A Pragmatic Procedure to Support the User-Centric Evaluation of Recommender Systems

Bart P. Knijnenburg
Department of Informatics
University of California, Irvine
Bart.K@uci.edu

Martijn C. Willemsen
Human-Technology Interaction Group
Eindhoven University of Technology
M.C.Willemsen@tue.nl

Alfred Kobsa
Department of Informatics
University of California, Irvine
Kobsa@uci.edu

ABSTRACT
As recommender systems are increasingly deployed in the real world, they are not merely tested offline for precision and coverage, but also "online" with test users to ensure good user experience. The user evaluation of recommenders is however complex and resource-consuming. We introduce a pragmatic procedure to evaluate recommender systems for experience products with test users, within industry constraints on time and budget. Researchers and practitioners can employ our approach to gain a comprehensive understanding of the user experience with their systems.

Categories and Subject Descriptors
H.1.2. **[Models and principles]**: User/Machine Systems–*software psychology*; H.5.2 **[Information Interfaces and Presentation]**: User Interfaces–*evaluation/methodology*; H.4.2. **[Information Systems Applications]**: Types of Systems–*decision support*

General Terms
Measurement, Experimentation, Human Factors, Standardization.

Keywords
Recommender systems, user experience, user-centric evaluation

1. INTRODUCTION
Until recently, the field of recommender systems primarily focused on testing and improving the accuracy of prediction algorithms [2,8,17]. Increasingly, industry and academic researchers agree that the ultimate goal of recommenders is to help users make better decisions, and that high accuracy in itself does not guarantee this aim [8,9,11,12]. In fact, the recommender with the highest accuracy may not even be the most satisfying [7,15]. A plethora of other factors, such as the composition of the recommendation set [1], the preference elicitation method [13], and personal considerations (e.g. privacy [14] and domain knowledge [4,5]) may also considerably influence the user experience (UX) with recommenders.

In [6], we introduced and validated a framework for the user-centric evaluation of recommender systems for experience products [10], i.e. products whose qualities are difficult to determine in advance but can be ascertained upon consumption (e.g., music, movies, books). This framework describes how different aspects of recommender systems influence users' experience and interaction. In contrast to (offline) algorithm evaluation, our approach analyzes the interaction of "live" test users with functional interactive systems and their reported experience. Whereas algorithmic accuracy can be described using objective and uniform metrics, UX is a complex interplay of subjective, psychological constructs that can often only be measured indirectly with comprehensive questionnaires, extensive logging of user behavior, and intricate statistical analysis.

However, when the goal is to merely get a basic idea of the factors influencing the UX with a system, the evaluation may be simplified. To that end, we take our validated framework and simplify the operationalizations of the constructs and the statistical analyses used to test the relations between them. The result is a *pragmatic procedure* for testing the effect of certain aspects of a recommender system on the UX with that system.

RecSys'11, October 23–27, 2011, Chicago, Illinois, USA.

2. TOWARDS A PRAGMATIC APPROACH
Our framework provides an empirically validated theoretical foundation for the concepts and relations underlying the UX with recommender systems. Below we discuss how its core can be adapted for our pragmatic procedure.

2.1 The Framework
Fig. 1 shows the framework described in [6], instantiated with the proposed components of the pragmatic procedure. At its center lies the user experience: the user's evaluation of the system (perceived system effectiveness and fun), system usage (usage effort and choice difficulty), and outcome of system usage (satisfaction with the chosen items).

The framework distinguishes itself from prior work [12] by acknowledging that the concept of UX is only useful if it can explain how specific aspects of the system (rounded rectangles in Fig. 1) cause differences in UX. Experiments based on the framework therefore assign participants to two or more versions of the same system ("conditions") that differ in one system aspect only. Differences between the reported experiences in these conditions can then be attributed to those objective system aspects.

When testing subtly different conditions, the link between objective system aspects and UX may not be particularly strong; e.g., some users may not even notice a change in algorithmic accuracy. Our framework therefore introduces subjective system aspects (e.g. perceived quality or variety) that mediate the link from objective system aspects to UX. They help explain how and why a system aspect might influence the UX.

The UX may also be influenced by personal characteristics (e.g. demographics, domain knowledge and general trust) and situational characteristics (e.g. privacy concerns). These characteristics are beyond the influence of the system, but they are important moderators to consider in user-experience evaluations.

