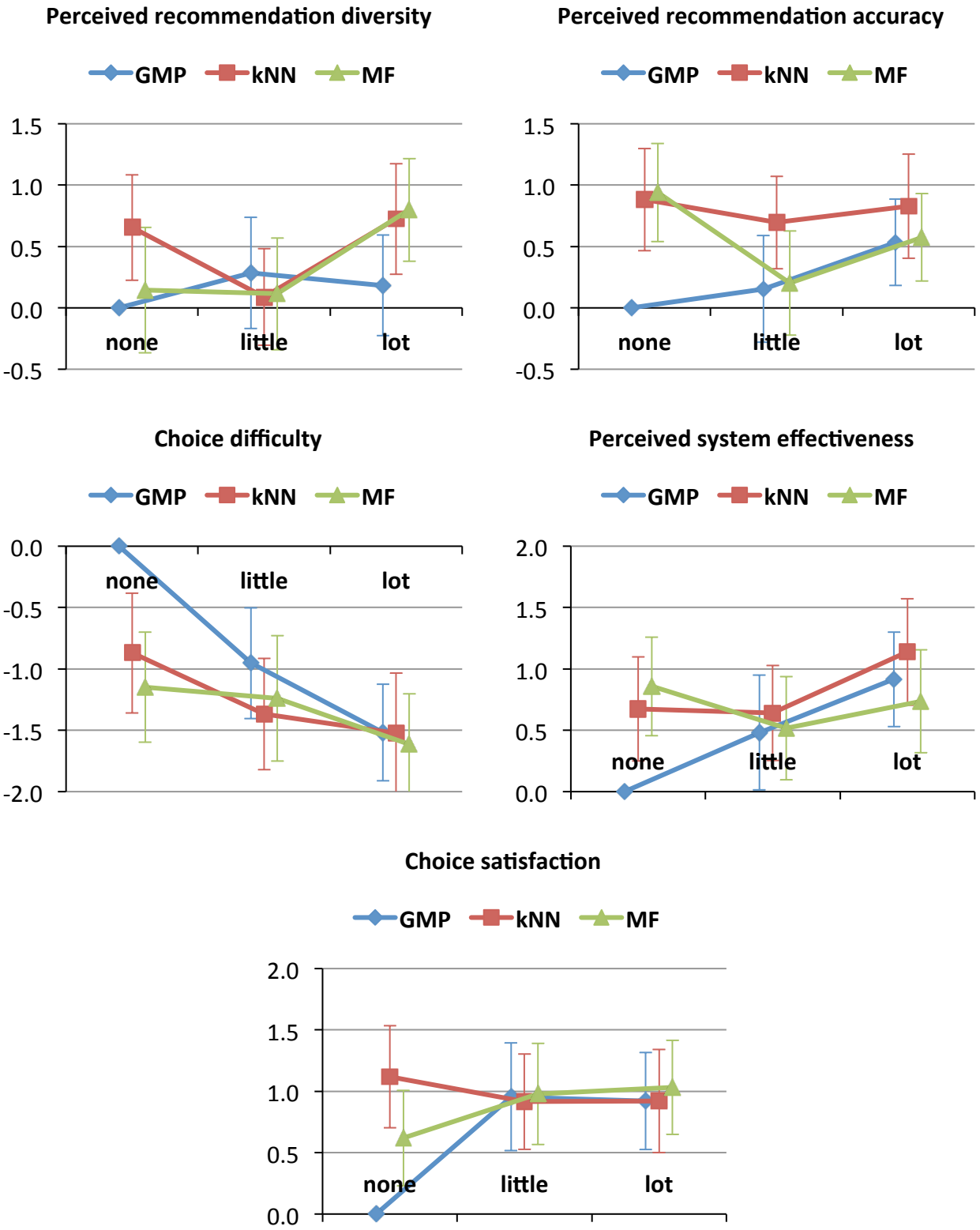

Figure 14: Tendency of marginal (direct and indirect) effects of algorithm and diversification on our subjective constructs. Error bars indicate +/-1 Standard Error. The value of GMP-none is fixed to zero; scales are in sample standard deviations.

4.8.3 Discussion of results

Like Ziegler et al. (2005), our results show a positive effect of diversification on the user experience (requirement 2, see section 3.7). The inclusion of several user experience constructs in our experiment enables us to show that this effect is mainly due to a lower choice difficulty and a higher perceived system effectiveness (requirement 5). Interestingly, this effect is partially mediated by perceived accuracy, but *not* by perceived diversity (requirement 4). Arguably, users do not necessarily see the diversified recommendations as more diverse. The higher perceived accuracy may have resulted from the fact that users evaluate the accuracy in our questionnaire not per item, but for the list of recommendations as a whole. Users may have noticed that more diverse lists more accurately reflect their diverse tastes.

Figure 14 shows that the "GMP-none" condition (i.e. the non-diversified condition without diversification) stands out; it is the only condition that results in a significantly worse user experience. Also, Figure 13 shows that the effects of diversification and algorithm are similar: they both increase recommendation accuracy and decrease choice difficulty. In other words: diversification of GMP recommendations might be just as effective in improving the user experience as introducing a recommendation algorithm. Moreover, despite the absence of an interaction effect in the model, Figure 14 suggests that these effects are not additive: introducing diversification *and* personalization does not improve the user experience beyond introducing either of them separately. Possibly, there is a ceiling-effect to the improvement of the perceived recommendation quality, at least in the between subjects design we employed in this study in which participants do not compare between conditions themselves directly. The result seems to suggest that providing a diversified but non-personalized list of recommendations may result in an equally good user experience as providing personalized recommendations, but we would suggest a replication with other datasets to confirm this surprising observation.

We are puzzled by the fact that our diversification manipulation did not result in a higher perceived diversity of the recommendations. Our diversification algorithm reduces the similarity between movies in terms of genre. This may not fit the definition of similarity as the users of the system judge it. Alternatively, our measurement of this concept could have been too weak to pick up on the subtle differences between the recommendation lists. Finally, the effects of expertise and gender show support for requirement 8: users' personal characteristics do indeed influence their user experience.

5 Validation of the framework

The goal of this paper is to describe a generic framework for the user-centric evaluation of recommender systems. Such a framework should have merits beyond the scope of a single research project. Despite their limited scopes, the field trials and experiments described in this paper can provide an initial test of the general applicability of the framework; any (lack of) consistency in our results gives valuable cues concerning the external validity of the framework. Furthermore, the framework was designed to allow for ad-hoc findings that can be elaborated in future work. Taken together, the expected and unexpected findings of our field trials and experiments allow us to reflect on the merit of our framework as a guiding tool for user-centric recommender systems research.

Below, we consolidate the findings of our field trials and experiments by validating parts of the framework separately and discussing the merit of the framework as a whole. We also assess whether parts of the framework need to be reconsidered in future investigations.

The findings are summarized in Table 2. The table shows that most requirements (as defined in section 3.7) are covered by at least two studies. As the requirements arise from gaps in the existing literature, the results of our studies bridge these gaps, and at the same time provide a thorough validation of the framework itself. Moreover, most of the specific results are confirmed in more than one study, thereby providing evidence for the robustness of the results as well as the general applicability of the framework.

Based on the current findings and the relation of the framework to a broad range of existing literature, we believe that the general requirements are extensible beyond the scope of the current studies. However, specific results may very well only hold for collaborative filtering media recommenders. Further research is needed to extend these specific findings to other types of recommender systems.

