Evaluating Recommender Explanations:
Problems Experienced and Lessons Learned for the Evaluation of Adaptive Systems

Nava Tintarev[1], Judith Masthoff[2]

[1] Telefonica Research, nava@tid.es
[2] University of Aberdeen, j.masthoff@abdn.ac.uk

Abstract. We describe the methodological considerations that arose over a series of experiments evaluating the effectiveness of explanations for recommendations. In particular, we look at issues relating to: criteria, metrics, product domain used, choice of materials, possible confounding factors, and approximation of experience versus real experience. We generalize the problems we found and the solutions that we applied to adaptive systems. We illustrate the learned lessons with examples from our previous work on adaptive systems (ranging from adaptive learning to persuasive technologies).

1 Introduction

The evaluation of adaptive systems is not easy, and several researchers have pointed out potential pitfalls when evaluating adaptive systems. Examples of pitfalls mentioned in [1] and [2] include:

- Difficulty in attributing cause: is it the adaptation which is causing the measured effect or something else (such as system usability)?
- Insignificant results due to too much variance between participants. Adaptation is typically used when individual participants differ. However, individual differences are likely to lead to a large variance in results, and this makes it harder to get statistically significant results.
- Difficulty in defining the effectiveness of adaptation. It is sometimes hard to define what constitutes a good adaptation.
- Allocation of insufficient resources. You often need many participants to fully evaluate an adaptive system (in part due to the expected variance between participants mentioned above).
- Too much emphasis on summative rather than formative evaluation. Evaluations often measure only how good or bad a system is rather than providing information on where the problems are and how the system can be improved.

The difficulty in evaluating adaptive systems has led to a series of workshops on this topic such as [3, 4]. This paper contributes to the debate by identifying a number of problems related to and expanding those listed above. This is done

in the context of a case study, where we discuss the problems experienced when evaluating explanations of recommended items, and the solutions applied. We also discuss how these problems and solutions are more widely applicable to the evaluation of adaptive systems.

2 Background to Case Study

Recommender systems such as Amazon offer users recommendations, or suggestions of items to try or buy. These recommendations can then be explained to the user, e.g. *"You might (not) like this item because..."*. In experiments, our system generates (given a simple user model) explanations for items using data retrieved from the Amazon website. In a between-subject design, we compared three degrees of personalization in a series of experiments. The example below illustrates the three conditions for *one* experiment in the movie domain:

1. **Baseline:** The explanation is neither personalized, nor describes item features: e.g. *"This movie is one of the top 250 in the Internet Movie Database"*.
2. **Non-personalized, feature based:** e.g. *"This movie belongs to the genre(s): Drama. Kasi Lemmons directed this movie."* The feature 'director' was not particularly important to this participant.
3. **Personalized, feature based:** e.g. *"Unfortunately this movie belongs to at least one genre you do not want to see: Action & Adventure. Also it belongs to the genre(s): Comedy, Crime, Mystery and Thriller. This movie stars Jo Marr, Gary Hershberger and Robert Redford."* For this user, the most important feature is leading actors, and the explanation considers that the user does not like action and adventure movies.

3 Problems, solutions and generalizations

This section discusses some of the problems we encountered in our experiments evaluating explanations, and the solutions we adopted. It also elaborates on what other researchers can learn from our experiences for the evaluation of adaptive systems in general. We illustrate how the lessons can be generalized with example evaluations where the problem appears, and where solutions were applied. Although these problems are common to many evaluations of adaptive systems, it seemed fairer to select examples from our own work, as we have *also* been affected by these problems in our previous research.

3.1 Criteria to use

Problem. The first issue we had to resolve was what we meant by a *good* explanation.

Solution. We surveyed the literature and discovered that explanations can serve multiple aims, such as increasing transparency, trust, satisfaction, efficiency, persuasiveness, and effectiveness [5]. We decided to focus primarily on effectiveness:

how helpful explanations are for users to make good decisions. According to [6], an effective explanation minimizes the difference between the user's rating of an item based on the explanation, and the user's rating of the item after experiencing it. Therefore, in our experiments, participants rated the items based on the explanation, then re-rated the items after having experienced them[3].