Differences in the UX are likely to be related to differences in behavior. In a commercial setting, certain behaviors (e.g. purchases, exposure to ads) are of primary importance. We frequently

observe however that online behavior is less robust than questionnaire responses. Moreover, while behavior captures the present interaction, UX has often been used as a predictor for adoption [3]. Behavior is also harder to interpret than subjective responses [16]: when a user inspects more recommendations, does this indicate liking or a desperate search for better items? To resolve these tensions, our framework triangulates behavioral data with subjectively measured experience concepts.

The framework provides a chain of effects, a link between objective aspects and objective user behavior, mediated by several subjective constructs. The subjective constructs explain how and why the user's experience with the recommender system comes about. This explanation constitutes the main value of the framework.

2.2 Simple Subjective Measures

In [6] we suggested measuring the subjective concepts by asking users multiple questions. A single question per concept is not advisable, as users may interpret it differently. Moreover, multiple questions allow the researcher to validate untested concepts, and to merge and split concepts when needed. After repeated validation, we can now identify a number of stable concepts though (rectangles in Fig. 1), as well as one or two questions per concept (text in italics) that measure them reasonably well.

2.3 Substituting Process Data

Behavioral measures may at times be ambiguous and less robust than subjective measures. We suggest that subjective measures be used whenever possible, which is not always the case. Some recommenders are designed to be inconspicuous and not to claim users' attention; asking questions may ruin this experience. Moreover, it may sometimes be nearly impossible for users to fill out a questionnaire (e.g. on a TV). In Fig. 1, we postulate several process data measures (ovals) that have shown robust correlations (double arrows) with certain subjective concepts [6].

2.4 From SEM to T-tests and Correlations

In [6] we use structural equation models (SEMs) to test the causal relations among measured concepts. SEMs test the robustness of our measures, the fit and invariance of the proposed model, and the ad hoc inclusion of unexpected effects. Due to their complexity, SEMs usually require a large amount of data to fit a model.

In Fig. 1 we postulate a number of simple correlations between concepts (arrows) that have been repeatedly validated in our previous work. These correlations can be used as hypotheses to guide researchers in testing the effect of their manipulations on perceptions (e.g., that a new algorithm will have a higher perceived recommendation quality), and how these perceptions influence the UX (e.g., that perceived recommendation quality correlates with perceived system effectiveness).

3. THE PRAGMATIC PROCEDURE

Participants should first be randomly assigned to one of the conditions of the study (Section 3.1). They should then be informed about its basic goals, but this explanation should not influence their behavior (e.g., they should not know which condition they are in). Participants then interact with the system, and their behavior is logged (Section 3.2). Afterwards, users are asked a set of questions (Section 3.3) to measure their perceptions and experiences using the system, and their personal and situational characteristics that may influence the UX. Finally, the collected data can be analyzed using standard tools such as Excel (Section 3.4).

3.1 Assign Participants to Conditions

The goal of the procedure is to measure the effect of specific aspects of a recommender (*objective system aspects*) on the UX. This can be accomplished through experimental manipulation: participants are randomly assigned to use one of several versions of the system that differ in one aspect only ("conditions"). We prefer a between-subjects over a within-subjects approach, in order to prevent "spill-over" effects in the UX that often occur when evaluating entire systems. Our research indicates that the algorithm, the composition of the set of recommendations, and the preference elicitation method have a significant impact on the UX. These are however only examples; the study goals will dictate which objective system aspects to measure and manipulate.

When several aspects are manipulated at the same time, this should be done orthogonally. For example, when testing two algorithms A and B, and two preference elicitation methods X and Y, four conditions should be tested: A+X, A+Y, B+X and B+Y. If only two conditions were tested (e.g. A+X, B+Y), it is impossible to know which of the two aspects caused the differences in UX.

3.2 Log Interaction Behavior

Our research suggests that when it comes to interacting with a recommender system, browsing is bad and consumption is good. One can thus measure *browsing behavior* (clicks, time) and *consumption* (purchase item, use item) as behavioral proxies for UX. Moreover, choice difficulty can be measured via *acquisition time and frequency* (how long and how often users inspect items) [1]. Note however that these behavioral measures may not have the