Requirement	Results	Studies
1. Algorithm	Turning recommendations on or off has a noticeable effect on user experience.	FT1, FT2, FT4, EX2
	Experience differences between algorithms are less pronounced.	EX2
2. Recommendation set composition	Recommendation set size, quality and diversity have a significant impact on the user experience.	EX1, EX2
	Sets of different sizes and quality may end up having the same choice satisfaction due to choice overload.	EX1
	Users do not perceive diversified recommendation sets as more diverse, but they do perceive them as are more accurate.	EX2
3. Preference input data	Explicit feedback leads to more diverse recommendations, which subsequently leads to increased perceived quality; implicit feedback increases the perceived recommendation quality directly.	FT4
4. Perceived aspects as mediators	Perceived aspects, particularly perceived recommendation quality and variety, provide a better understanding of the results.	all
5. User experience evaluation	Usage effort and choice difficulty are measured as process-related experience.	FT2, EX1, EX2
	Perceived system effectiveness is measured as system-related experience.	all except EX1
	Choice satisfaction is measured as outcome-related experience.	FT1, FT2, EX1, EX2
	Process-related experience causes system- or outcome-related experience.	FT2, EX1, EX2
	System-related experience causes outcome-related experience.	FT2, EX2
6. Providing feedback	Intention to provide feedback is a trade-off between trust/privacy and experience.	FT1, FT2, FT4
	Users' feedback behavior may not always be correlated with their intentions to provide feedback.	FT1, FT2
7. Behavioral data	A positive personalized user experience is characterized by reduced browsing behavior and increased consumption.	FT1, FT2
8. Personal and situational characteristics	Users' experience and behaviors change over time.	FT2, FT3
	Age, gender and domain knowledge have an influence on users' perceptions, experience and behaviors.	FT4, EX1, EX2

Table 2: A summary of the results of our field trials and experiments, listed per requirement (as defined in section 3.7)

5.1 Objective System Aspects (OSA) and Subjective System Aspects

We defined Objective System Aspects (OSAs) as the features of a recommender system that may influence the user experience, and as things to manipulate in our studies. Most work in the field of recommender systems is focused on algorithms. In our studies we therefore considered algorithms as a manipulation (requirement 1), but we also considered the type of preference input data (requirement 2) and the composition of the recommendation list (requirement 3). We reasoned that users would notice a change in an OSA: these subjective observations are the Subjective System Aspects (SSAs). Specifically, we argued that SSAs would mediate the effect of the OSAs on the user's experience (EXP; requirement 4).

Taken together, the manipulations in FT1, FT2, FT4 and EX2 show that users are able to perceive higher quality recommendations (OSA → SSA) and that this perception mediated the effect of the recommendation quality on the user experience (SSA → EXP), even though this mediation is in some cases only partial (i.e. in FT2, EX1, EX2).
EX1 and EX2 show similar results for the recommendation set composition. Size, quality and diversification of the recommendation set each influence the user experience via subjective perceptions (OSA → SSA → EXP), even though the effect on choice difficulty is mainly direct. Surprisingly, our diversification algorithm increased perceived accuracy but not perceived diversity. Despite this, diversification was as effective in improving the user experience as the introduction of a good algorithm.
In terms of preference input, FT4 shows that even though implicit feedback recommendations are based on more data than explicit feedback recommendations (all user behavior versus ratings only), they are not always better from a user's perspective, especially if one takes into account the variety of the recommendations.
Concluding, the core component of our framework – the link from algorithm or preference input data (OSA) to subjective recommendation quality (SSA) to experience (EXP) – is upheld throughout every conducted study. Some direct effects (OSA → EXP) occur, which could indicate that not all possible SSAs were measured. In general the SSAs mediate the effects of OSA on EXP, thereby explaining the effects of the OSAs in more detail (i.e. the why and how of improved user experience). The capability to explain both surprising and straightforward findings makes SSAs an essential part of our evaluation.

5.2 The different objects of experience (EXP)

We reasoned that user experience could be process-related, system-related and outcome-related. We argued that different recommender system aspects could influence different types of experience, and that one type of experience could influence another. Previous research typically includes just one type of user experience in their evaluation. In most of our studies we therefore considered more than one type of experience (requirement 5).
Our research confirms that these different types of user experience indeed exist; the studies discern the following user experience constructs: perceived usage effort (process-EXP; FT2), choice difficulty (process-EXP; EX1, EX2), perceived system effectiveness (system-EXP; all studies, except EX1) and choice satisfaction (outcome-EXP; FT1, FT2, EX1, EX2). In all cases, the main mediator that causes these experience variables is the perceived recommendation quality. Structurally,

process-related experiences often cause system-related experiences, which in turn cause outcome-related experiences (process-EXP → system-EXP → outcome-EXP).

5.3 Behavioral data and feedback intentions

We indicated that behavioral data could be triangulated with subjective experience data to improve the interpretation of behavioral results, and to ground self-report measures in actual behavior (requirement 7). We also reasoned that the specific behavior of providing preference feedback is of particular interest to recommender systems researchers, because in many systems this feedback is needed to provide good recommendations (requirement 6).
FT1, FT2 and FT4 show that feedback behavior is a trade-off between the users' trust or privacy concerns, and the user experience. The actual feedback behavior is not always highly correlated with the intention to provide feedback.
Furthermore, all field trials show several significant triangulations. Consistently, reduced browsing and increased consumption are indicators of effective systems. In some cases behaviors are directly related to the OSAs. This probably means that we did not measure the specific SSAs that would mediate the effect.

5.4 Personal and situational characteristics (PC and SC)

We argued that personal and situational characteristics may influence the users' experience and interaction with the system (requirement 8). Our research (FT1 and FT2) addresses trust in technology (PC) and system-specific privacy concerns (SC), and shows how these concepts influence users' feedback intentions (see requirement 6). Furthermore, the experience and interaction of the recommender systems in our research changes over time (SC; see FT2 and FT3). Concluding, several contextual effects seem to influence the users' interaction and experience with the recommender system. Our research addresses trust, privacy, time, age, gender and expertise (domain knowledge).
Our initial conception of the effect of personal and situational characteristics seems to have been too restrictive: We only allowed these characteristics to influence the experience (EXP) and interaction (INT). Based on the results of our research, we acknowledge that these characteristics can sometimes influence not only the evaluation, but also the perception of the system (i.e. influence the subjective system aspects, SSAs). We suggest including an additional arrow from personal and situational characteristics to subjective system aspects (PC → SSA and SC → SSA), and we encourage researchers to investigate this connection in more detail in future experiments.

Our research on the users' intention to provide feedback also alludes to possible changes in the framework. The results of FT1 and FT2 are not entirely consistent in terms of the factors that influence the intention to provide feedback, and further research is needed to specifically delineate what causes and inhibits users to provide preference feedback to the system (see Pommeranz et al., 2012, for a good example).

6 Conclusion

The framework provides clear guidance for the construction and analysis of new recommender system experiments. It allows for an in-depth analysis that goes beyond algorithmic performance: it can explain *why* users like a certain recommender system and *how* this user experience comes about.
The framework also puts emphasis on the integration of research. When testing with real users one cannot study algorithms in isolation; several system aspects (and personal and situational characteristics) have to be combined in a single experiment to gain a full understanding of the user experience.
For industry researchers, the user-centric focus of the framework provides a step closer to the customers, who may not consider the accuracy of the algorithm the most important aspect of their experience. Questionnaire-taking and A/B testing (the industry term for testing several versions of a certain system aspect) are an accepted form of research in web technology. For academic researchers, the framework provides an opportunity to check the real-world impact of the latest algorithmic improvements. Moreover, interesting effects of situational and personal characteristics, as well as behavioral correlates can be used as input for context-aware recommender engines. Even more so, when evaluating the relative merit of novel recommendation approaches such as context-aware algorithms (Adomavicius et al., 2005) and recommenders using social networks (Kautz et al., 1997), one has to rely on more sophisticated ways of measuring the full user experience, and our framework could serve as a guideline for such evaluations.