There is some evidence to suggest that personalization may increase persuasion, or acceptance of recommendations [7]. In contrast, we wanted to investigate whether personalization would increase *effectiveness*. In addition to measuring effectiveness, we decided to also measure user satisfaction with the explanations.

Generalization of problem. Sometimes when adaptive systems are evaluated, there is a shortage of information about what exactly they are being evaluated on [1]. For example, an adaptive instruction system can be evaluated on how well it keeps the learner motivated, how much it improves the understanding of a weak learner, how much it is appreciated etc. All of these could serve as valid aims for a "good" system, but they are likely to require different types of evaluations. Also, often the discussion on other criteria that may have been relevant is limited, and results are presented as one system outperforming another one without saying on which criterion.

Generalization of solution. In order to achieve a goal of optimization, the optimum or the main evaluation criterion needs to be explicitly formulated prior to evaluation. As optimization in one criterion may damage another, it is also important to understand how the criteria relate to one other. For example, a study of computer generated reports of babies in intensive care found that doctors preferred graphical reports (satisfaction), but made better decisions with textual reports (effectiveness) [8]. In that evaluation, focusing on just one criterion, such as satisfaction, would not have provided a full picture of the situation.

3.2 Avoiding confounding factors

Problem. The measured impact of explanations on effectiveness may be confounded with the impact of the accuracy of the recommender system. If recommended items are meant to be liked by the user (i.e. we do not give predictions for items the user may not like), and the recommender system has poor accuracy, it would be hard to distinguish between the effects of poor explanations and poor accuracy. Most likely, we would not be able to tell which factor was the main contributor to a large change in valuation of an item. This is a problem because in a real recommender system, it would be hard to guarantee comparable recommendation accuracy between participants and between conditions. Also, to obtain reasonable accuracy, participants would need to use the system for a non-trivial amount of time prior to evaluation, in order to provide their

[3] Initially we used approximations of real experience, see Section 3.4 for a justification and further discussion.

preferences.

Solution. We decided not to use a recommender system, and used random item selection instead. This meant that we did not require the experiments to include a training period. Instead, participants' preferences were explicitly requested in order to personalize the explanations. Also, our metric for effectiveness (briefly described in Section 3.1) can be used regardless of whether participants liked or disliked the items; as long as they were able to make an initial assessment based on the explanation, we were able to study their change of opinion after experiencing the item. It was not a problem for us to offer explanations for items participants might not like as participants were told that the explanations were aimed at helping them make decisions (to try or *not* to try) rather than make (only positive) recommendations. This also meant that we did not have to control for recommendation accuracy.

Generalization of problem. As mentioned in the introduction (Section 1), difficulty in attributing cause has been previously discussed as a problem in the evaluation of adaptive systems [1].

Generalization of solution. Layered evaluation has been mentioned as a way to help overcome this problem [9]. In our solution, we effectively used the dicing approach proposed in [1]: we focused on the functionality of interest, and evaluated that functionality in isolation. In a similar way, we have previously evaluated parts of an adaptive learning system [1, 10].

Another solution for an adaptation taking time would be to run an experiment in several installments - first training the system on participants, and then using the adapted system for further evaluations, such as in [11]. If this approach is used, the evaluation design needs to consider that participants may drop out (i.e. consider retention rates), and one may still need to control for confounding factors.

3.3 Domain to use

Problem. The effectiveness of explanations and degree of personalization may well depend on the domain used.

Solution. We surveyed the literature and found that there has been a great deal of debate about classification of products into different categories in economics (see [12] for examples, and [13] for an elaboration on our chosen classifications). For our research, we decided that we should at least distinguish between:

- products which are relatively easy to evaluate objectively and those which commonly require an experiential and subjective judgment
- products which are relatively cheap and those which are more expensive

Ideally, we would have evaluated the explanations in four domains, considering each of the four combinations along these two dimensions. However, this would

have been very resource intensive: requiring not only substantially more participants but also detailed investigation into aspects of each domain (such as product features, appropriate material selection, selection of baseline explanations, etc). Lack of resources (e.g. time, sufficient and suitable participants) is often an issue in evaluations [2]. We decided to use a middle ground: instead of fully exploring all domain options, we chose to use two that differed with regard to both of the dimensions mentioned above: movies (cheap and subjective) and cameras (more expensive and objective). We had to perform two user-centered investigations to find appropriate features of movies and cameras (see e.g. [14]). We also had to decide what materials and baseline explanations to use in both domains. However, the additional effort involved in studying two domains was justified: we found the same results for both movie and cameras, providing us with more confidence that our results generalize across domains.