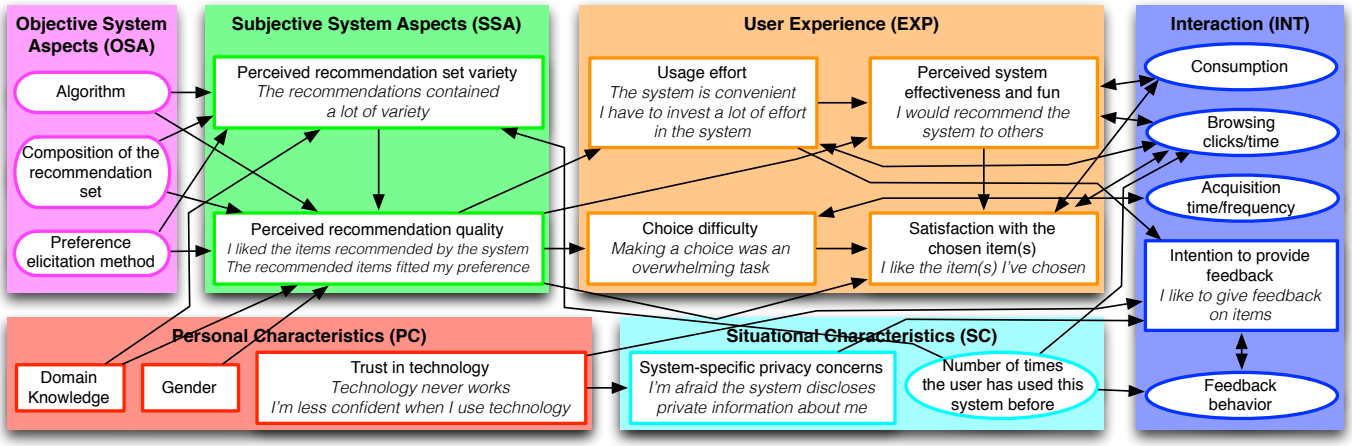

Figure 1. The framework for user-centric evaluation of recommender systems. Shown are the concepts of our procedure (boxes), relations between them (arrows), and key questionnaire items for measuring them (italics)

same effect in every system. It is therefore advisable to also gauge the subjective experience and relate it to the observed behavior.

One may also investigate which factors influence the *amount of feedback* (e.g. ratings) that users provide. Users should then also be asked about their *intention to provide feedback*, as their intention may not always be in line with their actual behavior. A good item for measuring feedback intentions is "I like to give feedback on items". Feedback behavior (and intentions) can be related to the UX, and to users' trust and privacy concerns.

3.3 Measure the Subjective Experience

After using the system, participants receive questionnaires that measure UX concepts. They are asked to expresss agreement or disagreement with certain statements on a five-point scale ranging from "I totally disagree" to "I totally agree". Below we outline the concepts established in our previous work, as well as the items that best measure these concepts, and the observed correlations between them. Researchers can select a subset of measures that is most relevant to the goals of their study.

3.3.1 Subjective system aspects

Subjective system aspects measure whether users perceive the introduced manipulations. Two subjective system aspects identified in our research are *perceived recommendation set variety* (item: "The recommendations contained a lot of variety") and *perceived recommendation quality* (items: "I liked the items recommended by the system" and "The recommended items fitted my preference"). Perceived recommendation quality and variety can measure the perception of differences in algorithm quality, recommendation set compostion, or preference elicitation method. Moreover, we have repeatedly confirmed that recommendation set variety can influence perceived recommendation quality.

3.3.2 User Experience

UX measurements should clearly distinguish the evaluation objects defined in the framework: process, system and outcome.

Our research has established two process-related experience concepts: *usage effort* and *choice difficulty*. Usage effort concerns the time and effort needed to operate the system (items: "The system is convenient" and "I have to invest a lot of effort in the system"). Choice difficulty can be measured by the item: "Making a choice was an overwhelming task". Usage effort is negatively related to perceived recommendation quality, while choice difficulty is typically positively related to perceived recommendation quality.

Perceived system effectiveness is the system-related experience concept established in our research (item: "I would recommend the system to others"). If the system usage is supposed to be entertaining, one can add the item "I have fun when I'm using the system". System effectiveness is usually influenced by perceived recommendation quality and usage effort.

The outcome-related experience variable identified in our research is *satisfaction with the chosen item(s)* (item: "I like the item(s) I've chosen"). Satisfaction with the chosen item(s) can be related to perceived system effectiveness, choice difficulty, and perceived recommendation quality.

3.3.3 Personal and situational characteristics

Personal and situational characteristics can also influence the experience and interaction with a recommender system. *Gender* and *domain knowledge* can have an effect on perceived recommendation quality and variety. With *repeated system use*, perceived variety and feedback decrease while browsing increases. Two important characteristics in the light of feedback behavior are *general trust in technology* and *system-specific privacy concerns*. The former is a personal characteristic that can be measured with the items "Technology never works" or "I'm less confident when I use technology". The latter is a situational characteristic that depends on the system at hand. It can be measured with the item "I'm afraid that the system discloses private information about me". We propose these characteristics based on what we established in prior work. The study goals at hand may suggest other relevant personal and situational characteristics.