7 Future research

This paper has argued that measuring algorithmic accuracy is an insufficient method to analyze the user experience of recommender systems. We have therefore introduced and validated a user-centric evaluation framework that explains how and why the user experience of a recommender system comes about. With its mediating variables and its integrative approach, the framework provides a

structurally inclusive foundation for future work. Our research validates the framework and, beyond that, produced some unanticipated results.
Still, our work represents merely the tip of the iceberg of user-centric recommender systems research. Our research is limited in scope: we only tested a small number of media-oriented recommender systems. To determine the scope of applicability of our framework, further validation of the framework should consider other content types, specifically "search products". Moreover, some of our results are inconclusive and require further investigation.
The link between algorithmic accuracy and user experience is a fundamental question that currently still functions as the untested premise of a considerable part of the recommender systems research. The same holds for users' intention to provide feedback. And whereas our research shows some interesting results concerning the use of explicit versus implicit feedback, it is by no means exhaustive in this respect. We furthermore believe that future work could investigate the effects of other personal and situational characteristics, and the results of these studies could be used to personalize not only the recommendations of the system, but also the system itself (Knijnenburg & Willemsen, 2010; Knijnenburg & Willemsen, 2009, Knijnenburg et al., 2011).
Because of the pioneering nature of our work, we took a very thorough approach in our evaluation. The proposed methodology of measuring constructs with multiple questionnaire items and analyzing the results with exploratory factor analyses and structural equation models improves the external validity of our results. We realize, however, that this methodology may be infeasible when testing recommender systems in a fully operational industry setting. We therefore created a pragmatic procedure that allows the measurement of specific concepts of recommender system user experience with just a few key questionnaire items and a simplified (possibly automated) statistical evaluation (Knijnenburg et al., 2011a). Our current results provided useful input for the development of this procedure.
Concluding, our framework provides a platform for future work on recommender systems, and allows researchers and developers in industry and academia to consistently evaluate the user experience of their systems. The future of user-centric recommender systems research is full of exciting opportunities.

8 Acknowledgements

We would like to thank Mark Graus for programming the recommender systems used in EX1 and EX2, Steffen Rendle for implementing the explicit feedback MF algorithm, Niels Reijmer, Yunan Chen and Alfred Kobsa for their comments at several stages of this paper, and Dirk Bollen for allowing us to incorporate the

results of his choice overload experiment (EX1) in this paper. We also thank the three anonymous reviewers for their extensive comments on the initial submission. We gratefully acknowledge the funding of our work through the European Commission FP7 project MyMedia (www.mymediaproject.org) under the grant agreement no. 215006. For inquiries please contact info@mymediaproject.org.

9 Bibliography

Ackerman, M., Cranor, L., and Reagle, J.: 1999. "Privacy in e-commerce: examining user scenarios and privacy preferences". *Conference on Electronic commerce*, Denver, CO, pp. 1-8.

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 transactions on knowledge and data engineering* **17**, 734-749.

Adomavicius, G., Sankaranarayanan, R., Sen, S., and Tuzhilin, A.: 2005. "Incorporating contextual information in recommender systems using a multidimensional approach". *ACM Transactions on Information Systems* **23**, 103-145.

Anderson, J. C. and Gerbing, D. W.: 1988. "Structural equation modeling in practice: A review and recommended two-step approach". *Psychological bulletin* **103**, 411-423.

Bagozzi, R. and Yi, Y.: 1988. "On the evaluation of structural equation models". *Journal of the academy of marketing science* **16**, 74-94.

Baudisch, P. and Terveen, L.: 1999. "Interacting with recommender systems". *SIGCHI Conference on Human Factors in Computing Systems*, Pittsburgh, PA, p. 164.

Bechwati, N. and Xia, L.: 2003. "Do computers sweat? The impact of perceived effort of online decision aids on consumers' satisfaction with the decision process". *Journal of Consumer Psychology* **13**, 139-148.

Bentler, P. M. and Bonett, D. G.: 1980. "Significance tests and goodness of fit in the analysis of covariance structures". *Psychological bulletin* **88**, 588-606.

Berendt, B. and Teltzrow, M.: 2005. "Addressing users' privacy concerns for improving personalization quality: Towards an integration of user studies and algorithm evaluation". *IJCAI 2003 Workshop on Intelligent Techniques for Web Personalization*, Acapulco, Mexico, LNAI **3169**, 69-88.

Bharati, P. and Chaudhury, A.: 2004. "An empirical investigation of decision-making satisfaction in web-based decision support systems". *Decision Support Systems* **37**, 187-197.

Bhatnagar, A. and Ghose, S.: 2004. "Online information search termination patterns across product categories and consumer demographics". *Journal of Retailing* **80**, 221-228.

Bollen, D., Knijnenburg, B., Willemsen, M., and Graus, M.: 2010. "Understanding choice overload in recommender systems". *Fourth ACM conference on Recommender systems*, Barcelona, Spain, pp. 63-70.

Bradley, K. and Smyth, B.: 2001. "Improving recommendation diversity". *Twelfth Irish Conference on Artificial Intelligence and Cognitive Science*, Maynooth, Ireland, pp. 85-94.

Brodie, C., Karat, C., and Karat, J.: 2004. "Creating an E-commerce environment where consumers are willing to share personal information". In: C-M. Karat, J. O. Blom and J. Karat (eds.): *Designing Personalized User Experiences in eCommerce*. Dordrecht, The Netherlands: Kluwer, pp. 185-206.

Burke, R.: 2002. “Hybrid recommender systems: Survey and experiments”. *User Modeling and User-Adapted Interaction* **12**, 331-370.

Cena, F., Vernero, F., and Gena, C.: 2010. “Towards a Customization of Rating Scales in Adaptive Systems”. *18th International Conference on User Modeling, Adaptation, and Personalization*, Big Island, HI, LNCS **6075**, 369-374.

Chellappa, R. and Sin, R.: 2005. “Personalization versus privacy: An empirical examination of the online consumer's dilemma”. *Information Technology and Management* **6**, 181-202.

Chen, L. and Pu, P.: 2008. “A cross-cultural user evaluation of product recommender interfaces”. *2008 ACM conference on Recommender systems*, Lausanne, Switzerland, pp. 75-82.

Chen, L. and Pu, P.: 2009. “Interaction design guidelines on critiquing-based recommender systems”. *User Modeling and User-Adapted Interaction* **19**, 167-206.

Chen, L. and Pu, P.: 2012. “Critiquing-based Recommenders: Survey and Emerging Trends”. *User Modeling and User-Adapted Interaction* **22**.

Chin, D.: 2001. “Empirical evaluation of user models and user-adapted systems”. *User Modeling and User-Adapted Interaction* **11**, 181-194.

Cooke, A., Sujan, H., Sujan, M., and Weitz, B. A.: 2002. “Marketing the unfamiliar: the role of context and item-specific information in electronic agent recommendations”. *Journal of Marketing Research* **39**, 488-497.

Cosley, D., Lam, S., Albert, I., Konstan, J., and Riedl, J.: 2003. “Is seeing believing?: how recommender system interfaces affect users' opinions”. *SIGCHI conference on Human factors in computing systems*, Ft. Lauderdale, FL, pp. 585-592.

Cramer, H., Evers, V., van Someren, M., Ramlal, S., Rutledge, L., Stash, N., Aroyo, L., and Wielinga, B.: 2008. “The effects of transparency on trust and acceptance in interaction with a content-based art recommender”. *User Modeling and User-Adapted Interaction* **18**, 455-496.

Cramer, H., Evers, V., van Someren, M., Ramlal, S., Rutledge, L., Stash, N., Aroyo, L., and Wielinga, B.: 2008. “The effects of transparency on perceived and actual competence of a content-based recommender”. *CHI’08 Semantic Web User Interaction Workshop*, Florence, Italy.

Csikszentmihalyi, M.: 1975. “Beyond boredom and anxiety”. San Fransisco, CA: Jossey-Bass Publishers.

Davis, F.: 1989. “Perceived usefulness, perceived ease of use, and user acceptance of information technology”, *MIS Quarterly* **13**, 319–340.

Davis, F., Bagozzi, R., and Warshaw, P.: 1989. “User Acceptance of Computer Technology: A Comparison of Two Theoretical Models”. *Management Science* **35**, 982-1003.

Diehl, K., Kornish, L., and Lynch Jr, J.: 2003. “Smart agents: When lower search costs for quality information increase price sensitivity”. *Journal of Consumer Research* **30**, 56-71.

Felfernig, A., Teppan, E., and Gula, B.: 2007. “Knowledge-Based Recommender Technologies for Marketing and Sales”. *International Journal of Pattern Recognition and Artificial Intelligence* **21**, 333-354.

