Evaluating the effectiveness of explanations for recommender systems

Methodological issues and empirical studies on the impact of personalization

Nava Tintarev · Judith Masthoff

Abstract When recommender systems present items, these can be accompanied by explanatory information. Such explanations can serve seven aims: effectiveness, satisfaction, transparency, scrutability, trust, persuasiveness, and efficiency. These aims can be incompatible, so any evaluation needs to state which aim is being investigated and use appropriate metrics. This paper focuses particularly on effectiveness (helping users to make good decisions) and its trade-off with satisfaction. It provides an overview of existing work on evaluating effectiveness and the metrics used. It also highlights the limitations of the existing effectiveness metrics, in particular the effects of under- and overestimation and recommendation domain. In addition to this methodological contribution, the paper presents four empirical studies in two domains: movies and cameras. These studies investigate the impact of personalizing simple feature-based explanations on effectiveness and satisfaction. Both approximated and real effectiveness is investigated. Contrary to expectation, personalization was detrimental to effectiveness, though it may improve user satisfaction. The studies also highlighted the importance of considering opt-out rates and the underlying rating distribution when evaluating effectiveness.

Keywords Recommender systems · Metrics · Item descriptions · Explanations · Empirical studies

N. Tintarev
University of Aberdeen
Tel.: +44-1224-274620
Fax: +44-1224-273422
E-mail: n.tintarev@abdn.ac.uk

J. Masthoff
University of Aberdeen
Tel.: +44-1224-272299
E-mail: j.masthoff@abdn.ac.uk

1 Introduction

While recommender systems have traditionally been evaluated in terms of recommendation accuracy, in recent years, interest has increased in more user-centered evaluation metrics such as user satisfaction (e.g. McNee et al 2006a; Tintarev and Masthoff 2007b, 2010; Ricci et al 2010). These metrics can be influenced by more than just the recommendations, such as by explanations, the way recommendations are presented, and the method of interacting with recommendations (Tintarev and Masthoff 2010). This paper investigates the role of explanations. It is sometimes erroneously assumed that explanations should always justify why items have been recommended. A popular definition of explain is *"to justify"*. However, to explain also means *"to make clear by giving a detailed description"* [Oxford concise dictionary]. So, an explanation can be an item description that helps the user to understand the qualities of the item well enough to decide whether it is relevant to them or not.

In their work on explanations, Herlocker et al (2000) noted that many recommender systems provided no transparency into the working of the recommendation process, nor offered any additional information to accompany the recommendations. Since then, the body of research on explanations in recommender systems has continued to grow (e.g., Bilgic and Mooney 2005; McCarthy et al 2005b; McSherry 2005; Pu and Chen 2007; Cramer et al 2008a; Tintarev and Masthoff 2008b; Guy et al 2009a; Vig et al 2009; Tintarev and Masthoff 2010).[1]

As indicated above, explanations can serve multiple aims. For example, explanations can provide transparency, exposing the reasoning and data behind a recommendation. This is the case with some of the explanations hosted on Amazon.com, such as: *"Customers Who Bought This Item Also Bought . . . "*. Alternatively, explanations can be more focused on helping users make decisions (about the items) that they are happy with: effectiveness. An effective explanation may be formulated along the lines of *"You might (not) like this item because..."*. In contrast to the Amazon example above, this explanation does not *necessarily* describe how the recommendation was selected - in which case it is not transparent: *"It is a funny comedy"* could be an effective explanation even when the recommendation was based on collaborative filtering.

Table 1 shows seven possible aims for explanations. These aims can be complementary (e.g. effectiveness may increase trust) or contradictory (e.g. persuasiveness may decrease effectiveness). Relations between aims are not always clear-cut. For example, transparency may lead to an increase or decrease in trust, depending how much confidence users have in the internal working of the system shown to them. The main explanatory aims need to be decided prior to the design and evaluation of (optimal) explanations. As optimization in one criterion may damage another, it is important to consider how the criteria relate to one other (Tintarev and Masthoff 2009).

---

[1] The current interest in explanation-aware computing extends beyond recommender systems (e.g. Roth-Berghofer et al 2008, 2009, 2010).

Table 1 Explanatory aims

Aim	Definition
Transparency	Explain how the system works
Scrutability	Allow users to tell the system it is wrong
Trust	Increase users' confidence in the system
Effectiveness	Help users make good decisions
Persuasiveness	Convince users to try or buy
Efficiency	Help users make decisions faster
Satisfaction	Increase the ease of use or enjoyment

Table 2 provides an overview of recommender systems with explanations, focusing on papers that evaluated the generated explanations. See Tintarev and Masthoff (2010) for more detailed definitions, examples, and a discussion on how factors such as the degree of interaction with the recommender system and the way in which recommendations are presented may impact evaluations on these aims.

This paper investigates the effectiveness of explanations: what makes an explanation effective (in the sense of helping users to make good decisions), and how effectiveness can best be evaluated. However, optimizing effectiveness may negatively impact other criteria. Therefore, two criteria (effectiveness and satisfaction) will be considered in our studies below, to investigate potential trade-offs.

We consider explanations for all items, not just for items predicted to be liked by the user. As discussed in Tintarev and Masthoff(2007b), there are alternative ways recommender systems can present their results instead of, or in addition to, a top-N list (where only the highest recommended items are presented). For example, systems such as MovieLens[2] show all items with a number of stars signifying how much the user is predicted to like each item. Showing all items may improve users' sense of control, transparency and trust. Similarly, recommender systems that use critiquing (e.g., Reilly et al 2004; McCarthy et al 2005a) may show items with positive and less positive feature values, such as cameras that are reasonably priced but have low resolution. Group recommender systems (e.g., Ardissono et al 2003; Masthoff 2004) may present items that are disliked by a user, even when presented as a top-N list, because they are liked by other group members. In all of these cases, an effective recommender system may need to highlight relevant positive *and* negative information about items.

We consider a particular type of explanations, namely feature-based explanations. Feature-based explanations will not be able to mention all item features. Firstly, there are likely space restrictions, for example when a user is surveying a list of recommendations or using a portable device such as a mobile phone. Secondly, mentioning many item features would be detrimental to the speed with which users can make decisions (i.e. efficiency). To provide good decision support, the selection of features may need to be personalized,

---

[2] `http://www.movielens.org/`, retrieved May 2011

as users may differ in terms of which features they find important, and have individual tastes with regard to these features. For example, people may differ in the degree to which the author is a decisive factor for reading or enjoying a book. A study by Herlocker et al (2000) on explanations found a strong *persuasive* effect for an explanation referring to a particular movie feature, namely "favorite actor or actress". This feature may have been more important to some users than others, since a high variance in acceptance for this type of explanation was found. In the real-estate domain, Carenini and Moore (2001) found that user-tailored evaluative arguments (such as including *"the house has a good location"* for a user who cares a lot about location) increased users' likelihood to adopt a particular house compared to non-tailored arguments.

There has already been substantial research on generating personalized feature-based item descriptions, for example for the ILEX system in the museum domain (Dale 1998; Oberlander and Mellish 1998). Integrating recommendations with personalized item descriptions has also been proposed, for example for the INTRIGUE system in the tourism domain (Ardissono et al 2003). Billsus and Pazzani (1999) also looked at tailored explanations (including some feature-based ones), for a news recommender. The novelty of the work presented in this paper is in the analysis of the impact of these kinds of explanations on the user, in particular on effectiveness.

While similar, our work also differs from previous studies on the impact of explanations which primarily considered the *persuasive* power of arguments and explanations, but did not study effectiveness (Carenini and Moore 2001; Herlocker et al 2000). Arguably, Carenini and Moore (2001) varied the polarity (i.e. good vs. bad) of the evaluative arguments, but given the domain (real-estate) it was difficult for them to consider the final valuation of the item, i.e. whether the user would really like the house once they bought it. Others (e.g. Bilgic and Mooney 2005) evaluated effectiveness but did not consider the role of personalization. We will investigate how personalization of item features can affect explanation effectiveness and user satisfaction.

Using a user-centered approach, we have conducted user studies to elicit which features users use to make decisions about whether or not to watch movies, or buy digital cameras. We then used the elicited item features in a prototype natural language generation system, using commercial meta-data, to dynamically generate explanations. We inquired whether personalization helps increase effectiveness. Sections 4 and 5 describe two experiments, in the movie and camera domains, using approximated effectiveness (based on reading online reviews). To check the validity of these results, Section 6 presents a study in the movie domain in which participants saw the movies in question. Section 7 concludes with comments on the effects of personalization in both domains and whether it makes sense to explain at all. It also reflects on lessons learned about measuring effectiveness, and presents future work inspired by the limitations of the work presented.

Table 2 The goals for which explanations in recommender systems have been evaluated. System names are mentioned if available. Works that have no clear explanation goal stated, or have not *evaluated* the system on the stated goal, are omitted. Note that while a system may have been evaluated for several goals, it may not have achieved all of them. Also, for completeness a distinction is made between multiple studies using the same system.

SYSTEM: type of items	Transp.	Scrutab.	Trust	Effectiveness	Persuasiveness	Efficiency	Satisf.
Internet providers (Felfernig et al 2007)			X		X		X
Digital cameras, notebook computers (Pu and Chen 2006)			X				
Digital cameras, notebook computers (Pu and Chen 2007)			X	X			
Music (Sinha and Swearingen 2002)			X				
Movies (Tintarev and Masthoff 2008b)				X	X		X
ADAPTIVE PLACE ADVISOR: restaurants (Thompson et al 2004)				X		X	
ACORN: movies (Wärnestål 2005)							X
CHIP: artworks (Cramer et al 2008a)	X		X	X			
CHIP: artworks (Cramer et al 2008b)	X		X				X
FRINGE: people (Guy et al 2009a)				X	X		X
iSUGGEST-USABILITY: music (Hingston 2006)	X			X			
LIBRA: books (Bilgic and Mooney 2005)				X			
MOVIELENS: movies (Herlocker et al 2000)					X		X
MOVIEXPLAIN: movies (Symeonidis et al 2008)				X			X
myCAMERAADVISOR: cameras (Wang and Benbasat 2007)			X				
QWIKSHOP: cameras (McCarthy et al 2005a)				X		X	
SASY: holidays (Czarkowski 2006)	X	X					X
TAGSPLANATIONS: movies (Vig et al 2009)	X			X			
Social software items (Guy et al 2009b)				X	X		

2 Effectiveness

Getting the metric right is a key part of any evaluation (Tintarev and Masthoff 2009). Sometimes when adaptive systems are evaluated, there is a shortage of information about what exactly they are being evaluated on (McNee et al 2006b). This section discusses related work in evaluating the effectiveness of explanations, the metric that will be used in the studies in this paper, and the potential impact of recommendation domain on effectiveness.