Generalization of problem. Evaluations of adaptive systems tend to focus on evaluating the system in one particular domain, often without mentioning the limitations this puts on the results. For example, when evaluating different algorithms for a group recommender system, we used the domain of video clips [15]. We drew conclusions on what algorithms people preferred (such as avoiding misery for others). However, it is more difficult to say if similar results would have been found if we had used another domain, such as news items, courses of a seven-course meal, etc. The expected duration for items and the expected impact of experiencing an unliked item is likely to differ between domains, and may well affect the final results.

Generalization of solution. The solution of surveying domains, and evaluating in multiple domains is applicable to adaptive systems in general. For example, in an adaptive e-learning system, we evaluated an adaptive item sequencing strategy for two learning tasks: a paired-associate learning task and a concept learning task [10]. This does not cover all possible learning tasks in the learning domain, but does give some insight into how generalizable the results are. So, whenever feasible, adaptive systems should be evaluated in multiple domains. If resources do not permit, at least a discussion of the possible impact of other domains should be included.

3.4 Approximation of experience versus real experience

Problem. It can be very difficult and time consuming for participants to really experience the recommended items. For example, it may take too long (or be too expensive) for participants to read a recommended book or go on a recommended holiday. Participants may also require time to *fully* experience a product (for example, a real experience of a camera may involve using it over a couple of days, so that initial technical difficulties do not overly influence the participants' final evaluation).

Solution. Previous work has approximated experience of the recommended item, for example by letting participants read online reviews [6]. In our initial experiments (for both the movie and camera domains) we used the approximation of reading on-line Amazon reviews (see e.g. [16]). Deciding on an appropriate approximation required careful consideration. For example, for movies we considered using trailers, but decided against this, as these are typically made to persuade people to see the movie rather than help them make informed decisions (which is what we wanted to achieve when we defined our aim as effectiveness). However, online reviews may also be positively biased. Therefore, after several experiments using approximation, we decided to run another experiment where users really experienced the items. We used the movie domain as movies are relatively cheap, and it is easier and faster for participants to judge movies than e.g. cameras. So, our solution has been to approximate for a number of initial studies, in order to adjust and perfect the evaluation, before a costly and time-consuming real experience evaluation.

Generalization of problem. It can be very time consuming for participants to fully experience adaptive systems. For example, a realistic experience of an adaptive learning system involves learners using it over multiple sessions, learning something they would *normally* be learning in another setting. Instead, we often evaluate over one or two sessions, sometimes using a controlled artificial learning domain. For example, when evaluating adaptive navigation in a learning system, we have used the artificial domain of square dancing, with participants learning to operate dancers on the screen using multiple computer-based lessons, but all within a one-hour session [17].

Generalization of solution. Approximation of experience is often a reasonable thing to do in early evaluations, as it requires less time, allows better experimental control, and is sometimes better for ethical reasons. This can then be followed by an evaluation in more realistic settings, often of a longitudinal nature. We have used approximation in many of our studies. For example, we measured whether an emphatic embodied agent influenced participants' mood after inducing a negative mood in an artificial test (rather than say a real course assessment) [18].

3.5 Choice of materials

Problem. For the real experience valuation we needed items that participants had not yet experienced, that had enough interesting features to produce an explanation, would not take too long to evaluate, and that were ethically ok to use.

Solution. To reduce time needed per item, we used short movies instead of full movies. To avoid movies that participants had already seen, an I-have-already-seen-this-movie button that skipped to another movie was added. To avoid exposing participants to sensitive material such as (extreme) violence or sex, only

movies suitable for 15 years and over (PG-15) were used. By choosing non-offensive movies, there was a distinct risk that participants' ratings of movies would not be as well spread over possible rating values as they could be. We used a pilot study to confirm that while the distribution may not make full use of the possible values, participants still indicate values that differed sufficiently from the mid-point to warrant an interesting analysis. We also considered the presences of relevant features when selecting the movies. We included movies with actors (e.g. Rowan Atkinson) or directors (e.g. Tim Burton) that were likely to be known. As certification rating (e.g. PG - parental guidance advised) was used as a feature for explanation generation, we also selected movies that had an international certification, which is otherwise often missing for short movies.