Felix, D., Niederberger, C., Steiger, P., and Stolze, M.: 2001. “Feature-oriented vs. Needs-oriented Product Access for Non-Expert Online Shoppers”. *IFIP Conference on Towards the E-Society: E-commerce, E-business, and E-government*, Zürich, Switzerland, pp. 399-406.

Fishbein, M. and Ajzen, I.: 1975. “Belief, attitude, intention and behavior: An introduction to theory and research”. Reading, MA: Addison-Wesley.

Gena, C., Brogi, R., Cena, F., and Vernero, F.: 2011. “Impact of Rating Scales on User’s Rating Behavior”. *19th International Conference on User Modeling, Adaptation, and Personalization*, Girona, Spain, LNCS **6787**, 123-134

Harper, F., Li, X., Chen, Y., and Konstan, J.: 2005. "An Economic Model of User Rating in an Online Recommender System". *10th International Conference on User Modeling*, Edinburgh, UK, LNCS **3538**, 307-316.

Hassenzahl, M.: 2005. "The thing and I: understanding the relationship between user and product". In: M.A. Blythe, A.F. Monk, K. Overbeeke, and P.C. Wright (eds.): *Funology*. Dordrecht, The Netherlands: Kluwer, pp. 31-42.

Hassenzahl, M.: 2008. "User experience (UX): towards an experiential perspective on product quality". *20th International Conference of the Association Francophone d'Interaction Homme-Machine*, Metz, France, pp. 11-15.

Häubl, G. and Trifts, V.: 2000. "Consumer decision making in online shopping environments: The effects of interactive decision aids". *Marketing Science* **19**, 4-21.

Häubl, G., Dellaert, B., Murray, K., and Trifts, V.: 2004. "Buyer behavior in personalized shopping environments". In: C-M. Karat, J. O. Blom and J. Karat (eds.): *Designing Personalized User Experiences in eCommerce*. Dordrecht, The Netherlands: Kluwer, pp. 207-229.

Hauser, J., Urban, G., Liberali, G., and Braun, M.: 2009. "Website morphing". *Marketing Science* **28**, 202-223.

Hayes, C., Massa, P., Avesani, P., and Cunningham, P.: 2002. "An on-line evaluation framework for recommender systems". *AH'2002 Workshop on Recommendation and Personalization in E-Commerce*, Málaga, Spain, pp. 50-59.

Herlocker, J., Konstan, J., and Riedl, J.: 2000. "Explaining collaborative filtering recommendations". *2000 ACM Conference on Computer Supported Cooperative Work*, Philadelphia, PA, pp. 241-250.

Herlocker, J., Konstan, J., Terveen, L., and Riedl, J.: 2004. "Evaluating collaborative filtering recommender systems". *ACM Transactions on Information Systems* **22**, 5-53.

Ho, S. Y. and Tam, K. Y.: 2005. "An Empirical Examination of the Effects of Web Personalization at Different Stages of Decision Making". *International Journal of Human-Computer Interaction* **19**, 95-112.

Hostler, R., Yoon, V., and Guimaraes, T.: 2005. "Assessing the impact of internet agent on end users' performance". *Decision Support Systems* **41**, 313-323.

Hsu, C. and Lu, H.: 2004. "Why do people play on-line games? An extended TAM with social influences and flow experience". *Information & Management* **41**, 853-868.

Hu, L.-T. and Bentler, P.: 1999. "Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives". *Structural Equation Modeling: A Multidisciplinary Journal* **6**, 1-55.

Hu, R. and Pu, P.: 2009. "A comparative user study on rating vs. personality quiz based preference elicitation methods". *14th International Conference on Intelligent User Interfaces*, Sanibel Island, FL, pp. 367-371.

Hu, R. and Pu, P.: 2010. "A Study on User Perception of Personality-Based Recommender Systems". *18th International Conference on User Modeling, Adaptation, and Personalization*, Big Island, HI, LNCS **6075**, 291-302.

Hu, R. and Pu., P.: 2011. "Enhancing Recommendation Diversity with Organization Interfaces". *16th international conference on Intelligent user interfaces*, Palo Alto, CA, pp. 347-350.

Hu, Y., Koren, Y., and Volinsky, C.: 2008. "Collaborative Filtering for Implicit Feedback Datasets". *2008 Eighth IEEE International Conference on Data Mining*, Pisa, Italy, pp. 263-272.

Huang, P., Lurie, N. H., and Mitra, S.: 2009. "Searching for Experience on the Web: An Empirical Examination of Consumer Behavior for Search and Experience Goods". *Journal of Marketing* 73(2):55-69.

Iyengar, S. and Lepper, M.: 2000. "When choice is demotivating: Can one desire too much of a good thing?". *Journal of personality and social psychology* **79**, 995-1006.

Jones, N., Pu, P., and Chen, L.: 2009. "How Users Perceive and Appraise Personalized Recommendations". *17th International Conference on User Modeling, Adaptation, and Personalization Conference*, Trento, Italy, LNCS **5535**, 461-466.

Kamis, A. and Davern, M. J.: 2004. "Personalizing to product category knowledge: exploring the mediating effect of shopping tools on decision confidence". *37th Annual Hawaii International Conference on System Sciences*, Big Island, HI.

Kaplan, B. and Duchon, D.: 1988. "Combining Qualitative and Quantitative Methods in Information Systems Research: A Case Study". *Mis Quarterly* **12**, 571-586.

Kautz, H., Selman, B., and Shah, M.: 1997. "Referral Web: combining social networks and collaborative filtering". *Communications of the ACM* **40**, 63-65.

Lynch, J. G., Jr. and Ariely, D.: 2000. "Wine Online: Search Cost and Competition on Price, Quality, and Distribution," *Marketing Science*, **19**, 83-103.

Knijnenburg, B. P., Reijmer, N. J. M., and Willemsen, M. C.: 2011. "Each to His Own: How Different Users Call for Different Interaction Methods in Recommender Systems". *5th ACM Conference on Recommender Systems*, Chicago, IL.

Knijnenburg, B. P. and Willemsen, M. C.: 2009. "Understanding the effect of adaptive preference elicitation methods on user satisfaction of a recommender system". *Third ACM conference on Recommender systems*, New York, NY, pp. 381-384.

Knijnenburg, B. P. and Willemsen, M. C.: 2010. "The effect of preference elicitation methods on the user experience of a recommender system". *28th International Conference on Human Factors in Computing Systems*, Atlanta, GA, pp. 3457-3462.

Knijnenburg, B. P., Willemsen, M. C., and Kobsa, A.: 2011a. "A Pragmatic Approach to Support the User-Centric Evaluation of Recommender Systems". *5th ACM Conference on Recommender Systems*, Chicago, IL.

Knijnenburg, B. P., Meesters, L., Marrow, P., and Bouwhuis, D.: 2010. "User-centric evaluation framework for multimedia recommender systems". *First International Conference on User Centric Media*, Venice, Spain, LNICST **40**, 366-369.

Knijnenburg, B. P., Willemsen, M. C., and Hirtbach, S.: 2010a. "Receiving recommendations and providing feedback: The user-experience of a recommender system". *11th International Conference on Electronic Commerce and Web Technologies*, Bilbao, Spain, LNBIP **61**, 207-216.

Kobsa, A. and Teltzrow, M.: 2005. "Contextualized communication of privacy practices and personalization benefits: Impacts on users' data sharing and purchase behavior". *Workshop on Privacy Enhancing Technologies*, Toronto, Canada, LNCS **3424**, 329-343.

Komiak, S. Y. X. and Benbasat, I.: 2006. "The effects of personalization and familiarity on trust and adoption of recommendation agents". *Mis Quarterly* **30**, 941-960.

Konstan, J. A. and Riedl, J.: 2012. "Recommender Systems: From Algorithms to User Experience". *User Modeling and User-Adapted Interaction* **22**.

Koren, Y., Bell, R., and Volinsky, C.: 2009. "Matrix Factorization Techniques for Recommender Systems". *IEEE Computer* **42**, 30-37.