2.1 Related work on effectiveness and metrics

As mentioned above, the focus of this paper is on explanations which aim to help users make qualified decisions, i.e. effective explanations. Effectiveness is by definition dependent on the accuracy of the recommendation algorithm, i.e. it is hard for users to make correct decisions if the recommendations are poor. However, an effective explanation may help the user evaluate the quality of suggested items according to their own preferences. This increases the likelihood that users discard irrelevant options while helping them to recognize good ones. For example, a book recommender system with effective explanations helps users to buy books they actually enjoy reading. Previous work emphasizes the importance of measuring the ability of a system to assist the user in making accurate decisions about recommendations, and compared different explanation types for effectiveness (Bilgic and Mooney 2005). Effective explanations could also serve the purpose of introducing a new domain, or the range of products, to a novice user, thereby helping them to understand the full range of options (Felfernig et al 2007; Pu and Chen 2006).

Several studies that seem to investigate effectiveness have in fact investigated something subtly different:

- For their conversational recommender system, Thompson et al (2004) calculated the percentage of conversations in which the first item presented was acceptable to the user (hit-rate). They also considered the proportion of features about which the system asked but the user did not care (rejection-rate). This muddles the distinction between effectiveness on the one hand and efficiency and satisfaction on the other. Thompson et al's metrics do not really measure how good a decision the system helps the user to make, but rather how fast the user can make a decision, and what they may (dis)like about the interaction.
- Cramer et al (2008a) tested the same system with and without an explanation facility, and compared the system's perceived and actual competence. Actual competence was established by comparing the item features used by the recommender system, with the item features mentioned by users when asked why they had chosen their top items. So, actual competence

focused on the quality of the user modeling rather than on whether the explanations helped users to make good decisions. Similarly, perceived effectiveness focused on whether the *recommendations* were good rather than on whether the explanations helped users to make good decisions.

– Symeonidis et al (2008) compared three justification styles. Users rated which justification style they preferred, which seems to measure satisfaction. Additionally, they used a metric called 'coverage ratio', to measure the quality of the justifications objectively. For this, they measured how well the item features revealed in the justifications covered the relevant item features in the user's profile. So, similarly to the work by Cramer et al (2008a), they measure how good the user modeling underlying the justifications is rather than whether the explanations helped users to make good decisions.

Several metrics for measuring effectiveness have been proposed:

– *Perceived effectiveness before consumption.* Vig et al (2009) let participants judge the effectiveness of explanations by rating how much they agreed with statements such as *"This explanation helps me determine how well I will like this movie."*. Using a similar metric, Hingston (2006) studied the perceived effectiveness of explanations for a number of interfaces and recommendation algorithms. Participants were asked how useful (and understandable) they perceived the explanations to be, and to rank explanations in order of usefulness.

– *Perceived effectiveness after consumption.* A variation of the metric used by Vig et al (2009) would be for participants to rate how much they agree with statements such as *"This explanation helped me determine how well I would like this item."*.

– *Success rate in finding the best item.* Chen and Pu (2007) used a decision quality metric from marketing (Häubl and Trifts 2000) which measures the fraction of participants that switched their choice to another option after viewing all of the items. This metric considers how effectively the system supports users to find the single *best* possible item (rather than "good enough items" as above). Participants interacted with the system until they found the item they would buy. Next, they viewed all items and were able to change their choice. Effectiveness was measured by the fraction of participants who changed their choice (so, a lower fraction meant better effectiveness).

– *Acceptance of items known to the user.* Guy et al (2009a) counted the number of accepted recommendations. In their domain, this meant the number of recommended people that were added to a social network. As a general metric, this seems to measure persuasiveness rather than effectiveness. However, if the recommended items (people in their domain) are already known to the user – as was the case in their study – then the fraction of accepted recommendations may indeed measure effectiveness.

– *Use of the explanations.* McCarthy et al (2005a) investigated compound critique-based explanations that make users more aware of the items avail-

able beyond the currently suggested item (e.g. other cameras with more memory but heavier). They measured among other things how often users selected compound critiques, regarding an explanation as more effective if the critiques were used more often. This measured whether the explanations helped users in making a choice, but not whether this actually resulted in users being happy with the item chosen after trying it. They also measured how satisfied users were with the camera they decided to purchase. However, this was done without the user actually trying the camera, and was merely done to ensure that the users who made little use of the compound critiques were not simply more difficult to satisfy.

– *Similarity between liking items before and after consumption.* Bilgic and Mooney (2005) used as metric the *absence of a difference* between the liking of the recommended item prior to, and after, consumption. The metric compares two item ratings: one after receiving an explanation, and a second after experiencing the item. If the opinion on the item did not change much, the explanation was considered effective.

These metrics differ on three dimensions:

– *Timing: Measuring effectiveness before or after consumption.* After-consumption effectiveness takes into account the user's opinion *after* experiencing the item. The metrics used by Vig et al (2009), Hingston (2006), and McCarthy et al (2005a) measure before-consumption effectiveness, while Bilgic and Mooney's metric approximates after-consumption effectiveness (they used an approximation of really experiencing the item). Chen and Pu's metric is somewhere in between, as participants could explore the whole catalog of items, presumably getting more information about the items, but not really fully experiencing them.

– *Items considered: Measuring effectiveness using only the top recommended item or all items.* Bilgic and Mooney's metric considers all items across the rating spectrum (or a representative set of such items). Cramer et al's metric considers a subset of items, namely the six items most preferred by users. Chen and Pu's metric considers the item users deemed *best* before having tried it.

– *Type of measurement: Measuring objective or perceived effectiveness.* Perceived effectiveness measures the user's perception of the system's effectiveness (e.g. through self-reporting), while objective effectiveness measures effectiveness directly (e.g. through comparing the difference between before and after ratings for items)[3]. Vig et al's and Hingston's metrics measure perceived effectiveness, while the metrics used by Bilgic and Mooney, Chen and Pu, and McCarthy et al measure objective effectiveness. The dimensions are not independent: objective effectiveness tends to be measured after consumption, though McCarthy et al (2005a) show that it is possible to measure it before consumption. Perceived effectiveness can be measured both before and after consumption.

---

[3] Parallels can be drawn with the distinction made between perceived and actual accuracy of recommendations as discussed in the usability framework proposed by Pu et al (2012).

2.2 Metric used in this paper

We will use Bilgic and Mooney's (2005) metric, as it measures objective effectiveness after consumption, and considers all items. We measure after-consumption effectiveness, as before-consumption effectiveness may overlap with persuasiveness. We measure objective effectiveness, as perceived effectiveness may overlap with satisfaction. We consider all items (across the rating spectrum, so including items the user may not like), as this provides a more comprehensive view of effectiveness. A metric considering a subset of items lacks granularity. For example, suppose that in a particular recommender system users tend to select item A when no explanations are present, and item B when explanations are present. Suppose that the item most suited was in fact item C. According to the metric used by Chen and Pu (2007), these systems have equal effectiveness, independent of how suitable items A and B were. So, even if item B were a lot more suitable – suggesting that explanations led to a decision a user was happier with – the effectiveness metric would not show this. It could be argued that this could be solved by creating a new metric which takes the difference between ratings for the best item before consumption and the best item after consumption (so, using an adapted version of Bilgic and Mooney's metric). However, this would still assume that only the best item(s) matter. As discussed in the introduction, in some circumstances, it is important that users get an accurate impression even of items less suitable to their tastes. This is for example the case in a group recommender system, where users may have to compromise given varying tastes in the group (Masthoff 2004).

For the reasons outlined, we measure effectiveness using the metric suggested by Bilgic and Mooney (2005):

1. **(Rating1)** The user rates the item on the basis of the explanation
2. The user tries the item
3. **(Rating2)** The user re-rates the item

Effectiveness can then be measured by the discrepancy between steps 1 and 3 ($Rating1 - Rating2$). According to this metric, an effective explanation is one which minimizes the gap between these two ratings. If an explanation helps users make good decisions, getting more (accurate and balanced) information or trying the product should not change their valuation of the product greatly. Bilgic and Mooney approximated step 2, by letting the users view reviews of the items (books) online. For other domains, other types of approximation may be possible, such as trailers for movies.

When the user rates several items, one possibility is to study the mean of the difference of these two ratings (Bilgic and Mooney 2005). In this case, 0 is the best possible mean. In a normal distribution, with as much over- as underestimation, this effectiveness metric will be close to 0, but this does not mean the explanations are effective. Bilgic and Mooney remedy this by *also* looking at the correlation between the first and second rating, with a high and significant correlation reflecting strong effectiveness.

Alternatively, one can take the mean of the *unsigned* difference (as well as studying the correlation) between the two ratings. We have included this additional analysis to our experiments, and survey the signed mean to see if there is a greater degree of over or underestimation.

2.3 Over- and underestimation and the role of domain

The metric introduced above does not distinguish between under- and overestimation. For example, a difference of 2 between an item's rating before consumption and after (overestimation) will have the same magnitude of contribution to the effectiveness score as a difference of -2 (underestimation). The question arises whether an explanation leading to an overestimation is indeed equally bad as one leading to a similarly big underestimation. Likewise, one wonders if the location on the scale matters. For example, should the impact on effectiveness of a pre-rating of 3 and a post-rating of 5 (on a scale from 1 to 5) really be equal to the impact of a pre-rating of 1 and a post-rating of 3? There may also be an impact of the recommender domain: the expected duration of experiencing the items and the expected impact of experiencing a disliked item is likely to differ between domains, and may well affect how effectiveness is perceived.

In economics, there has been a great deal of debate about classification of products into different categories. There is a distinction between experience goods, or goods that consumers learn about through experience, and "search goods" which they do not need to learn about through direct experience (Shapiro 1983). Similarly, there has been a distinction between sensory products and non-sensory products (Cho et al 2003). In Tintarev and Masthoff (2008a), we proposed an interpretation of these categories which distinguishes between products which are easy to evaluate objectively (e.g. light bulbs and cameras) and those which commonly require an experiential and subjective judgment (e.g. holidays and movies).

Another common categorization in economics involves investment or cost. Often this is a complex construct. For example, Murphy and Enis (1986) discuss *perceived* price in terms of the dimensions of risk and effort. This construct of risk includes financial risk but also psychological, physical, functional and social risk. The construct of effort considers purchase price, but also time that the purchase takes. Perceived price has also been defined in terms of non-monetary effort and degree of involvement (Cho et al 2003). Others narrow down the definition of cost to the objective measure of the purchase price of an item (Laband 1991). For simplicity, we will also use a definition of investment which only considers purchase price (for example, light bulbs and movies are low-investment, while holidays and cameras are high-investment).

In our previous work, we found that users considered overestimation to be less helpful than underestimation (Tintarev and Masthoff 2008a). We also found that the negative impact of a discrepancy between pre- and post-ratings on perceived effectiveness was significantly higher for high investment domains

than for low investment ones. In particular overestimation had a significantly higher negative impact on perceived effectiveness in high investment domains.

There was also a trend towards over- and underestimation having a higher negative impact on perceived effectiveness in objective compared to subjective domains. This trend was confirmed by participant comments (e.g. a wrong suggestion about subjective evaluations of products, such as for movies or holidays, should not determine a severe bad judgment of the website.). We also found that the effect of prediction errors on perceived effectiveness varied depending on where on the scale the prediction error occurred. Gaps on the negative end of the scale (e.g. a pre-rating of 1 and a post-rating of 3 and vice versa) had a higher negative impact on perceived effectiveness than gaps on the positive end (e.g. $3 \leftrightarrow 5$), and gaps which cross over between the positive and negative ends of the scale (e.g., $2 \leftrightarrow 4$) for both over- and underestimation. Gaps which cross over in turn were perceived less effective than positive gaps.