We also found that most short movies use less famous actors/directors, and therefore, despite our best efforts, there was a higher frequency of unknown actor and director names than in our previous experiments [16]. Most likely as a consequence of this, we found that participants were less happy with personalized explanations (actors and directors are the most common preferred features aside from genre). Additionally, the ethical considerations had the side effect that the majority of movies watched belonged to the genres comedy, animation and children. So, there is not always a perfect solution, but at least being aware of the potential impact of materials can help explain results and understand the limitations of a study.

Generalization of problem. Fields such as psychology have a common practice of carefully selecting the materials they use in experiments. This is done either because the material may affect the outcome (we also discuss avoiding confounding factors in Section 3.2) or for other reasons such as ethical ones. In the evaluation of adaptive systems, we are often told what materials were used, but not on the basis of which criteria they have been chosen, and whether pilot studies have been done to validate their appropriateness.

Generalization of solution. The criteria on the basis of which materials are chosen need to be clearly defined and stated. In addition, pilot studies need to be performed to test the suitability of materials. For example, when studying the effect of adding a doctor's photo on website credibility, we needed the photo to contain an image of a doctor that would be considered credible in this domain. We found such a photo by running pilot studies in which participants judged the profession of the person depicted, and rated their domain credibility using validated metrics [19].

3.6 Appropriate measurement

Problem. To measure true effectiveness, we needed to distinguish between participants not having formed an opinion of the item, and participants believing the item is kind of average (middle of the scale).

Solution. We decided to add a separate opting-out option as an alternative for rating, for participants who were not able to formed an opinion. In our analyses of effectiveness, we excluded ratings of participants who had opted-out. One thing we noticed in all our experiments was that while baseline explanations did surprisingly well for effectiveness, they also led to a very high opt-out frequency: participants were unable to provide a rating. This means that only considering the change between the before and after rating for those people who opt-in is not a true reflection of effectiveness.

Despite the opportunity to opt-out, we still found that some participants seemed to use the middle of the scale to indicate that they had no real opinion about a movie. This could be seen in the higher frequency of middle of the scale ratings for baseline explanations, or when participants were asked to give a rating based on only a movie title (without the explanation).

Problem. The effectiveness metric from [6] considers the difference between the before and after ratings. However, they do not discuss the effects of over- and underestimation (which could lead respectively to trying an item you may not end up liking and missing an item you may have liked). So, the question arises whether an explanation leading to an overestimation is as bad as one leading to a similarly big underestimation. And how about the position of the gap? Does it matter whether (on a scale from 1 to 5) the pre-rating is 3 and the post-rating 5, compared to a pre-rating of 1 and a post-rating of 3? To complicate matters further, does it all differ per domain type?

Solution. We investigated these questions [13], and found that:

- Overestimation was considered more severely than underestimation
- Overestimation was considered more severely in high investment domains compared to low investment domains (see also Section 3.3 on domain effects).
- Gaps which remained in the negative half of the scale were considered more severely than gaps which crossed over from good to bad (or vice-versa), and gaps which remained in the positive half of the scale.

Generalization of problems. In the evaluation of all systems, including adaptive ones, one has to take care that the metric used is really measuring what you want to find out. For example, when assessing personalisation in interactive TV often the time spent watching a programme is used as an indication of user interest. However, longer viewing times may well have been caused by other factors such as the viewer having a coffee or even being so bored that they fell asleep. Learners spending more time on a lesson may mean that they are more motivated or that they find the lesson harder to understand.

Generalization of solutions. A critical analysis is required of all metrics used, asking whether there are situations when the value given by the metric is not accurate. For a metric to be good, the same value should have the same meaning independent of the circumstances. For example, our experience shows that the

effectiveness metric of [6] falls short of this, as e.g. over- and underestimation may lead to the same value while having a different effect on users.