Koren, Y.: 2010. "Factor in the neighbors: Scalable and accurate collaborative filtering". *Transactions on Knowledge Discovery from Data* **4**, 1-24.

Koufaris, M.: 2003. "Applying the technology acceptance model and flow theory to online consumer behavior". *Information systems research* **13**, 205-223.

Kramer, T.: 2007. "The effect of measurement task transparency on preference construction and evaluations of personalized recommendations". *Journal of Marketing Research* **44**, 224-233.

Krishnan, V., Narayanashetty, P., Nathan, M., Davies, R., and Konstan, J.: 2008. "Who predicts better?: results from an online study comparing humans and an online recommender system". *2008 ACM conference on Recommender systems*, Lausanne, Switzerland, pp. 211-218.

Law, E., Roto, V., Hassenzahl, M., Vermeeren, A., and Kort, J.: 2009. "Understanding, scoping and defining user experience: a survey approach". *27th International Conference on Human Factors in Computing Systems*, Boston, MA, pp. 719-728.

Marrow, P., Hanbidge, R., Rendle, S., Wartena, C., and Freudenthaler, C.: 2009. "MyMedia: Producing an Extensible Framework for Recommendation". *Networked Electronic Media Summit 2009*, Saint-Malo, France.

McNamara, N. and Kirakowski, J.: 2006. "Functionality, usability, and user experience: Three areas of concern". *ACM Interactions* **13**, 26-28.

McNee, S., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S., Rashid, A., Konstan, J., and Riedl, J.: 2002. "On the recommending of citations for research papers". *2002 ACM Conference on Computer Supported Cooperative Work*, New Orleans, LA, pp. 116-125.

McNee, S., Riedl, J., and Konstan, J.: 2006. "Being accurate is not enough: how accuracy metrics have hurt recommender systems". *24th International Conference Human factors in computing systems*, Montréal, Canada, pp. 1097-1101.

McNee, S., Riedl, J., and Konstan, J.: 2006. "Making recommendations better: an analytic model for human-recommender interaction". *24th International Conference Human factors in computing systems*, Montréal, Canada, pp. 1103-1108.

Meesters, L., Marrow, P., Knijnenburg, B. P., Bouwhuis, D., and Glancy, M.: 2008. "MyMedia Deliverable 1.5 End-user recommendation evaluation metrics.". http://www.mymediaproject.org/Publications/WP1/MyMedia_D1.5.pdf

Murray, K. and Häubl, G.: 2008. "Interactive consumer decision aids". In: B. Wierenga (ed.): *Handbook of Marketing Decision Models,* Heidelberg: Springer-Verlag, pp. 55-77.

Murray, K. and Häubl, G.: 2009. "Personalization without interrogation: Towards more effective interactions between consumers and feature-based recommendation agents". *Journal of Interactive Marketing* **23**, 138-146.

Muthen, B.: 1984. "A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators". *Psychometrika* **49**, 115-132.

Nelson, P.: 1970, "Information and Consumer Behavior," *Journal of Political Economy* **78**, 311-29.

Netemeyer, R. and Bentler, P.: 2001. "Structural Equations Modeling and Statements regarding Causality". *Journal of Consumer Psychology* **10**, 83-85.

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

Ozok, A. A., Fan, Q., and 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* **29**, 57-83.

Paramythis, A., Weibelzahl, S., and Masthoff, J.: 2010. "Layered Evaluation of Interactive Adaptive Systems: Framework and Formative Methods". *User Modeling and User-Adapted Interaction* **20**, 383-453.

Pathak, B., Garfinkel, R., Gopal, R. D., Venkatesan, R., and Yin, F.: 2010. "Empirical Analysis of the Impact of Recommender Systems on Sales". *Journal of Management Information Systems* **27**, 159-188.

Pedersen, P.: 2000. "Behavioral effects of using software agents for product and merchant brokering: an experimental study of consumer decision-making". *International Journal of Electronic Commerce* **5**, 125-141.

Pommeranz, A., Broekens, J., Wiggers, P., Brinkman, W.-P., and Jonker, C. M.: 2012. "Designing Interfaces for Explicit Preference Elicitation: A User-Centered Investigation of Preference Representation and Elicitation Process". *User Modeling and User-Adapted Interaction* **22**.

Preece, J., Rogers, Y., and Sharp, H.: 2002. "Interaction design: beyond human-computer interaction". New York: Wiley.

Pu, P. and Chen, L.: 2006. "Trust building with explanation interfaces". *11th International Conference on Intelligent User Interfaces*, Sydney, Australia, pp. 93-100.

Pu, P. and Chen, L.: 2007. "Trust-inspiring explanation interfaces for recommender systems". *Knowledge-Based Systems* **20**, 542-556.

Pu, P. and Chen, L.: 2010. "A User-Centric Evaluation Framework of Recommender Systems". *ACM RecSys 2010 Workshop on User-Centric Evaluation of Recommender Systems and Their Interfaces*, Barcelona, Spain, pp. 14-21.

Pu, P., Chen, L., and Hu, R.: 2012. "Evaluating Recommender Systems from the User's Perspective: Survey of the State of the Art". *User Modeling and User-Adapted Interaction* **22**.

Pu, P. and Kumar, P.: 2004. "Evaluating example-based search tools". *5th ACM conference on Electronic commerce*, New York, NY, pp. 208-217.

Pu, P., Chen, L., and Kumar, P.: 2008. "Evaluating product search and recommender systems for E-commerce environments". *Electronic Commerce Research* **8**, 1-27.

Rendle, S. and Schmidt-Thieme, L.: 2008. "Online-updating regularized kernel matrix factorization models for large-scale recommender systems". *2008 ACM conference on Recommender systems*, Lausanne, Switzerland, pp. 251-258.

Rendle, S., Freudenthaler, C., Gantner, Z., and Schmidt-Thieme, L.: 2009. "BPR: Bayesian personalized ranking from implicit feedback". *Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, Montreal, Canada, pp. 452-461.

Resnick, P. and Varian, H.: 1997. "Recommender systems". *Communications of the ACM* **40**, 56-58.

Scheibehenne, B., Greifeneder, R., and Todd, P.: 2010. "Can there ever be too many options? A meta-analytic review of choice overload". *Journal of Consumer Research* **37**, 409–25.

Schwartz, B.: 2004. "The paradox of choice: Why more is less". New York, NY: HarperCollins Publishers.

Senecal, S. and Nantel, J.: 2004. "The influence of online product recommendations on consumers' online choices". *Journal of Retailing* **80**, 159-169.

Sheeran, P.: 2002. "Intention-Behavior Relations: A Conceptual and Empirical Review". *European Review of Social Psychology* **12**, 1-36.

Simonson, I. and Tversky, A.: 1992. "Choice in Context: Tradeoff Contrast and Extremeness Aversion". *Journal of Marketing Research* **29**, 281-295.

Spiekermann, S., Grossklags, J., and Berendt, B.: 2001. "E-privacy in 2nd generation E-commerce: privacy preferences versus actual behavior". *3rd ACM conference on Electronic Commerce*, Tampa, FL, pp. 38-47.

Stolze, M. and Nart, F.: 2004. "Well-integrated needs-oriented recommender components regarded as helpful". *22nd International Conference on Human Factors in Computing Systems*, Vienna, Austria, p. 1571.

Tam, K. Y. and Ho, S. Y.: 2005. "Web Personalization as a Persuasion Strategy: An Elaboration Likelihood Model Perspective". *Information Systems Research* **16**, 271-291.

Teltzrow, M. and Kobsa, A.: 2004. "Impacts of user privacy preferences on personalized systems". *Human-Computer Interaction Series* **5**:315-332.

Tintarev, N. and Masthoff, J.: 2012. "Evaluating the Effectiveness of Explanations for Recommender Systems". *User Modeling and User-Adapted Interaction* **22**.

Torres, R., McNee, S., Abel, M., Konstan, J., and Riedl, J.: 2004. "Enhancing digital libraries with TechLens+". *4th ACM/IEEE-CS joint conference on Digital libraries*, Tucson, AZ, pp. 228-236.