In this paper, we will keep the metric as defined in Section 2.2. Because of the findings in this subsection, we will consider over- and underestimations in the experiments, and we will present studies in two different domains: a low investment and subjective domain (movies) and a high investment and objective domain (cameras).

3 Experimental setup

The four experiments outlined in the next sections share a common experimental setup, following that of our study in Tintarev and Masthoff (2008b). This section discusses this setup. Small divergences will be highlighted in the experiments when they occur.

3.1 Methodology

3.1.1 Experimental design

Participants were randomly allocated to one of three conditions in a between-subjects design. The conditions differed in the type of explanations provided, in particular on whether the explanations were personalized and feature-based.

We used the layered evaluation framework, which advocates that adaptation needs to be decomposed and assessed in layers in order to be evaluated effectively (Paramythis et al 2010). Instead of evaluating a full recommender system, we focused the evaluation on the combination of the Decide Upon Adaptation (DA), and Apply Adaptation (AA) layers. So, we wanted to know whether the decision to personalize the explanations (DA) and the way this was done (AA) would indeed lead to increased effectiveness and satisfaction. Following the advice in Paramythis et al (2010), we needed to guarantee accurate input to the layers under evaluation, so an accurate user model. Participants

therefore provided the user model directly, using the features that were available for a commercial recommender.

We also did not use a real recommender algorithm to provide recommendations. There are two reasons for this. Firstly, we wanted to evaluate the effectiveness of the explanations, rather than the recommendation algorithm. Secondly, as discussed above, we were interested in the effectiveness of explanations across the spectrum: not just how well explanations would support decisions related to good items, but also for weaker items. So, instead of using a recommender system to select the items, we used a random selection from a fixed set of items. Isolating explanations is in accordance with the 'dicing' approach, as advocated in Masthoff (2002). While layered evaluation 'slices' the adaptation process into its components (e.g. Decide Upon Adaptation and Apply Adaptation), dicing isolates the various functionalities that are being adapted (e.g. recommendation algorithm and explanations).

3.1.2 Independent variables

Type of explanation. Three types of explanations were used:

1. **Baseline.** The explanation was neither personalized, nor described item features. E.g. *"This movie is one of the top 100 movies in the Internet Movie Database."* In all studies, baseline explanations were chosen so that they could be automatically retrieved and were comparable to common practice on commercial sites.
2. **Non-personalized, feature-based.** The explanation described item features, but the features were not tailored to the user. E.g. *"This movie belongs to the genre(s): Drama. Kasi Lemmons directed this movie."*
3. **Personalized, feature-based.** The explanation described item features, and tailored them to the user's interests. E.g. *"Although this movie does not belong to any of your preferred genres(s), it belongs to the genre(s): Documentary. This movie stars Ben Kingsley, Ralph Fiennes and Liam Neeson your favorite actor(s)."* For this user, the most important feature is leading actors and these actors were mentioned as favorites.

Details of the explanations used will be given separately for each study.

3.1.3 Dependent variables

Effectiveness. Effectiveness was measured using the metric described in Section 2.2. Participants rated the items twice (e.g., *"How much do you think you would like this camera?"*), using a 7-point Likert scale (from "not at all" to "a lot"), both before and after experiencing them. The metric considers how the user's valuation of the items changes.

Satisfaction. Satisfaction was measured through rating the explanations (*"How good do you think this explanation is?"*) on a 7-point Likert scale (from really bad to really good).

3.1.4 Procedure

The procedure consisted of the following steps:

1. Participants provided background information about themselves, such as demographic information.
2. Participants rated the importance of different product features and entered their preferences for each feature, resulting in a simple user model. For example, for experiments 1, 2 and 4, participants rated the importance of the following features: actors, director, MPAA rating (suitable for children, adult content etc), genre and average rating by other users. They also selected their favorite actors and directors and indicated which genres they were in the mood for and for which ones they were not.
3. A number of items were selected at random from a pre-selected set. If participants were already familiar with an item, they could request another one by clicking on a button (*e.g., "I might know this movie, please skip to another one"*). Participants evaluated the items and their explanations. For each item in turn:
   (a) Participants were shown the item and explanation, and rated:
      – *How much they would like this item.*
      – *How good the explanation was.*
      They could opt out by saying they had "no opinion", and could give qualitative comments to justify their response.
   (b) Participants tried the item. In the first three experiments this step was approximated by participants reading user and expert reviews on Amazon.com; care was taken to differentiate between our explanation facility and Amazon.
   (c) They re-rated the item, and the explanation.

3.1.5 Statistical Analysis

To analyze the results for effectiveness and satisfaction, mixed linear effects models were fitted, with participant as random factor, trial as repeated factor (as each participant evaluated multiple items), AR(1) as co-variance structure[4], type of explanation as fixed factor, and absolute effectiveness and satisfaction as dependent variables respectively. We used the SPSS v19 procedure mixed models linear, with Bonferroni correction for the pair-wise comparisons between explanation types.

For the analysis of effectiveness, opt-outs ("no opinion") were treated as missing values. However, opt-outs are not arbitrary missing data; they indicate that participants did not have enough information to decide how much they would like the item. One could argue that treating opt-outs as missing values biases the data towards better effectiveness, and that this may bias our results when the opt-out rates differ per condition. This is not always a problem: if a

---

[4] The statistical inferences about the explanation types remained the same regardless of the particular variance structure selected.

condition with better effectiveness also has a lower opt-out rate, then we can assume that the better effectiveness is a real result. However, if a condition with better effectiveness has a higher opt-out rate, then the result is less solid.

A solution may be to replace the missing values with the worst possible effectiveness value (6). However, the assumption that opt-outs result in artificially better effectiveness is not necessarily correct. If participants had been forced to make a decision instead of being allowed to opt-out, they are likely to have chosen the neutral middle of the scale rating. This may in fact have led to *better* effectiveness, as middle of the scale ratings would result in a worst possible effectiveness score of 3 rather than 6 (see Section 7.3.1 for more on middle of the scale ratings). So, replacing missing values with 6 is not representative of participant behavior, and artificially diminishes the effectiveness of explanations with higher opt-out rates. This may lead to such explanations looking worse than "misleading" explanations (where participants think they can judge the item, but judge it completely wrongly). Replacing missing values with 3 is not ideal either. Participants opting out means that they had insufficient information to decide, which is a clear indication of poor effectiveness. Replacing missing values with 3 may hide this, and make explanations with high opt-out rates look better than they deserve. Therefore, we have decided not to replace missing values, but to analyze opt-out rates separately and discuss their potential implications.

To analyze the opt-out rates, generalized linear mixed effect models were fitted, with participant as random factor, trial as repeated factor, AR(1) as co-variance structure, type of explanation as fixed factor, and opt-out[5] as dependent variable with a binomial distribution and logit link function. We used the SPSS v19 procedure GLMM, with sequential Bonferroni correction for the pair-wise comparisons between explanation types.

3.2 Hypotheses

We hypothesized that:

- **H1:** Personalized feature-based explanations will be more effective than non-personalized feature-based and baseline explanations.
- **H2:** Users will be more satisfied with personalized feature-based explanations compared to non-personalized feature-based and baseline explanations.

---

[5] Opt-outs were represented as a binary variable: "0" for opt-outs, and "1" when a rating was given.

Table 3 Participants in the experiments who were included in the analyses

Experiment	Total N	Age Mean (StD)	Gender
1. MoviesI	46	26.54 (8.13)	25 male, 21 female
2. MoviesII	33	24.58 (6.58)	26 male, 7 female
3. Cameras	47	24.17 (5.85)	31 male, 16 female
4. Final Evaluation	48	26.17 (7.24)	21 male, 27 female

3.3 Participants

Participants were recruited from university staff and students. They received a gift voucher (£5-10 depending on the study duration) to compensate them for their time. The studies took place in a lab under controlled conditions. Table 3 shows the number of participants and demographics for the four studies discussed in this paper. These are the participants whose data was used in the analyses. There were some additional participants that were excluded from the analyses. The number of participants excluded and the reasons for their exclusion will be discussed in each study.

3.4 Motivation of the choice of explanations

The explanations used in these experiments are short and simple. There are three good reasons for this. Firstly, brevity is important in a context where the user has to review many possible options. Secondly, the features that are currently available in existing commercial services are limited in both diversity and depth[6]. Thirdly, algorithms which consider simple features such as actor and director names already exist (Symeonidis et al 2008).

Thus, the questions we are investigating are if it makes sense for the developers of explanations in recommender systems to change their algorithms to explain by using item features, and if it makes sense to personalize the features to present. Even a simplistic change can be a large investment, and so an experiment of this type saves a considerable potential cost. As this sort of change would not make much sense unless it made a difference, let us now see if this is the case.

4 Experiments 1 and 2: Approximated effectiveness for movies

The aims of the initial experiments in the movie domain was to see if using movie features (e.g. lead actors), and personalization in explanations could affect their effectiveness and user satisfaction. We chose to use item features that realistically could be extracted from a real world system such as Amazon Web Services as these were freely available via an API for a number of domains,

---

[6] In these experiments we have chosen to use Amazon Web Services as a representative example, although similar limitations are likely to occur with other commercial services.

while considering the features extracted from our user studies (e.g. Tintarev and Masthoff 2007a). This resulted in the features: *genre, cast, director, MPAA rating (e.g. rated R) and average rating.* Our user studies also suggested that genre information is important to most if not all users, so both feature based conditions contain a sentence regarding the genre.

Experiment 1 has been reported in full in Tintarev and Masthoff (2008b), and is therefore only summarized here. Since there was no significant difference between conditions w.r.t. effectiveness (and the trend was even for non-personalized explanations to be more effective), we considered potentially confounding factors and repeated the experiment with some modifications. This section shows the results of this new experiment (Experiment 2) in the context of the results of Experiment 1.

4.1 Example explanations for Experiment 2

Table 4 summarizes the types of explanations that were given in the three conditions in Experiment 2. Explanations could contain negative as well as positive information, and the genre information could vary in terms of polarity: positive if the movie belongs to only preferred genres, negative if it belonged to any genre the participant did not want to see, and neutral if it contained neither preferred nor 'disliked' genres.[7].
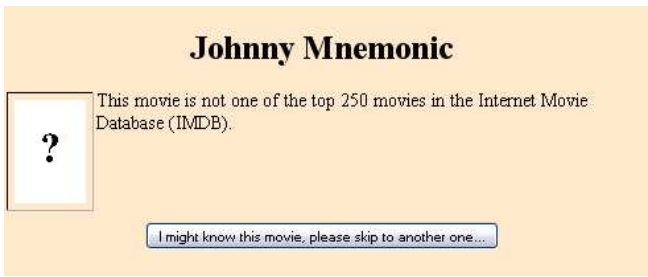
4.2 Differences between Experiments 1 and 2

- *Use of cover image.* In Experiment 1, the movie cover was shown in all conditions, and this may have provided unintentional additional information about the movie. In Experiment 2, the cover was not shown.
- *Distinction between personalized and non-personalized conditions.* In Experiment 1, the information about genre would always say whether the genre was a preferred or non-preferred genre. This can be seen as a form of personalization. In Experiment 2, this was only mentioned in the personalized condition.[8]
- *Genre information.* We ensured that the information about genres was more detailed and complete in Experiment 2, as participants in Experiment 1 complained that genre information automatically retrieved from Amazon was occasionally incorrect and incomplete[9]. The genre information was annotated by hand, and the generated explanations describe all the genres a movie belongs to.