3.4 Analyze the Collected Data

When sufficient data is collected (at least 20 users per condition are typically required for adequate statistical power) the data can be analyzed by testing the significance and size of a subset of the effects outlined in Fig. 1. Spreadsheet or statistical software can be used to calculate a T-value (statistics), p-value (significance) and correlation (effect size) for each effect.

Let us consider the hypothetical data in Table 1. The experiment tested two algorithms X and Y with 10 participants[1] (N=10), and measured their perceived recommendation quality ("I liked the items recommended by the system"), the perceived system effectiveness ("I would recommend this system to others"), and the consumption of recommendations (in terms of the number of recommendations that were eventually followed up). Table 2 shows the Excel formulas for the analysis.

Table 1. Example data set

User ID	Algorithm	Perc. rec. quality	Perc. system effectiveness	Recs. followed up
1	X	3	2	5
2	X	2	2	11
3	X	3	2	9
4	X	4	3	3
5	X	1	1	6
6	Y	5	5	21
7	Y	4	4	4
8	Y	5	3	10
9	Y	4	2	15
10	Y	5	5	24

Table 2. Excel formulas for analysis of the example data set

	Independent samples T-test	Pearson correlation
Statistic (T)	`=T.INV(p, N-1)`	`=r*SQRT ((N-2)/(1-r^2))`
Significance (p)	`=T.TEST(C1:C5, C6:C10, 2, 3)`	`=1-T.DIST (T, N-1, TRUE)`
Effect size (r)	`=T^2/(T^2+N-1)`	`=CORREL (C1:C10, D1:D10)`

We first test whether users of the two algorithms judge the recommendation quality differently. The mean response to the item measuring this concept is 2.6 for algorithm X and 4.6 for Y. An independent-samples t-test shows that this difference is significant with a large effect size[2]: [t(9) = 2.65, p = .013, r = .439].

The next step is to test whether the perceived system effectiveness is indeed related to the perceived recommendation quality. A

[1] Fewer than suggested since the example is only for illustration.

[2] A typical threshold for significance is $p < .05$, meaning that the chance of incorrectly rejecting the null hypothesis of no effect is smaller than 5%. Accepted interpretations of effect size are: small/weak: r=.1, medium: r=0.3, large/strong: r=0.5.

Pearson correlation test shows that the answers of the two questions measuring these concepts are indeed very strongly and significantly correlated [r = .819, p = .0015].

We finally test whether the difference in perceived effectiveness is related to the number of recommendations being followed up. The correlation is strong and significant [r = .597, p = .0324].

We can also directly test for a difference in the number of recommendations that users followed per algorithm, but this difference is not significant [t(9) = 1.43, p = 0.093]. The difference in system effectiveness per algorithm is significant [t(9) = 2.07, p = 0.034], but this effect is mediated by perceived recommendation quality[3]. It is important to establish this mediation effect, as it explains why algorithm Y leads to a higher system satisfaction.

Based on these results we can draw the conclusion that algorithm Y has a higher perceived recommendation quality than algorithm X, which leads to a higher system satisfaction, which in turn leads to more recommendations being followed up (see Fig. 2).

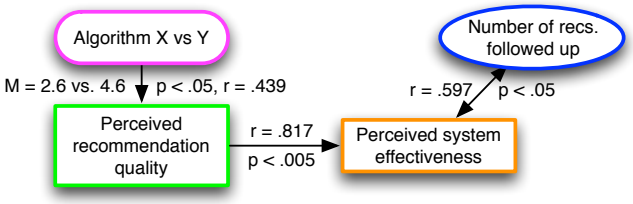

Figure 2. Graphical presentation of the analysis results

4. CONCLUSION AND FUTURE WORK

Grounded in our validated evaluation framework, our procedure provides a pragmatic approach to the user-centric evaluation of recommenders for experience products. Researchers and practitioners can use it to proceed from accuracy towards a more comprehensive understanding of users' experience with their systems. Practitioners can run their tests using Google Website Optimizer[4], which provides basic functionalities for randomized A/B testing and logging. Researchers of recommendation algorithms can incorporate the approach into existing research recommenders (such as MovieLens[5]), plug in their new algorithm, and see how it compares to other algorithms in terms of user experience.