Van Velsen, L., Van Der Geest, T., Klaassen, R., and Steehouder, M.: 2008. "User-centered evaluation of adaptive and adaptable systems: a literature review". *Knowledge Engineering Review* **23**, 261-281.

Venkatesh, V., Morris, M., Davis, G., and Davis, F.: 2003. "User Acceptance of Information Technology: Toward a Unified View". *Mis Quarterly* **27**, 425-478.

Viappiani, P., Faltings, B., and Pu, P.: 2006. "Preference-based search using example-critiquing with suggestions". *Journal of Artificial Intelligence Research* **27**, 465-503.

Viappiani, P., Pu, P., and Faltings, B.: 2008. "Preference-based search with adaptive recommendations". *AI Communications* **21**, 155-175.

Vijayasarathy, L. R. and Jones, J. M.: 2001. "Do Internet Shopping Aids Make a Difference? An Empirical Investigation". *Electronic Markets* **11**, 75-83.

Wang, W. and Benbasat, I.: 2007. "Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs". *Journal of Management Information Systems* **23**, 217-246.

Willemsen, M. C., Knijnenburg, B. P., Graus, M. P., Velter-Bremmers, L. C. M., and Fu, K.: 2011. "Using Latent Features Diversification to Reduce Choice Difficulty in Recommendation Lists". *RecSys'11 Workshop on Human Decision Making in Recommender Systems*, Chicago, IL.

Xiao, B. and Benbasat, I.: 2007. "ECommerce Product Recommendation Agents: Use, Characteristics, and Impact". *Mis Quarterly* **31**, 137-209.

Yu, J., Ha, I., Choi, M., and Rho, J.: 2005. "Extending the TAM for a t-commerce". *Information & Management* **42**, 965-976.

Ziegler, C., McNee, S., Konstan, J., and Lausen, G.: 2005. "Improving recommendation lists through topic diversification". *14th international World Wide Web Conference*, Chiba, Japan, pp. 22-32.

Zins, A. and Bauernfeind, U.: 2005. "Explaining online purchase planning experiences with recommender websites". *International Conference on Information and Communication Technologies in Tourism*, Innsbruck, Austria, pp. 137-148.

Appendix A: Methodology

Statistical analysis of our field trials and experiments involved two main steps: validating the measured latent concepts using exploratory factor analysis (EFA) and testing the structural relations between the manipulations, latent concepts and behavioral measurements using structural equation modeling (SEM). In this section we provide a detailed description of our methodological approach by means of a

hypothetical example. See Knijnenburg et al. (2011a) for a more pragmatic procedure to evaluate recommender systems.
In the example, we test two algorithms (A_1 and A_2) against a non-personalized baseline (A_0). We measure perceived recommendation quality (Q) with 5 statements (Q_1…Q_5) to which participants can agree or disagree on a 5-point scale. Satisfaction with the system (S) is measured with 6 items (S_1…S_6). Finally, we measure user behavior (B) in terms of the number of clips watched from beginning to end.

Exploratory Factor Analysis (EFA)

The first step is to confirm whether the 13 items indeed measure the predicted two latent concepts. This is done using exploratory factor analysis. This technique extracts common variance between the measured items and distributes this variance over a number of latent factors. We use the software package Mplus[18] to do this analysis with the command "analysis: type = efa 1 3; estimator = wlsmv;". This runs the exploratory factor analysis with 1, 2 and 3 factors, and uses a weighted least squares estimation procedure with mean- and variance-adjusted chi-square tests.
The factor analytical model can be represented as Figure 15. Each item is essentially a regression outcome[19], predicted by two unobserved latent variables. $I_{1,1}$ to $I_{2,11}$ are called the loadings of the items Q_1…S_6 on the factors F_1 and F_2. The model tries to estimate these loadings so that the paths match the covariance matrix of the items Q_1 to S_6 as closely as possible (e.g. $cov_{1,2} \approx I_{1,1}*I_{1,2} + I_{2,1}*I_{2,2} + I_{1,1}*w_{1,2}*I_{2,2} + I_{1,2}*w_{1,2}*I_{2,1}$). Intuitively, factor analysis tries to model the "overlap" between items. The part of the variance that does not overlap is excluded (and represented by the arrows at the bottom of Figure 15). The more overlap extracted, the more reliable the factor (the arrows at the top). The factors may be correlated with each other ($w_{1,2}$). The solution has no standard coordinate system, so it is rotated so that each question loads on only one factor as much as possible.
If the measurement tool was specified correctly, then the model has a good least-squares fit, and only the paths from F_1 to Q_1…Q_5 and from F_2 to S_1…S_6 are significantly larger than zero (i.e. only the darker paths in Figure 15). In that case F_1 measures recommendation quality, and F_2 measures satisfaction with the system. However, the following problems may arise:

- A one-factor solution fits almost as well as a two-factor solution: In this case, we must conclude that Q and S are essentially the same concept.

[18] http://www.statmodel.com/

[19] Since these outcomes are measured on a 5-point scale, this regression model is an ordinal response model.

- A three-factor solution fits much better than a two-factor solution: In this case, this questionnaire measures three concepts (usually because either S or Q is in fact a combination of two factors).
- A certain item does not have enough in common with the other items to load on any of the factors significantly ($p < .05$): In this case the item has "low communality" and should be removed from the analysis.
- A certain item loads on the wrong factor, or on both factors: The item has a high "cross-loading", and unless we can come up with a good reason for this to occur, it should be removed from the analysis.

Once an adequate factor solution is established, the remaining items are used in the second step of the analysis.

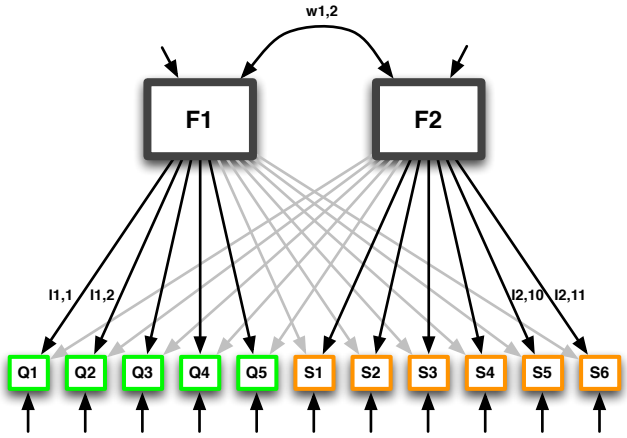

Figure 15: Representation of the exploratory factor analysis example. Two factors, F_1 and F_2, are extracted from the questionnaire data (items $Q_1 \ldots Q_5$ and $S_1 \ldots S_6$). The arrows from the factors to the questions ($I_{1,1} \ldots I_{2,11}$) represent the factor loadings. The arrows at the bottom represent the portion of the variance not accounted for by the analysis. Factors have a certain reliability (represented by the arrows at the top), and may be correlated (represented by $w_{1,2}$). If the grey arrows are close to zero, F_1 effectively measures recommendation quality (Q) and F_2 effectively measures satisfaction with the system.

Structural Equation Modeling (SEM)

The second step is to test the structural relations between our manipulation (A), the latent concepts (Q and S) and the behavior measurement (B). Our manipulation has three conditions (A_0, A_1, and A_2), so we create two dummy variables (A_1 and A_2), which are tested against the baseline. The dummies are coded in such a way that they represent the different conditions: For participants in the baseline $A_1 = 0$, and $A_2 = 0$; for participants with algorithm 1, $A_1 = 1$, and $A_2 = 0$; for participants with algorithm 1, $A_1 = 0$, and $A_2 = 1$. As a result of this coding, the effect of A_1 is therefore the effect of algorithm 1 compared to the baseline, and the effect of A_2 is the effect of algorithm 2 compared to the baseline.

The resulting model is tested using structural equation modeling. The specification of the model defines the factors and the items with which they are measured (in

Mplus: "Q by Q1-Q5; S by S1-S6;"), and the structural relations among the factors and other variables (in Mplus: "S on Q A1 A2; Q on A1 A2;"). This creates a model that can be represented as Figure 16 (but, as of yet, without B).