---

[7] While not included in the results here, polarity was found to significantly correlate with users' initial ratings of movies, but there was no significant correlation between polarity and effectiveness (Tintarev and Masthoff 2008b).

[8] While this was a flaw, the personalized condition in Experiment 1 was still more personalized than the other feature-based condition.

[9] This will have affected both feature-based conditions to the same extent, so does not invalidate the results regarding the effect of personalization.

Table 4 Experiment 2: Example explanations per condition in the movie domain. The general interface was similar in appearance in all three conditions, as exemplified by the screenshot for the baseline condition. Both feature-based explanations describe the movie genres, but the personalized explanation also relates them to the user's preferences. The personalized explanation also describes the most important feature for this user (actors), while the non-personalized explanation used a random feature (average rating).

Non-personalized:	*"This movie belongs to the genre(s): Action & Adventure and Comedy. On average other users rated this movie 4/5.0"*
Personalized:	*"Unfortunately, this movie belongs to at least one genre you do not want to see: Action & Adventure. It also belongs to the genre(s): Comedy. This movie stars Jo Marr and Robert Redford."*
Baseline:	

- *Number of trials.* Experiment 1 typically took around 45 minutes to complete, which may have led to participant fatigue. For this reason, we reduced the number of trials from ten in Experiment 1 to three in Experiment 2.
- *Baseline condition* The baseline condition in Experiment 2 mentions movies in the top 250 (rather than top 100) in the Internet Movie Database (IMDB).

4.3 Materials

In Experiment 2, the 85 movies were distributed evenly among 17 genres. As a movie belongs to multiple genres, they were balanced according to the main genre. In this experiment, movies were also selected for having a high degree of variation of rating. High variation is more likely to lead to polarized views leading to an even distribution of initial ratings of movies. We used the measure of rating variation (entropy) described in Rashid et al (2002), based on the MovieLens 100k ratings data set[10], which considers both variation and number of ratings for that movie (to avoid very obscure movies). Fourteen of the movies were in the top 250 of IMDB.

---

[10] http://www.grouplens.org/node/12#attachments

4.4 Results

In Experiment 2, seven participants (out of twelve) in the baseline condition
had to be removed from analysis for clicking through the experiment (finish-
ing in under 90 seconds) or dropping out altogether. We also found that the
remaining participants in the baseline condition opted out of giving Movie
Before ratings in 56% of cases. There was a significant effect of condition on
opt-out rates (F(2,94)=17.31, p<0.001), with the baseline condition having
significantly more opt-outs than the other conditions (sequential Bonferroni
corrected, p<0.05). This suggests that explanations such as our baseline with-
out cover images could damage user satisfaction considerably. Because of these
problems, the baseline condition is not included in the analysis of Experiment
2.

Are personalized explanations more effective (H1)? As in Experiment 1, ex-
planations in the non-personalized feature-based condition appear to be most
effective in Experiment 2 (see Table 5), but this difference was not signifi-
cant (F(1,57.61)=2.23, p=.14). It is also in this condition that participants
opted out the least (4.3% compared to 15.2% in the personalized condition,
sequential Bonferroni corrected, p<0.05).

Table 5 Experiments 1 and 2: Means (StD) of the two movie ratings (excluding opt-outs)
and effectiveness per condition per experiment. "Before" and "After" denote the two movie
ratings before and after viewing Amazon reviews. Effectiveness is better the closer it is to
zero.

	Condition	Movie Before	Movie After	Effectiveness (absolute)	Effectiveness (signed)
Exp.1	Baseline	3.45 (1.26)	4.11 (1.85)	1.38 (1.20)	-0.69 (1.69)
	Non-personalized	3.85 (1.87)	4.43 (2.02)	1.14 (1.30)	-0.57 (1.64)
	Personalized	3.61 (1.65)	4.37 (1.93)	1.40 (1.20)	-0.77 (1.68)
Exp.2	Non-personalized	3.84 (1.95)	3.93 (1.95)	0.96 (0.81)	-0.09 (1.25)
	Personalized	3.75 (2.05)	4.00 (1.87)	1.33 (1.27)	-0.25 (1.85)

Are users more satisfied with personalized explanations (H2)? Table 6 shows
that the Explanation After ratings are mostly higher than the Explanation
Before ratings. This may be due to participants confounding our explanations
with the Amazon reviews when rating Explanation After. Since participants'
comments corroborate this, we did not include Explanation After in anal-
yses for the first three experiments (where Amazon reviews are used). The
mean rating for Explanation Before is low overall. In Experiment 1, there was
a significant difference in ratings between the conditions (F(2,449.36)=9.95,
p<0.001), with participants rating the first explanation significantly high-
est in the personalized condition (Bonferroni corrected, p<0.01). In Exper-
iment 2, the mean satisfaction for personalized explanations is again highest
(Explanation Before, see Table 6), though the difference between the non-
personalized and personalized conditions is not statistically significant this

time $(F=(1,75.15)=1.11, p=0.3)$. This suggests that while the personalized explanations may not help users make better decisions, users may still be more satisfied. This was confirmed by the qualitative comments given by participants.

Table 6 Experiments 1 and 2: Means (StD) of the two explanation ratings, excluding opt-outs, (on a scale from 1-7, 1=really bad, 7=really good) per condition per experiment. "Before" and "After" denote the two explanation ratings before and after viewing Amazon reviews.

	Condition	Explanation Before	Explanation After
Experiment 1	Baseline	2.38 (1.54)	2.85 (1.85)
	Non-personalized	2.50 (1.62)	2.66 (1.89)
	Personalized	3.09 (1.70)	3.15 (1.99)
Experiment 2	Non-personalized	2.72 (1.68)	2.83 (1.74)
	Personalized	3.31 (1.55)	2.97 (1.33)

General comments In all conditions, there was more underestimation than overestimation. In light of this underestimation we reconsider the fact that movie ratings, and Amazon reviews, may lean toward positive ratings. If Amazon reviews are overly positive, this may have affected our results.

5 Experiment 3: Approximated effectiveness for cameras

To investigate whether the results were due to the domain, or were more general, we repeated the experiment in another domain. For this purpose we chose a domain which was more objective and higher investment: digital cameras. Our intuition was that the movie domain suffers from being subjective in nature. So while it is possible to talk about the user's favorite actor starring in a movie, the actor's performance may be deemed as both good and bad depending on the user. Nor is an actor's performance likely to be consistent across their career, deeming this feature (most commonly selected by our participants) a poor indicator for decision support. We expected this effect to be smaller in a more objective domain such as digital cameras.

We have also seen that participants may be less forgiving of overestimation (persuasion) in high investment and (relatively) objective domains (Tintarev and Masthoff 2008a). We are interested to see how additional and personalized information in explanations influences users in the camera domain: whether this impacts effectiveness and user satisfaction.

5.1 Modifications

To perform the experiment in a second domain, several changes were needed.

5.1.1 Procedure

– Participants are less likely to be consumers of cameras than movies. To exclude participants that would never use or buy a camera, they indicated their photography expertise and likelihood of purchasing a camera.
– People buy fewer cameras than they see movies. This means that participants are unlikely to be familiar with a particular camera, especially because explanations were accompanied with a generic image of a camera (see below) and the name of the camera was not used. For this reason, we did not need a "I might know this item" button as in the movie experiments.
– Participants evaluated 4 cameras and explanations.
– No polarity was applied to the explanations. In the previous experiments, a movie could belong to a user's preferred genres (positive polarity) as well as disliked genres (negative polarity). Even though polarity may not impede effectiveness[11], in this experiment we chose to control for polarity as well.

5.1.2 Explanations used

Table 7 provides an overview of the explanations given in the three conditions described below:

– *Non-personalized condition.* This explanation describes the three most commonly selected camera features (as described in Section 5.1.4), which were camera type, brand, and price. Three features are mentioned to make these explanations comparable in length to the explanations in the movie domain which mentioned genre as well as a feature.
– *Personalized condition.* The explanation describes the three camera features that are **most** important to this participant. For example, if features 'price', 'brand' and 'zoom' are most important the explanation may be: *"This camera costs 679.95£. This camera is a Nikon. It has an optical zoom of 11.0x."*
– *Baseline condition.* There is no equivalent to IMDB for cameras, so the baseline was changed to a bar chart which summarizes review ratings of the camera categorized into good, ok, and bad (see Table 7). This sort of bar chart exists on the Amazon website, and is similar to the explanations given on several commercial sites. It is similar to the bar chart used in (Herlocker et al 2000), but considers all reviews rather than similar users only.

---

[11] In the previous experiments, polarity was found to significantly correlate with users' initial ratings of movies, but there was no significant correlation between polarity and effectiveness (Tintarev and Masthoff 2008b).

Table 7 Experiment 3: example explanations for three conditions in the camera domain. The general interface was similar in appearance in all three conditions, as exemplified by the screenshot for the baseline condition. In the personalized condition, the explanation describes the three camera features that are **most** important to this participant, but in the non-personalized the same features (type, brand and price) are always used. While simple, the baseline is similar (but not identical) to information supplied by Amazon and used in the study by Herlocker et al 2000.

Non-personalized:	e.g. *"This camera costs 179.0£. This camera is a Panasonic. This camera is a 'point and shoot camera'."*
Personalized:	e.g. *"This camera costs 679.95£. This camera is a Nikon. It has an optical zoom of 11.0x."*
Baseline:	

5.1.3 Hypotheses

The changed baseline results in a third hypothesis. Camera reviews are strongly biased toward positive ratings: there are more positive ratings than negative and neutral. Bilgic and Mooney (2005) found that a positively biased bar chart is likely to lead to overestimation of items. For this reason we also hypothesize that:

– H3: Users are more likely to overestimate their rating of the camera in the baseline condition compared to the two feature-based explanations (persuasion).

5.1.4 Materials

We wanted to elicit which features are generally considered important when purchasing a camera. As a starting point, we surveyed which features existing recommender systems in the camera domain have used (Chen and Pu 2007; Felfernig et al 2008; McCarthy et al 2004, 2005b). We shortlisted the following features: brand, optical zoom, price, resolution, weight, memory and type of camera (SLR or point and shoot). From these memory was excluded as modern cameras usually have external memory that can be added on. The re-

Table 8 Experiment 3: Range of features over the 22 cameras used in this experiment.