5. REFERENCES

[1] Bollen, D. et al. 2010. Understanding choice overload in recommender systems. *Proc. of the 4th ACM conf. on Recommender systems*. RecSys'10. ACM, New York, NY, 63-70. DOI= http://doi.acm.org/10.1145/1864708.1864724.

[2] Cosley, D. et al. 2003. Is seeing believing? *Proc. of the SIGCHI conf. on Human factors in computing systems*. ACM, New York, NY, 585-592. DOI= http://doi.acm.org/10.1145/642611.642713

[3] Davis, F.D. 1989. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*. 13, 3 (Sep. 1989), 319-340. DOI= http://dx.doi.org/10.2307/249008

[4] Knijnenburg, B.P. and Willemsen, M.C. 2010. The effect of preference elicitation methods on the user experience of a recommender system. *Extended abstracts on Human factors in computing systems*. ACM, New York, NY, 3457-3462. DOI= http://doi.acm.org/10.1145/1753846.1754001

[5] Knijnenburg, B.P. and Willemsen, M.C. 2009. Understanding the effect of adaptive preference elicitation methods on user satisfaction of a recommender system. *Proc. of the 3rd ACM conf. on Recommender systems*. ACM, New York, NY, 381-384. DOI= http://doi.acm.org/10.1145/1639714.1639793

[6] Knijnenburg, B.P. et al. Explaining the User Experience of Recommender Systems. Accepted to *User Modeling and User-Adapted Interaction*. http://t.co/cC5qPr9

[7] McNee, S.M. et al. 2002. On the recommending of citations for research papers. *Proc. of the 2002 ACM conf. on Computer supported cooperative work*. ACM, New York, NY, 116-125. DOI= http://doi.acm.org/10.1145/587078.587096

[8] McNee, S.M. et al. 2006. Being accurate is not enough. *Extended abstracts on Human factors in computing systems*. ACM, New York, NY, 1097-1101. DOI= http://doi.acm.org/10.1145/1125451.1125659

[9] Murray, K.B. and Häubl, G. 2008. Interactive Consumer Decision Aids. *Handbook of Marketing Decision Models*. B. Wierenga, ed. Springer, Heidelberg, Germany, 55-77. DOI= http://dx.doi.org/10.1007/978-0-387-78213-3_3

[10] Nelson, P. 1970. Information and Consumer Behavior. *Journal of Political Economy*. 78, 2, 311-329. DOI= http://dx.doi.org/10.1086/259630

[11] Ozok, A.A. et al. 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, 1 (Jan. 2010), 57-83. DOI= http://dx.doi.org/10.1080/01449290903004012

[12] Pu, P. and Chen, L. 2010. User-Centric Evaluation Framework of Recommender Systems. *Proc. ACM RecSys 2010 Workshop on User-Centric Evaluation of Recommender Systems and Their Interfaces*. CEUR-WS Vol-621. 14-21.

[13] Pu, P. et al. 2008. Evaluating product search and recommender systems for E-commerce environments. *Electronic Commerce Research*. 8, 1-2 (May. 2008), 1-27. DOI= http://dx.doi.org/10.1007/s10660-008-9015-z

[14] Teltzrow, M. and Kobsa, A. 2004. Impacts of user privacy preferences on personalized systems. *Designing personalized user experiences in eCommerce*. C.-M. Karat, J. Blom, J. Karat, eds. Springer, Heidelberg, Germany, 315-332. DOI= http://dx.doi.org/10.1007/1-4020-2148-8_17

[15] Torres, R. et al. 2004. Enhancing digital libraries with TechLens+. *Proceedings of the 2004 joint ACM/IEEE conference on Digital libraries*. ACM, New York, NY, 228-236. DOI= http://doi.acm.org/10.1145/996350.996402

[16] van Velsen, L. et al. (2008). User-centered evaluation of adaptive and adaptable systems: a literature review. *Knowledge Engineering Review*. 23, 3, 261-281. DOI= http://dx.doi.org/10.1017/S0269888908001379

[17] Ziegler, C.-N. et al. 2005. Improving recommendation lists through topic diversification. *Proc. of the 14th intl. conf. on World Wide Web*. ACM, New York, NY, 22-32. DOI= http://doi.acm.org/10.1145/1060745.1060754

[3] http://www.davidakenny.net/cm/mediate.htm gives an excellent explanation of mediation analysis

[4] http://www.google.com/analytics/siteopt

[5] http://www.movielens.org