Like any regression model, structural equation models make assumptions about the direction of causality in the model. From a modeling perspective, an effect (Q→S) and its reverse (S→Q) are equally plausible. By including the manipulation(s) in our model, we are able to "ground" the causal effects: participants are randomly assigned to a condition, so condition assignment cannot be caused by anything in the model (i.e. A_1→Q is possible, but not Q→A_1). Furthermore, the framework provides hypotheses for the directionality of causal effects (since in the framework we hypothesize that SSA→EXP and not EXP→SSA, we can limit ourselves to testing Q→S).

For each regression path in the model, a regression coefficient is estimated. As the values of the latent constructs are standardized (by means of the factor analysis), the regression coefficient between the two latent constructs (Q→S) shows that a 1 standard deviation difference in Q causes a 0.60 standard deviation difference S. The standard error of the coefficient (0.15) can be used in a z-test to test whether the path is significant ($p = z[.60/.15] = .00003 < .001$). The dummy variables A_1 and A_2 are not standardized, but represent the presence (1) or absence (0) of a certain condition. Therefore, the coefficient on the arrow A_1→Q shows that Q is 0.50 standard deviations higher for participants in A_1 than for those in A_0.

In a typical model, not all initially specified paths are significant. This means that some effects will be fully mediated. For example: if the paths from A_1 and A_2 to S (the lighter paths in Figure 16) are not significant, the effect of A_1 and A_2 on S is fully mediated by Q (i.e. A→Q→S). Otherwise, there is only partial mediation. Non-significant effects are removed from the model and the model is ran again to create a more parsimonious result[20].

Variables that have no hypothesized effect (such as B in this model) are initially included in the model without specifying any structural relation for it. Mplus can provide a "modification index" for this variable, thereby showing where it best fits in the model. Since such effect is ad-hoc, it is only to be included if it is highly significant.

[20] Not all non-significant effects are excluded from the models. For instance, when two conditions are tested against a baseline, and one of the conditions does not significantly differ from the baseline but the other does, then the non-significant effect is retained to allow a valid interpretation of the significant effect. Furthermore, in EX2, the interactions between diversity and algorithm, though not significant, are retained, as they allow a valid interpretation of the significant conditional main effects.

Variables for which we have no hypothesized direction of effect (again, variable B in this model, for which we don't know whether B→S or S→B) are included as a correlation. This means that no assumption about the direction of causality is made. The final model (after excluding non-significant effects and including ad-hoc effects) can be tested as a whole. The Chi-square statistic tests the difference in explained variance between the proposed model and a fully specified model. A good model is not statistically different from the fully specified model ($p > .05$). However, this statistic is commonly regarded as too sensitive, and researchers have therefore proposed other fit indices (Bentler & Bonett, 1980). Based on extensive simulations, Hu and Bentler (1999) propose cut-off values for these fit indices to be: CFI > .96, TLI > .95, and RMSEA < .05. Moreover, a good model makes sense from a theoretical perspective. The model shown in Figure 16 has this quality: The algorithms (A_1 and A_2) influence the perceived recommendation quality (Q), which in turn influences the satisfaction with the system (S). The satisfaction (S) is in turn correlated with the number of clips watched from beginning to end (B).

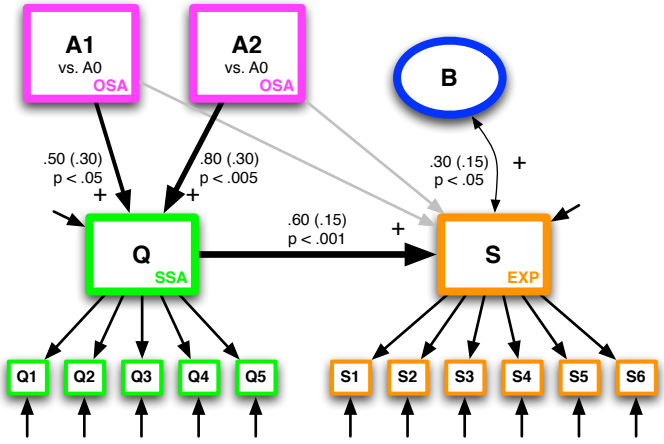

Figure 16: Representation of the structural equation modeling example. The algorithms (A_1 and A_2) influence the perceived recommendation quality (Q), which in turn influences the satisfaction with the system (S). The satisfaction (S) is in turn correlated with the number of clips watched from beginning to end (B). Q is measured by (Q_1...Q_5) and S is measured by (S_1...S_6). In the models in the main text the questionnaire items (Q_1...S_6) are hidden in order to get a less cluttered representation.

Appendix B: Questionnaire items for each construct

This appendix lists all the questionnaire items shown to the participants of the field trials and experiments. It makes a distinction between those items that were included in the final analysis, and those items that did not contribute to stable constructs and were therefore deleted. Included questions are in order of decreasing factor loading.

FT1 EMIC pre-trial

Perceived recommendation quality

Included

- I liked the items recommended by the system.
- The recommended items fitted my preference.
- The recommended items were well-chosen.
- The recommended items were relevant.
- The system recommended too many bad items.
- I didn't like any of the recommended items.
- The items I selected were “the best among the worst”.

Perceived system effectiveness

Included:

- I would recommend the system to others.
- The system is useless.
- The system makes me more aware of my choice options.
- I make better choices with the system.
- I can find better items without the help of the system.
- I can find better items using the recommender system.

Not included:

- The system showed useful items.

Choice satisfaction

Included:

- I like the items I’ve chosen.
- I was excited about my chosen items.
- I enjoyed watching my chosen items.
- The items I watched were a waste of my time.
- The chosen items fit my preference.
- I know several items that are better than the ones I selected.
- Some of my chosen items could become part of my favorites.
- I would recommend some of the chosen items to others/friends.

Intention to provide feedback

Included:

- I like to give feedback on the items I’m watching.
- Normally I wouldn’t rate any items.
- I only sparingly give feedback.
- I didn’t mind rating items.
- In total, rating items is not beneficial for me.

General trust in technology

Included:

- Technology never works.
- I'm less confident when I use technology.
- The usefulness of technology is highly overrated.
- Technology may cause harm to people.

Not included:

- I prefer to do things by hand.
- I have no problem trusting my life to technology.
- I always double-check computer results.

System-specific privacy concern

Included:

- I'm afraid the system discloses private information about me.
- The system invades my privacy.
- I feel confident that the system respects my privacy.
- I'm uncomfortable providing private data to the system.
- I think the system respects the confidentiality of my data.

FT2 EMIC trial

Perceived recommendation quality

Included:

- I liked the items shown by the system.
- The shown items fitted my preference.
- The shown items were well-chosen.
- The shown items were relevant.
- The system showed too many bad items.
- I didn't like any of the shown items.

Not included:

- The system showed useful items.
- The items I selected were "the best among the worst".

Effort to use the system

Included:

- The system is convenient.
- I have to invest a lot of effort in the system.
- It takes many mouse-clicks to use the system.

Not included:

- Using the system takes little time.
- It takes too much time before the system provides adequate recommendations.

Perceived system effectiveness and fun

Included:

- I have fun when I'm using the system.
- I would recommend the system to others.
- Using the system is a pleasant experience.
- The system is useless.
- Using the system is invigorating.
- The system makes me more aware of my choice options.
- Using the system makes me happy.
- I make better choices with the system.
- I use the system to unwind.
- I can find better items using the recommender system.

Not included:

- I can find better items without the help of the system.
- I feel bored when I'm using the system.

Choice satisfaction

Included:

- I like the items I've chosen.
- I was excited about my chosen items.
- I enjoyed watching my chosen items.
- The items I watched were a waste of my time.
- The chosen items fit my preference.

Not included:

- I know several items that are better than the ones I selected.
- Some of my chosen items could become part of my favorites.
- I would recommend some of the chosen items to others/friends.