Feature	Range	Mean (StD)	Mode
Price	£106-1695	448.40 (489.23)	225.73
Resolution	5-12 megapixels	9.45 (2.21)	10
Zoom	1-10 x	5.77 (4.80)	3
Weight	118-930 g	421.59 (286.86)	334
Camera 'type'	SLR (9), "point-and-shoot" (13)		
Brands	Panasonic (4), Nikon (4), Canon (4), Olympus (4), Fujifilm (3), Sony (3)		

Table 9 Experiment 3: Total number of ratings, and mean number of reviews per camera (StD), by category of rating.

Good		Ok		Bad	
Total n	Mean (StD)	Total n	Mean (StD)	Total n	Mean (StD)
326	16.09 (13.10)	19	0.86 (1.22)	25	1.32 (2.69)

maining six features are all readily available on Amazon's Web Service. Next, 11 members of university staff or members of the university photography club (1 female, 10 male; average age 44.67, range 29-62) rated the importance of these six features. The purpose of the questionnaire was twofold. Firstly, we wanted to know whether there are features that are commonly considered important. Secondly, we wanted to find out if there was a case for personalization, i.e. do different people find different features important. Overall, type of camera, brand, and price were found to be most important. However, this is not a complete consensus, people do rate the features differently. It is not the case that any one feature was rated highest by each participant.

Twenty-two cameras have been hand-picked from the Amazon website. Specifications for SLR cameras were defined by the lens that came with them per default. Table 8 summarizes the range for each of the features. It is also possible to select the cameras automatically via an API, but handpicking the items enabled us to control the range for each feature better.

As seen in Table 9, there were by far more good ratings (4's and 5's) than ok (3s) and bad (1s and 2s), which is a pre-existing bias for the cameras which had at least 3 reviews.

The explanation was accompanied with an identical image of a camera, with a semitransparent letter (A-D) superimposed to differentiate the four cameras.

5.2 Results

Five participants were removed from analysis: one for not completing the experiment, three for being "unlikely to buy a camera" and one for saying they "knew nothing about photography".

5.2.1 Enough to form an opinion?

As in the previous experiments we inquire if the short explanations are sufficient for users to form an opinion. Table 10 shows that in the baseline there was a large percentage (23.9%) of opt-outs for the first camera rating. There was a significant effect of condition on opt-out rates ($F(2,185)=18.30$, $p<0.001$). The baseline and non-personalized conditions had significantly more opt-outs than the personalized condition (sequential Bonferroni corrected, $p<0.01$).

Table 10 Experiment 3: Percentage of opt-outs for the two camera ratings and the two explanation ratings. "Before" and "After" denote the ratings before and after viewing Amazon reviews.

Condition	Camera Before	Camera After	Explanation Before	Explanation After
Baseline	23.9%	7.5%	6%	6%
Non-personalized	16.7%	8.3%	3.3%	0%
Personalized	1.6%	0%	3.2%	1.6%

Table 11 Experiment 3: Means (StD) of the two camera ratings (excluding opt-outs) and effectiveness per condition. "Before" and "After" denote the two camera ratings before and after viewing Amazon reviews.

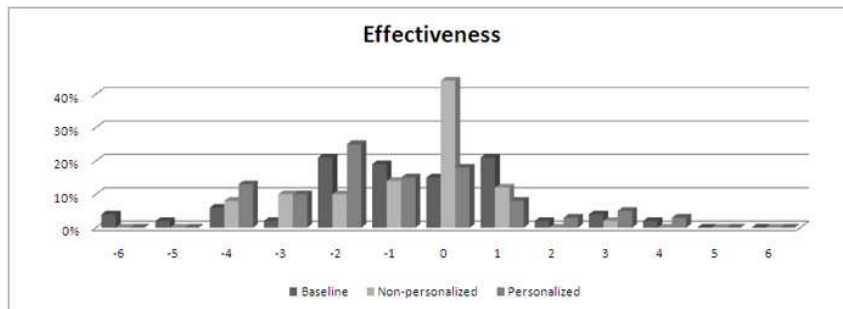| Condition	Camera Before	Camera After	Effectiveness (absolute value)	Effectiveness (signed value)
Baseline	3.94 (1.47)	4.75 (1.73)	1.77 (1.50)	-0.77 (2.20)
Non-personalized	3.88 (1.62)	4.78 (1.75)	1.14 (1.32)	-0.78 (1.57)
Personalized	3.83 (1.86)	4.95 (1.77)	1.88 (1.34)	-1.08 (2.05)

Fig. 1 Experiment 3: Distribution of (signed) effectiveness per condition for cameras (excluding opt-outs).

Camera Before

Camera After

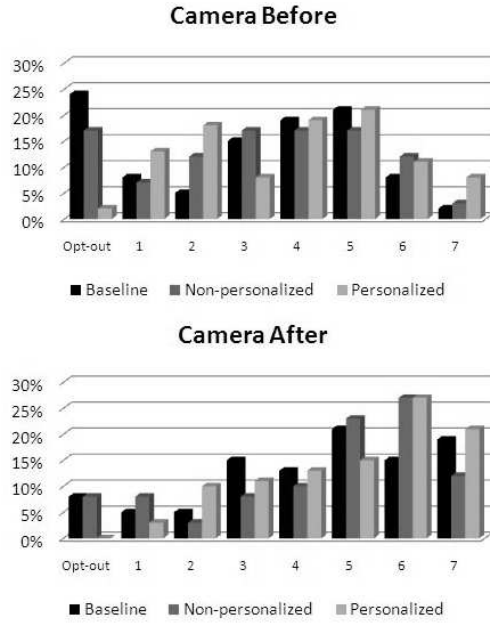**Fig. 2** Experiment 3: Camera ratings before and after, per condition.

5.2.2 Are personalized explanations more effective? (H1)

Table 11 shows the ratings of the cameras and effectiveness per condition. Effectiveness appears to be best (lowest value) in the non-personalized condition. Comparing between conditions we found a significant difference in (absolute values for) effectiveness (F(2,151.42)=4.83, $p < 0.01$). Post-hoc tests showed that effectiveness was significantly better in the non-personalized condition than the personalized condition (Bonferroni corrected, $p < 0.05$) and marginally better in the non-personalized condition than in the baseline condition (Bonferroni corrected, $p = 0.056$). That is, non-personalized explanations were more effective. Figure 1 shows that almost 45% of explanations in this condition lead to perfect effectiveness (i.e. Rating1 - Rating2 = 0). In other words there is no support for H1, personalized explanations are not most effective. The lower opt-out rates in the personalized condition indicate that that the personalized explanations did help participants form an opinion, though not necessarily an accurate one.

5.2.3 Do baseline explanations lead to more overestimation? (H3)

We also hypothesized that participants would be more likely to overestimate their ratings of cameras in the baseline condition. Firstly, in Table 10 we see that a large number of participants in this condition have opted out. In Table

11, we see however that with the opt-out ratings omitted, the initial ratings for cameras are comparable between the three conditions. Figure 2 also shows the distribution of these initial camera ratings per condition. Moreover, the signed value of effectiveness in Table 11 suggests a small underestimation in the baseline condition. These findings are contrary to our third hypothesis, H3.

It is surprising that the baseline has reasonably good effectiveness. The distribution of initial ratings in the baseline condition (Figure 2) suggests that users are less susceptible to persuasion than one might initially think. We return to this when discussing users' qualitative comments.

5.2.4 Are users more satisfied with personalized explanations? (H2)

We compared the users' ratings for the initial explanations (Explanation Before, see Table 12), and found a significant difference between conditions ($F(2,168.38)=6.04$, $p<0.01$). Post-hoc tests support H2; participants were significantly more satisfied with personalized explanations than non-personalized (Bonferroni corrected, $p<0.01$). There was, however, no significant different between the baseline and the personalized condition.

Table 12 Experiment 3: Means (StD) of the two explanation ratings, excluding opt-outs, (on a scale from 1-7, 1=really bad, 7=really good) per condition. "Before" and "After" denote the two explanation ratings before and after viewing Amazon reviews.

Condition	Explanation Before	Explanation After
Baseline	2.83 (1.44)	3.80 (1.87)
Non-personalized	2.38 (1.64)	2.87 (1.94)
Personalized	3.27 (1.27)	2.67 (1.56)

5.2.5 Distribution of features

Participants in the non-personalized condition were all shown the same three features: price, brand and type. We were interested in the distribution of features chosen in the personalized condition to see how it compared to these three features. The combination of features used in the non-personalized condition only received 8% of the votes, and there is a great deal of variability in the features participants found most important.

5.2.6 Limitations and qualitative comments

As we mentioned in our description of the three conditions, the bar chart we used is not identical to the bar chart used in (Herlocker et al 2000). This is also noted in user comments; *"...doesn't give you information about what kind of customers rated it (a complete newbie wanting to buy a 'point-and-shoot' will rate things differently than amateur buying a SLR)"*; *"No clear indication*

of the audience that's declared it 'good'. "

Participants reacted to the number of ratings and polarity of the baseline explanations, which may explain why the baseline performed surprisingly well yet again. Participants were not easily persuaded, and did not rate cameras in this condition consistently highly. For example, a bar chart using too few ratings was considered insufficient information: *"...since it only has the review from six people, it's hard to base my decision on just this explanation...."*; *"Too small a test group for a clear set of results"*.

The majority of reviews were positive even for cameras with many reviews, which by some participants was perceived as poor information as well: *"There are no other opinions except for the people's who are in favor of the camera. This is a poor display of statistics"*; *"everybody cannot possibly rate this good, there has to be some opposers."*. Explanations were taken more seriously when the distribution of ratings was more even; *"The ratings have a larger review base, with some dissenting into "ok", broader review"*.

There were also two small issues, but neither should have affected the comparison between conditions: some participants were influenced by the accompanying image (which was always the same) and some compared cameras with each other.

5.3 Summary of Experiment 3

Participants made better decisions in the non-personalized condition, but preferred the personalized explanations. This result is similar to what we found in the movie domain, although we did not expect to find this in an objective, high investment domain.

6 Experiment 4: True effectiveness for movies

So far, we have been looking at explanation effectiveness in two domains: movies and cameras. These experiments were limited by the approximation we used for evaluating the items: reading online reviews. In particular, the reviews used are likely to have been positively biased, and so this raises the question if the same results would be found if users actually tried the items. It is possible that non-personalized explanations cause an overestimation that correlates well with the positive bias of the reviews, but that these explanations are not in actual fact effective. In this experiment, participants experience the items. This experiment is conducted in the movie domain as it is easier, and less costly, to collect suitable materials. This also simplifies the users' valuation of items, as it is easier to let users watch movies than evaluate digital cameras.