4 Conclusions

This paper illustrates some of the problems we have encountered when investigating the effectiveness of explanations in a recommender system. Our investigations consisted of a series of experiments, and in each experiment we improved our understanding of the evaluation design. As we have discussed, many of these issues are generalizable to other types of adaptive systems. In any evaluation, it is important to:

- Decide which criteria to use
- Avoid confounding factors
- Take into account domain effects
- Build up the experiment gradually, and consider limited resources
- Take into account the effects of the material you select
- Consider if a metric really measures what you want

From this paper it would perhaps be easy to conclude that it is hard (and probably impossible) to design the perfect evaluation. In retrospect, there is always something else that could impact the results. Clearly defining your goals, metrics and refining your design through a sequence of experiments, and gradually investigating different aspects, will however help you avoid the most common pitfalls - may your next evaluation be a successful one!

References

1. Masthoff, J.: The evaluation of adaptive systems. In: Adaptive evolutionary information systems. Idea Group publishing (2002) 329–347
2. Weibelzahl, S.: Problems and pitfalls in evaluating adaptive systems. In: Fourth Workshop on the Evaluation of Adaptive Systems in conjunction with UM'05. (2005) 57–66
3. Weibelzahl, S., A., P., Masthoff, J., eds.: Fifth Workshop on User-Centred Design and Evaluation of Adaptive Systems, associated with AH'06, Dublin (2006)
4. Weibelzahl, S., Paramythis, A., Masthoff, J., eds.: Fourth Workshop on the Evaluation of Adaptive Systems, associated with UM'05, Edinburgh (2005)
5. Tintarev, N., Masthoff, J.: A survey of explanations in recommender systems. In: WPRSIUI associated with ICDE'07. (2007) 801–810
6. Bilgic, M., Mooney, R.J.: Explaining recommendations: Satisfaction vs. promotion. In: Proceedings of the Wokshop Beyond Personalization, in conjunction with IUI. (2005) 13–18
7. Carenini, G., Moore, D.J.: An empirical study of the influence of user tailoring on evaluative argument effectiveness. In: IJCAI. (2001)
8. Law, A.S., Freer, Y., Hunter, J., Logie, R., McIntosh, N., Quinn, J.: A comparison of graphical and textual presentations of time series data to support medical decision making in the neonatal intensive care unit. J Clin Monit Comput. **19(3)** (2005) 183–94

9. Paramythis, A., Totter, A., Stephanidis, C.: A modular approach to the evaluation of adaptive user interfaces. In Weibelzahl, S., Chin, D.N., Weber, G., eds.: Evaluation of Adaptive Systems in conjunction with UM'01. (2001) 9–24
10. Masthoff, J.: An Agent-Based Interactive Instruction System. PhD thesis, Eindhoven University of Technology (1997)
11. Amatriain, X., Pujol, J.M., Oliver, N.: I like it... i like it not: Evaluating user ratings noise in recommender systems. In: UMAP. (2009)
12. Cho, Y., Im, I., Hiltz, J.F.S.R.: The impact of product category on customer dissatisfaction in cyberspace. Business Process Managment Journal **9 (5)** (2003) 635–651
13. Tintarev, N., Masthoff, J.: Over- and underestimation in different product domains. In: Workshop on Recommender Systems associated with ECAI. (2008) 14–19
14. Tintarev, N., Masthoff, J.: Effective explanations of recommendations: User-centered design. In: Recommender Systems. (2007) 153–156
15. Masthoff, J.: Group modeling: Selecting a sequence of television items to suit a group of viewers. User Modeling and User Adapted Interaction **14** (2004) 37–85
16. Tintarev, N., Masthoff, J.: Personalizing movie explanations using commercial meta-data. In: Adaptive Hypermedia. (2008) 204–213
17. Masthoff, J.: Design and evaluation of a navigation agent with a mixed locus of control. In Cerri, S., Gouardres, G., Paraguau, F., eds.: Intelligent Tutoring Systems, Berlin, Springer Verlag (2002)
18. Nguyen, H., Masthoff, J.: Designing empathic computer: The effect of multimodal empathic feedback using animated agent. In: Persuasive Technology, Claremont, USA, Springer Verlag (2009)
19. Nguyen, H., Masthoff, J.: Is it me or is it what i say? source image and persuasion. In: Persuasive Conference, Springer Verlag (2007)