Intention to provide feedback

Included:

- I like to give feedback on the items I'm watching.
- Normally I wouldn't rate any items.
- I only sparingly give feedback.
- I didn't mind rating items.

Not included:

- In total, rating items is not beneficial for me.

General trust in technology

Included:

- Technology never works.
- I'm less confident when I use technology.
- The usefulness of technology is highly overrated.

- Technology may cause harm to people.

System-specific privacy concern

Included:

- I'm afraid the system discloses private information about me.
- The system invades my privacy.
- I feel confident that the system respects my privacy.

Not included:

- I'm uncomfortable providing private data to the system.
- I think the system respects the confidentiality of my data.

FT3 BBC pre-trial

Perceived recommendation quality

Included:

- The recommended items were relevant.
- I liked the recommendations provided by the system.
- The recommended items fitted my preference.
- The MyMedia recommender is providing good recommendations.
- I didn't like any of the recommended items.
- The MyMedia recommender is not predicting my ratings accurately.
- The recommendations did not include my favorite programmes.

Perceived recommendation variety

Included:

- The recommendations contained a lot of variety.
- All the recommended programmes were similar to each other.

Perceived system effectiveness

Included:

- The MyMedia recommender is useful.
- I would recommend the MyMedia recommender to others.
- The MyMedia recommender has no real benefit for me.
- I can save time using the MyMedia recommender.
- I can find better programmes without the help of the MyMedia recommender.
- The MyMedia recommender is recommending interesting content I hadn't previously considered.

Not included:

- The system gave too many recommendations.

FT4 BBC trial

Perceived recommendation quality

Included:

- The MyMedia recommender is providing good recommendations.
- I liked the recommendations provided by the system.
- The recommended items fitted my preference.
- The recommended items were relevant.
- I didn't like any of the recommended items.
- The MyMedia recommender is not predicting my ratings accurately.
- The recommendations did not include my favorite programmes.

Perceived recommendation variety

Included:

- The recommendations contained a lot of variety.
- The MyMedia recommender is recommending interesting content I hadn't previously considered.
- The recommendations covered many programme genres.
- All the recommended programmes were similar to each other.
- Most programmes were from the same genre.

Perceived system effectiveness

Included:

- The MyMedia recommender has no real benefit for me.
- I would recommend the MyMedia recommender to others.
- The MyMedia recommender is useful.
- I can save time using the MyMedia recommender.
- I can find better programmes without the help of the MyMedia recommender.

Not included:

- The system gave too many recommendations.

EX1 Choice overload experiment

Perceived recommendation variety

Included:

- The list of recommendations was varied.
- The list of recommendations included movies of many different genres.
- Many of the movies in the list differed from other movies in the list.
- All recommendations seemed similar.

Not included:

- No two movies in the list seemed alike.
- The list of recommendations was very similar/very varied.

Perceived recommendation quality

Included:

- The list of recommendations was appealing.
- How many of the recommendations would you care to watch?
- The list of recommendations matched my preferences.
- I did not like any of the recommendations in the list.

Choice difficulty

Included:

- Eventually I was in doubt between … items.
- I changed my mind several times before making a decision.
- I think I chose the best movie from the options.
- The task of making a decision was overwhelming.

Not included:

- How easy/difficult was it to make a decision?
- How frustrating was the decision process?

Choice satisfaction

Included:

- My chosen movie could become one of my favorites.
- How satisfied are you with the chosen movie?
- I would recommend the chosen movie to others.
- I think I would enjoy watching the chosen movie.
- I would rather rent a different movie from the one I chose.
- I think I chose the best movie from the options.

Not included:

- The list of recommendations had at least one movie I liked.

Expertise

Included:

- I am a movie lover.
- Compared to my peers I watch a lot of movies.
- Compared to my peers I am an expert on movies.

EX2 Diversification experiment

Perceived recommendation accuracy

Included:

- The recommended movies fitted my preference.
- Each of the recommended movies was well-chosen.
- I would give the recommended movies a high rating.
- The provided recommended movies were interesting.
- I liked each of the recommended movies provided by the system.
- Each of the recommended movies was relevant.

Not included:

- I did not like any of the recommended movies.

Perceived recommendation variety

Included:

- Several movies in the list of recommended movies were very different from each other.
- The list of recommended movies covered many genres.
- The list of recommended movies had a high variety.
- Most movies were from the same type.
- The list of recommended movies was very similar/ very varied.

Not included:

- All recommended movies were similar to each other.

Choice difficulty

Included:

- Selecting the best movie was very easy / very difficult.
- Comparing the recommended movies was very easy / very difficult.
- Making a choice was overwhelming.
- I changed my mind several times before choosing a movie.

Not included:

- Making a choice was exhausting.
- Making a choice was fun.
- Making a choice was frustrating.
- Eventually I was in doubt between … movies.

Perceived system effectiveness

Included:

- The recommender system gave me valuable recommendations.
- I would recommend the recommender system to others.

- I can find better movies using the recommender system.
- I make better choices with the recommender system.
- The recommender system is useless.
- The recommender system makes me more aware of my choice options.
- I don't need the recommender system to find good movies.

Choice satisfaction

Included:

- My chosen movie could become one of my favorites.
- The chosen movie fits my preference.
- I will enjoy watching my chosen movie.
- I like the movie I have chosen.
- I will recommend the movie to others/friends.
- Watching my chosen movie will be a waste of my time.

Not included:

- I am excited about my chosen movie.

Expertise

Included:

- Compared to my peers I watch a lot of movies.
- Compared to my peers I am an expert on movies.
- I only know a few movies.

Not included:

- I am a movie lover.

Author Biographies

Bart P. Knijnenburg

University of California, Irvine, Department of Informatics, Donald Bren School of Information and Computer Sciences, Irvine, CA, USA 92697

Bart P. Knijnenburg is a Ph.D. candidate in Informatics at the University of California, Irvine. His work focuses on the user experience of recommender systems, adaptive preference elicitation methods, and privacy-aware interfaces for adaptive systems. He received his B.S. degree in Innovation Sciences and his M.S. degree in Human-Technology Interaction from Eindhoven University of Technology (TU/e), The Netherlands, and his M.A. degree in Human-Computer Interaction from Carnegie

Mellon University. The work described in the current paper was conducted while employed as a researcher on the MyMedia project at TU/e.

Dr. Martijn C. Willemsen

Eindhoven University of Technology (TU/e), Human-Technology Interaction group, P.O. Box 513, 5600MB Eindhoven, The Netherlands

Martijn Willemsen holds an M.S. in Electrical engineering and in Technology and Society, and a PhD degree in cognitive psychology (decision making) from the TU/e. Since 2002 he is an assistant professor in the HTI group at TU/e. From 2003-2004 he was a visiting PostDoc at Columbia University (New York) to work with Prof. Eric Johnson on online (web-based) process tracing methods. His expertise is in online consumer behavior and decision-making, and in user-centric evaluation. His current research includes the application of the psychology of decision making in recommender systems, and on this topic he served as an advisor in the MyMedia project from which the current paper originates.

Zeno Gantner

University of Hildesheim, Information Systems and Machine Learning Lab, Marienburger Platz 22, 31141 Hildesheim, Germany

Zeno Gantner is a researcher at University of Hildesheim's Information Systems and Machine Learning Lab. He received a Dipl.-Inf. degree (Master in computer science) from University of Freiburg, Germany. His research focus is machine-learning techniques for recommender systems. He is the main author of the MyMediaLite recommender algorithm library.

Hakan Soncu

European Microsoft Innovation Center GmbH, Ritterstrasse 23, 52072 Aachen, Germany

Hakan Soncu is a Software Development Engineer at EMIC. He received his M.S. in Software Systems Engineering from RWTH Aachen.

Dr. J.C. Newell

BBC R&D, Centre House, 56 Wood Lane, London, W12 7SB, UK

Dr. Newell is a Lead Technologist at BBC Research & Development. He received his B.A. in Physics and D.Phil. in Engineering from the University of Oxford. His primary interests are user interfaces and the associated metadata in digital broadcasting systems and he has contributed to the DVB, ID3 and TV-Anytime standards.