6.1 Modifications

6.1.1 Procedure

Given the strong result for baseline explanations in the previous experiments in the movie domain, we wanted to know how much the title influenced users' ratings of a movie and so included an initial rating (Movie Title). So, the procedure is revised such that:

1. The user rates the item on the basis of the title only **(Movie Title)**
2. The user rates the item on the basis of the explanation **(Movie Before)**
3. The user tries the item
4. The user re-rates the item **(Movie After)**

   Other changes are:

   – In addition to the usual *"I might know this movie, please skip to another one"* button, for ethical reasons an *"I don't want to watch this movie"* button was also included so that participants would not be forced to watch a movie they did not want to see.
   – Participants were asked if they disliked the genres children, comedy or animation, as most of our movies were (for ethical reasons) from those genres. This resulted in three participants being omitted. Participants were not explicitly told about the selection criteria for movies.
   – Participants were asked to watch as many short movies as was feasible within the duration of an hour, but no more than 3.
   – For ethical reasons, participants had full control over the movies (e.g. they could press the stop and pause buttons), and could terminate the experiment at any time.

6.1.2 Experimental design

The baseline condition was modified to adjust for the fact that we were using short movies (see Section 6.1.4): e.g. *"This movie is (not) in the top 50 short movies on IMDB (the Internet Movie Database)."*. The other conditions were the same as in the second experiment.

6.1.3 Hypotheses

Hypotheses H1 and H2 are as before, while H3 regards the camera experiment only. The addition of an initial movie rating leads to an additional hypothesis:

   – **H4:** Users will able to form an opinion more often after the first explanation (Movie Before) than after just seeing the title (Movie Title).

We believe that the title is not enough information for a user to form an opinion, while the additional information supplied in explanations will help to make a decision.

Table 13 Experiment 4: Short movies used, with genres and duration.

Title	Genres	Duration (minutes)
Wrong Trousers	Animation, Children, Comedy, Crime	29
Close Shave	Animation, Children, Comedy, Crime	30
Grand Day Out	Animation, Children, Comedy, Crime	23
Feed the Kitty	Animation, Children, Comedy	7
Vincent	Animation, Children, Fantasy	6
For the Birds	Animation, Children, Comedy	3
Mickey's Trailer	Animation, Adventure, Children, Comedy	8
The Rocks	Animation, Comedy, Fantasy	8
Mr. Bean's Christmas	Comedy	4
Rabbit Seasoning	Animation, Children, Comedy	7
Hedgehog in the Fog	Animation, Children, Drama, Fantasy, Mystery	10
Kiwi	Animation, Action, Adventure, Children, Comedy, Drama, Thriller	3
Strange to Meet You	Comedy, Drama	6
Jack Shows Meg His Tesla Coil	Comedy, Drama	7
Somewhere in California	Comedy, Drama	11

6.1.4 Materials

To decrease the duration of the experiment, short movies were chosen over full length features. Movies were selected as to be non-offensive. Fifteen movies were selected, out of which eleven are in the top 50 short movies in IMDB. The durations of the movies vary from 3-30 minutes (mean=11.73, StD=10.36). Table 13 summarizes the selected movies. The majority of movies belong to the genres comedy, animation, and children. Most of the movies have an international certification rating, and some have actors (e.g. Rowan Atkinson) or directors (e.g. Tim Burton) that are likely to be known. The movies also vary w.r.t. other factors, for example some are in foreign languages (English subtitles), the animations differ in style, and three of the movies are black and white.

The selection of non-offensive movies could result in users' ratings of movies being less well distributed. However, for an interesting analysis, it suffices that the distribution differs sufficiently from the mid-point, even if it does not make full use of the scale. This was confirmed in two pilot sessions.

6.2 Results

In this section we survey the effectiveness of, and satisfaction with, explanations when participants were able to trial the items (watch short movies).

6.2.1 How do titles compare to explanations? (H4)

Opt-outs. Looking at the percentages of opt-outs in Table 14 we see that, on average, participants opted out 35.6% of the time for Movie Title, compared to 15.9% after receiving an explanation (Movie Before). In Figure 3 we see the change in opt-outs for the three movie ratings across the conditions, and the noteworthy decrease of opt-outs from Movie Title to Movie Before in all three conditions. This suggests that explanations do help users to make decisions.

We investigated whether the difference in opt-outs between Movie Title and Movie Before was significant. We used a repeated measure "opt-out point" with two levels (one for Movie Title and one Movie Before) each associated with a binary opt-out value. We fitted a generalized estimating equation regression model, with participant as random factor, trial and opt-out point as repeated factors, AR(1) as co-variance structure, type of explanation and opt-out point as fixed factors, and opt-out as dependent variable with a binomial distribution and logit link function. We included the main effects of type of explanation and opt-out point, their interaction, and a random intercept in the model. We used the SPSS v19 procedure GEE. There was a significant effect of opt-out point (Wald Chi-Square=15.20, $p < 0.001$). Surveying the direction of changes, H4 is confirmed - more participants were able to make decisions with the explanations than with just the title.

At first glance the opt-out rates for Movie Title seem to differ per condition. However, further statistical analysis (GLM) shows that this difference is not significant (F(2,129)=2.03, p=0.14). Nevertheless, in Section 6.2.5 we discuss if the large number of opt-outs for Movie Title in the *baseline* could be an artifact of the movies shown to the participants or individual differences.

Table 14 Experiment 4: Percentage of opt-outs for the three movie ratings. "Title", "Before" and "After" denote the three movie ratings based on the movie title only, and with the explanation before and after viewing the movie.

Condition	Movie Title	Movie Before	Movie After
Baseline	51.1%	28.9%	0%
Non-personalized	20.9%	7.0%	0%
Personalized	34.1%	11.4%	0%
Average	35.6%	15.9%	0%

Mid-scale ratings. It is also worth surveying the proportion of ratings in the middle of the scale (value=4). While these are not as strong an indicator of (lack of) informativeness as opt-outs, a larger proportion in the middle of the scale could suggest that participants had difficulty in forming a strong opinion. Movie Before and Movie After were distributed beyond the mean rating of 4, suggesting that participants are able to form opinions across the scale. Movie Before and Movie After ratings are also skewed toward the higher end of the
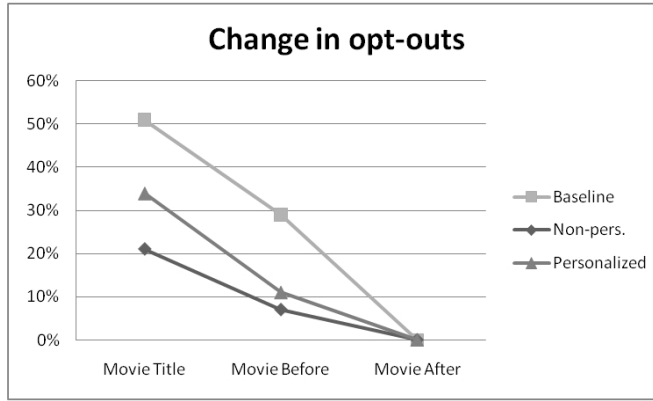
Fig. 3 Experiment 4: Change in opt-outs between the three movie ratings.

Table 15 Experiment 4: Means (StD) of the three movie ratings (excluding opt-outs). "Title", "Before" and "After" denote the three movie ratings based on the movie title only, and with the explanation before and after viewing the movie.

Condition	Movie Title	Movie Before	Movie After
Baseline	4.36 (0.95)	4.28 (0.81)	4.76 (1.67)
Non-personalized	4.12 (1.67)	4.45 (1.53)	4.58 (1.88)
Personalized	3.86 (1.23)	4.31 (1.26)	4.93 (1.86)

scale; there are more ratings of value 4 or above. This skew is not completely surprising given that we had selected movies that were unlikely to cause offense, as well as avoided genres and movies that participants did not want to see. There were more mid-scale ratings of 4 in the non-personalized and personalized conditions for Movie Title than for Movie Before and Movie After. The large number of opt-outs and mid-scale ratings suggest that users struggled to specify an opinion with the title alone.

6.2.2 Do explanations differ on how much they help to form an opinion?

We have seen that explanations help people form an opinion. We also investigated whether explanations differed on how much they helped. Table 14 shows the opt-out rates for Movie Before in the different conditions. There was a significant effect of condition on these opt-out rates ($F(2,129)=3.65$, $p<0.05$). The baseline condition had significantly more opt-outs than the other two conditions (sequential Bonferroni corrected, $p<0.05$). So, baseline explanations were the least helpful.

6.2.3 Are personalized explanations more effective? (H1)

The mean effectiveness in each condition is summarized in Table 16. Looking at the signed effectiveness we see that in all conditions the explanations

led to a slight underestimation. Comparing between conditions we found a significant difference in (absolute values for) effectiveness ($F(2,103.32)=4.39$, $p < 0.05$). Surprisingly, explanations in the baseline condition led to the "best" effectiveness (Bonferroni corrected, $p < 0.05$ for baseline compared to non-personalized, $p = 0.05$ for baseline compared to personalized). This finding is in stark contrast to the large number of opt-outs in this condition, which indicate that baseline explanations are clearly not helpful more than half of the time.

Table 16 Experiment 4: Means (StD) of effectiveness (excluding opt-outs) per condition.

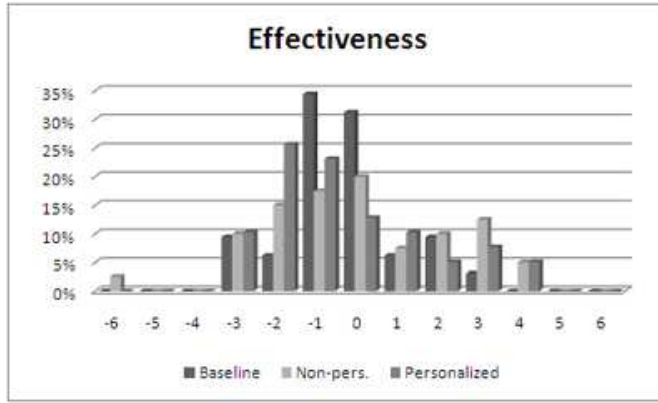| Condition	Effectiveness (absolute value)	Effectiveness (signed value)
Baseline	1.09 (1.00)	-0.41 (1.43)
Non-personalized	1.78 (1.37)	-0.08 (2.26)
Personalized	1.69 (1.08)	-0.41 (1.98)

Fig. 4 Experiment 4: Distribution of effectiveness, excluding opt-outs

One possible explanation for the comparatively high effectiveness of baseline explanations is that the baseline explanations biased the participants toward high ratings, as most of the short movies were in the top 50 in IMDB. As the selection of movies was guided by being acceptable to users, this also was likely to lead to a large proportion of high ratings. There was a skew toward high ratings in all conditions for Movie Before, but this skew was not the most severe in the baseline. Movie Before ratings are comparable for all conditions (no significant difference), suggesting that the baseline is not more strongly skewed toward positive ratings. We also surveyed how many of the shown movies in the baseline were in the top 50, and found that the relative proportion was comparable (44.4% were in the top 50, and 55.6% were not). So, it seems that the large number of mid-range ratings is a more plausible ex-

planation for good baseline effectiveness than a skew toward positive ratings. (We explain why a rating distribution with many mid-range ratings may lead to misleading measurements of effectiveness in Section 6.3.)

6.2.4 Are users more satisfied with personalized explanations? (H2)

We hypothesized that participants would prefer personalized explanations to non-personalized and baseline explanations. First, we look at the opt-out rates. In Table 17 we see that while the opt-out rates for the explanations in the two feature based conditions are comparable, the opt-out rate for explanations in the baseline is much higher.

Next, we investigate if participants preferred the personalized explanations

Table 17 Experiment 4: Means (StD) and percentage of opt-outs of the two explanation ratings, (on a scale from 1-7, 1=really bad, 7=really good) per condition. "Before" and "After" denote the two explanation ratings before and after viewing the movie. Means exclude opt-outs.

	Explanation Before		**Explanation After**	
Condition	**Mean (StD)**	**Opt-outs**	**Mean (StD)**	**Opt-outs**
Baseline	2.55 (1.43)	14.6%	2.89 (1.60)	2.1%
Non-personalized	3.51 (1.61)	4.5%	3.53 (2.00)	0.0%
Personalized	3.21 (1.46)	4.3%	3.16 (1.83)	2.2%

over the explanations in the other two conditions. First, we look at the initial explanation ratings (Explanation Before). There was significant effect of condition ($F_{(2,113.93)}$=4.05, $p < 0.05$). There was significant difference between the ratings for the baseline and non-personalized explanations (Bonferroni corrected, $p < 0.05$), but not for the ratings between the personalized condition and the other two conditions (although the mean rating for Explanation Before indicates that participants preferred non-personalized explanations over personalized ones). This contradicts our previous findings, where personalized explanations were preferred in both the camera and movie domain. We will discuss this further in Section 6.2.6.

In our previous experiments we were not able to compare the ratings for Explanation After as participants confused our test bed with Amazon itself. As this confusion no longer is a factor in this experiment, we can study the participants' opinion of the explanations after watching the movie. To our surprise we found no significant difference between the conditions, but note that the non-personalized explanations still have the highest mean rating, followed by the personalized explanations.

Table 18 Experiment 4: Number of times used and number of opt-outs for each movie per condition, and overall opt-out rate (%) for each movie across conditions. Movies have been sorted by overall opt-out rate, from lowest to highest.

	Overall	Baseline		Non-Pers.		Pers.	
Title	*%opt-outs*	*N*	*opt-outs*	*N*	*opt-outs*	*N*	*opt-outs*
Mr. Bean's Christmas‡	0	3	0	4	0	1	0
Wrong Trousers‡	0	2	0	0	-	1	0
Close Shave‡	0	1	0	0	-	1	0
Mickey's Trailer⋄	10	2	1	3	0	5	0
Hedgehog in the Fog†	20	4	1	2	0	4	1
Feed the Kitty⋄	29	1	1	2	0	4	1
Rabbit Seasoning	30	2	2	5	0	3	1
Jack shows Meg his Tesla coil	33	4	2	4	1	7	2
For the Birds†	43	5	2	6	3	3	1
Strange to Meet You	44	6	4	5	1	5	2
Kiwi⋄	50	1	1	1	0	0	-
Somewhere in California‡	50	6	3	6	3	4	2
Grand Day Out†	60	0	-	2	1	3	2
The Rocks†	63	5	3	1	0	2	2
Vincent	67	3	3	2	0	1	1

‡ = identical opt-outs, † = very similar opt-outs, ⋄ = difference due to 1 participant

6.2.5 Why did more participants opt out for Movie Title in the baseline condition and does it matter?

A larger proportion of participants opted out for Movie Title –before explanations had been given – in the baseline condition. We also see that the mean Movie Title rating is lowest in the personalized condition, despite users only being shown the title. We imagine two possible reasons for this: either the titles differ somehow, or there are notable individual differences between participants.

Difference in shown titles? If the movie titles shown to participants in the baseline were less informative (revealing less information such as the character in "Mr. Bean's Christmas") than in the other conditions, this could explain the presence of more opt-outs. Table 18 shows how often each movie was shown in each condition.[12] In general, the distribution of titles between the three conditions is comparable. However, there are some small differences which may have contributed to the higher opt-out rate for Movie Title in the baseline condition. For example, the baseline used The Rocks (which has a high overall opt-out rate) a bit more and Mickey's Trailer (which has a low overall opt-out rate) a bit less. Assuming that the overall opt-out rates per movie are representative and that participants are similar between conditions, the difference in movie distribution would result in a slightly higher expected

---

[12] Remember that movies shown were chosen randomly, but with participants requesting another movie if they had already seen it. It would have been difficult to ensure the same movies were used in each condition given we could not control what participants had already seen.

opt-out rate[13] for the baseline (38%) compared to the non-personalized (35%) and personalized (34%) conditions. So, the slightly different distribution of movies may have contributed, but is clearly not the only factor.

Differences between participants? It is harder to discuss the effects of individual differences as participants only rated up to three movies, but we did survey the rating patterns of individual participants. While participants in the baseline on average opted out more for Movie Title, this was not due to participants opting-out for all the titles they rated: there were many cases where participants opted out for some of the movies, but not all. Table 18 shows that for most movies Movie Title opt-out ratios are quite similar between conditions: they are identical for four movies, very similar for four movies, and differences are caused by only one participant for three movies. So, there does not seem to be a difference in general participant decisiveness between conditions. We conjecture that small deviations in participants' reactions to individual titles (e.g. somebody not thinking of Mickey Mouse when seeing the title Mickey's Trailer) caused differences in opt-outs.

So, overall, a likely explanation is that the difference in opt-outs for Movie Title may have been caused by a combination of a small difference in movies shown and some small individual differences.

Does the difference in Movie Title opt-outs matter? We do not expect the small difference in movies shown and small individual differences to have had much impact on our results. While it is likely that the higher opt-out rate in the baseline condition contributed to the high opt-out rate for Movie Before in the baseline, there is plenty of evidence to suggest that this high opt-out rate for Movie Before was also caused by the baseline explanations. Participants considered baseline explanations less informative, as evidenced by the high number of opt-outs for Explanation Before and the low ratings for Explanation Before (see Section 6.2.4). This is also corroborated by the large proportion of opt-outs for Movie Before in the baseline for all our previous experiments (see Table 21).

6.2.6 Qualitative comments

Participants may use the movie title to inform their opinion. Indeed, this is reflected in the participants' comments: *"..I considered what it might be like based on the title and the genre (personalized condition)"*, *"I thought the title (Rabbit Seasoning) suggested rabbits being killed but the light hearted nature of the film made me enjoy it more than I thought I would."*

We were surprised that personalized explanations in this experiment did not lead to higher satisfaction than non-personalized explanations. A large

---

[13] A condition's expected number of opt-outs has been calculated by summing over the movies, the product of the movie's overall opt-out rate and the number of times it was shown.

proportion of the explanations in the personalized condition described the actors, but the information is not as useful to the participants as they might have anticipated. The names used were largely unknown to participants, as it was harder to find short movies with known actors. For animations in particular, participants saw the names of actors whose voices were recorded, rather than the character they played (e.g. Bugs Bunny or Gromit). This could have decreased the satisfaction with the personalized explanations as participants were less likely to recognize them: ''*Also without knowing who the star is, this could still not mean a lot to the description.*". Likewise, some participants complained they did not recognize the director: *"the director's name is even less recognizable than the actors' names..."*; *"As I don't know the director, the rest of the description could easily belong to a totally different movie."*

Participants also commented on factors that were not considered so important in our focus groups (Tintarev and Masthoff 2007a), but which may have been identifying for the movies they were shown. For example, while our focus groups participants said they did not care about movie studio, this does affect the style of animation: *"...pretty much what I'd expect from a Pixar movie."*; *"Unlike the last movie I was not expecting a Walt Disney film..."*.

6.3 Summary of Experiment 4

The results of this experiment reinforce the importance of selecting relevant evaluation criteria. While the baseline explanations were found to be the most effective, they also had the lowest satisfaction, and led to most opt-outs and ratings in the middle of the scale.

The difference between opt-outs for the Movie Title and Movie Before ratings offers an argument in favor of explanations in recommendations: participants in all three conditions opted out a lot less after receiving even a simple explanation.

Both feature-based explanations were initially (Explanation Before) rated higher than baseline explanations, but only the difference between non-personalized and baseline explanations was significant. We believe that the weaker result for personalized explanations in this experiment compared to our previous experiments is due to the restricted choice of materials.

This experiment also brings into light two situations where our evaluation metric for effectiveness could fail: a) if a large proportion of ratings fall on the middle of the scale and b) if the explanations are biased in the same direction as the data. We discuss these further in our conclusion in Section 7.3.

We consider the effect of letting users watch the movies contra reading movie reviews on Amazon. In our case it is difficult to separate the effects of material choice from the effects of the change in design. The baseline explanations were the most effective in this experiment, but this does not seem to be due to an initial overestimation because Movie Before ratings were comparable between conditions (see Section 6.2.3). It is possible that the popularity of short movies is a better predictor than for long movies.

Our previous experiments in two domains used reading Amazon reviews as an approximation for real experience and led to repeated results. One could therefore also argue that using Amazon reviews leveraged the results for feature-based explanations w.r.t. effectiveness in our previous experiments. That is, reading reviews on Amazon caused an overestimation that correlated well with the (also overestimated) valuation of items after reading explanations. However, the average Movie After ratings in this experiment are also high, even though real experience replaced the reading of Amazon reviews. This bias would therefore also benefit from a presumed positive skew caused by feature-based explanations, but in this experiment no such bias is evident as the feature-based explanations show worse effectiveness. Our suggestion is therefore that the selection of movies was more likely to have affected our results than the change of design, in particular with regard to satisfaction. It is also worth considering that the baseline was more effective simply because IMDB is a good data source for short movies. Another alternative explanation for our results is that while the baseline did not offer the best possible explanations (inferring from the large number of opt-outs), the type of personalization we used in the personalized condition does not contribute to effectiveness. Naturally, further similar experiments with alternative item selection would be required to confirm that this really is the case.

7 Conclusions and future work

This paper has presented an overview of seven evaluation criteria for explanations in recommender systems and considerations related to choosing between them. We have discussed in detail the aim of effectiveness and how it can be measured, also highlighting limitations of the current metrics. We have presented a series of experiments in two domains investigating the impact of personalization and feature-based explanations on effectiveness. While there were issues related to the individual experiments (as discussed above), a meta-review of the series of experiments, presented below, allows us to draw overall conclusions with relative confidence. First, we address the question of why we should explain, or whether there is any point in explaining recommendations at all (Section 7.1). Then, we discuss if the type of explanations matters, or rather if our personalization of explanations increased their effectiveness (Section 7.2). Next, we summarize the lessons we have learned from the experiments about the used metric for effectiveness, and its relation to the underlying data (Section 7.3). Finally, we conclude with suggestions for future work.

7.1 The value of explanations

For a recommender system aiming at user satisfaction rather than decision support, well formed explanations can contribute positively: in Experiments

1-3 we saw that personalization of explanations does increase satisfaction compared to a baseline. Table 19 summarizes participant's satisfaction with the explanation, by condition, for each of the four experiments: MoviesI, MoviesII, Cameras and Final Evaluation.

Table 19 Summary: Means (StD) of initial satisfaction with explanations, excluding opt-outs, (on a scale from 1-7, 1=really bad, 7=really good) per condition per experiment.

Condition	MoviesI	MoviesII	Cameras	Final Evaluation
Baseline	2.38 (1.54)	-	2.83 (1.44)	2.55 (1.43)
Non-personalized	2.50 (1.62)	2.72 (1.68)	2.38 (1.64)	3.51 (1.61)
Personalized	3.09 (1.70)	3.31 (1.55)	3.27 (1.27)	3.21 (1.46)

Table 20 summarizes the change of opinion, where we hope to minimize this value[14]. We see that the average change (mean absolute effectiveness) for all explanations and experiments is reasonable: on the magnitude of 1 scale point on a 7 point scale. This suggests that explanations, our baselines included, offer relevant (albeit limited and imperfect) information, with the caveat that baseline explanations have led to more opt-outs for the initial rating in all four experiments (see Table 21 for a summary of opt-outs[15]). Part of the strong

Table 20 Summary: Means (StD) of absolute effectiveness (excluding opt-outs), per condition per experiment.

Condition	MoviesI	MoviesII	Cameras	Final Evaluation
Baseline	1.38 (1.20)	-	1.77 (1.50)	1.09 (1.00)
Non-personalized	1.14 (1.30)	0.96 (0.81)	1.14 (1.32)	1.78 (1.37)
Personalized	1.40 (1.20)	1.33 (1.27)	1.88 (1.34)	1.69 (1.08)

result for baseline explanations (when participants did not opt out) may have been due to the presence of a title for movies, but the replicated finding for cameras is promising (see also Table 20): explanations (with or without titles) can help in making decisions.

In the final evaluation, we allowed participants to rate the title alone, and then rate the item again once they saw the explanation. The number of opt-outs decreased *significantly* once participants received an explanation. That is, explanations also add to effectiveness in terms of increasing the number of items that users feel that they can evaluate.

---

[14] The lack of data for the baseline condition for MoviesII in Table 20 reflects a large opt-out rate, and participants removed for extremely short duration time.

[15] The smaller difference in opt-outs between the baseline and feature-based conditions in MoviesI is caused by the movie cover image being shown.

Table 21 Summary: Percentage of opt-outs for Item Before per condition per experiment

Condition	MoviesI	MoviesII	Cameras	Final Evaluation
Baseline	8.8	55.6	23.9	28.9
Non-personalized	7.2	4.3	16.7	7.0
Personalized	3.1	15.2	1.6	11.4

7.2 The value of personalized explanations

Table 22 summarizes the results related to effectiveness contra satisfaction across all experiments. In three initial experiments in two domains, we found that our method of personalization hindered effectiveness, but increased satisfaction with explanations. However, these experiments were based on an approximation of effectiveness where participants read reviews for items rather than trying them.

In our final evaluation, participants were able to watch the movies. In this case, the opt-out rate for the baseline explanation (Movie Before) was much higher than for the other two conditions (see also Table 21). For the remaining movie ratings, both feature-based explanation types were *less* effective than baseline explanations. Again, personalization did not improve effectiveness. Feature-based explanations led to higher satisfaction than baseline explanations. Contrary to our earlier results, personalized explanations were not preferred to non-personalized ones (the trend is even in the opposite direction). As discussed in 6.2.6, qualitative user comments suggest that the satisfaction with personalized explanations was decreased because participants recognized actors and directors less often (given that these tend to be more obscure in short movies). So, personalization of explanations only seems to increase satisfaction if it results in information that is meaningful to a user (rather than e.g. mentioning names of unknown actors).

Table 22 Overview of the results related to effectiveness and satisfaction across all experiments.

Experiment	Effectiveness	Satisfaction
MoviesI	Trend: non-personalized best	Significant: personalized highest
MoviesII	Trend: non-personalized best Significant: non-personalized had least opt-outs	Trend: personalized highest
Cameras	Significant: non-personalized best, personalized had least opt-outs	Significant: personalized higher than non-personalized Trend: personalized higher than baseline
Final Eval.	Significant: baseline best, but had most opt-outs	Significant: non-personalized higher than baseline Trend: personalized higher than baseline

While the results for the approximated and true effectiveness experiments are not entirely consistent, three conclusions can be drawn for all experiments:

1. Contrary to our initial hypothesis, personalization was in most cases clearly *detrimental* to effectiveness.
2. Users are likely to be more satisfied with feature-based than baseline explanations. If the personalization is perceived as relevant to them, then personalized feature-based explanations are preferred over non-personalized.
3. User satisfaction is also reflected in the proportion of opt-outs, which is highest for the baseline explanations in all experiments. This was the case despite the different types of baselines used in the two domains.

7.3 Lessons learned for effectiveness

Through our series of experiments, we have learned a few things about measuring effectiveness. Firstly, this metric does not consider opt-outs. As mentioned previously, the baseline explanations in our experiments suffered from a large number of opt-outs even if the effectiveness (measured as the absolute mean of the difference between the before and after ratings) was seemingly comparable with other conditions. For an explanation to be effective, it has to at the very least elicit some sort of rating (preferably one that reflects the user's preferences). An explanation that cannot help elicit *any* rating, by definition leads to poor effectiveness, and moreover is likely to result in user satisfaction so low that the system is likely to lose the user.

Secondly, the validity of the used metric (see Section 2.2) is dependent on the underlying data set. In Section 6.3 we highlighted two weaknesses of the used metric in relation to the underlying data set. We elaborate on them here.

7.3.1 If a large proportion of ratings fall on the middle of the scale.

Firstly, mid-scale ratings are ambiguous, we do not know if users are selecting this option because they cannot make a decision, or because they feel neutral (i.e. neither good, nor bad) about an item. The presence of an opt-out option helps clarify this, but only partially, as participants may supply a neutral value when they do not have enough information to form a polarized opinion. For example, in the final evaluation for the baseline Movie Before ratings we saw that participants gave more mid-scale ("4") ratings than in other conditions: this was more likely due to a lack of information than a large proportion of movies that were precisely ok.

Secondly, explanations that cause a majority of mid-scale ratings (for the rating before trying the item, e.g. Movie Before) are likely to lead to better effectiveness than explanations that result in more evenly distributed ratings. When the before rating is mid-scale, then even the most extreme change in opinion can only be as big as half of the scale. A wider distribution of the before ratings, assuming that the after ratings are normally distributed around

the middle of the scale, is likely to lead to greater divergences. Thus, smaller differences between before and after ratings might then (at least in part) be due to the poor distribution of the before ratings, rather than better effectiveness. So, explanations that lead to more polarized opinions may be more detrimental to the metric of effectiveness. This may have contributed to the downfall of the personalized explanations, which are likely to result in more polarized opinions. For example, personalized explanations may have led to users expecting to really like a movie when favorite actors were mentioned, only to be disappointed when those actors only had a minor role.

It is also arguable that if the initial ratings (Movie Before) are random but also follow a normal distribution (our current assumption for Movie After), the *signed* average difference would be 0. This is why we highlight the importance of measuring the *absolute* value of the difference.

7.3.2 If the explanations are biased in the same direction as the data.

Following on from the previous point, we can imagine an extreme scenario where a recommender system gives explanations that result in a large proportion of neutral ratings, and only recommends "safe" items that usually score around the middle of the scale. These explanations might appear to be effective[16], but are not likely to be particularly informative. This does not mean that the explanations *are* generally effective, as they would be misleading for non-neutral recommendations.

In addition, in our experiments we have seen that neither the before nor after ratings were centered around the middle of the scale. In this case, it makes more sense to consider the mean rating (e.g. 5 out of 7) for the after distribution (Movie After) rather than the middle of the scale (e.g. 4 out of 7). That is, false effectiveness may be found if there are many initial ratings (Movie Before) around the value that is the mean of the after ratings (Movie After). We can imagine explanations that inflate the initial valuation of items for a system that only recommends the most popular items; or explanations that devalue items and the system only presents unpopular items. In these cases our metric for effectiveness may result in high correlations, and a mean difference of 0 between the before and after ratings. However, this does not mean that the explanations are effective. For this reason, the underlying distribution of ratings should be presented alongside any measurement of effectiveness.

We caution that none of these situations per default imply a failed metric. The items may in fact be just ok, and a system that helps to identify this correctly should not be classified as faulty. Likewise, "slanted" explanations may be suitable if this fits the data e.g. positive explanations for items that the user is predicted to like. Baseline explanations like ours may make sense

---

[16] Assume the after ratings are normally distributed around the middle of the scale. Compare before ratings that are fixed at the middle of the scale, with before ratings that are fixed at another value (e.g. if all Movie Before ratings were equal to 5). The average difference between the before and after ratings would be smaller in the former case.

if they are based on many previous user opinions as is the case with the Internet Movie Database (IMDB). However, it would be prudent to assume that explanations cannot be ported between data sets, or domains without careful consideration. Any study using the same metrics for effectiveness would need to study the underlying distribution as well. For these reason we would encourage replication of this experiment with other materials and in different domains, to confirm which of our findings carry beyond our small selection of materials.

7.4 Future work

We found that while personalization could increase satisfaction in two domains, contrary to our hypothesis, it was detrimental to effectiveness. It may be the case that personalization in general does not increase effectiveness. We considered if this result is more specific to our studies, and discussed how our choice of experimental design, and type of explanations generated may have led to these surprising results. Of course, the outcome of the experiments may be dependent on the particular personalization chosen and the domains used.

Further studies are needed with more complex or simply longer explanations (e.g. based on deeper user models), using a different design (e.g. different materials), other domains, larger and less homogeneous participants' samples, and trying alternative ways of personalization. Of particular interest may be to compare explanations based on user models built using the different compositional preference elicitation methods described in Pommeranz et al (2012). Our suggestions for related future work also include using an implicitly learned user model given that users may not always know what information they need to make accurate decisions.

While the independence from a particular recommender system has allowed us to run controlled experiments, it would also be interesting to conduct studies with a live recommender system. That way one could for example conduct longitudinal studies, investigating e.g. the effect of explanations on trust, and see in which situations trust increases and decreases over time.

In addition, other researchers are starting to find that explanations are part of a cyclical process. The explanations affect a user's mental model of the recommender system, and in turn the way they interact with the explanations. In fact this may also impact the recommendation accuracy negatively (Ahn et al 2007; Cramer et al 2008b). For example, Ahn et al (2007) saw that recommendation accuracy decreased as users removed keywords from their profile for a news recommender system. Understanding this cycle will likely be one of the future strands of research.

References

Ahn JW, Brusilovsky P, Grady J, He D, Syn SY (2007) Open user profiles for adaptive news systems: help or harm? In: Proceedings of the 16th international conference on World Wide Web, Banff, Alberta, Canada, pp 11–20

Ardissono L, Goy A, Petrone G, Segnan M, Torasso P (2003) INTRIGUE: Personalized recommendation of tourist attractions for desktop and handheld devices. Applied Artificial Intelligence 17:687–714

Bilgic M, Mooney RJ (2005) Explaining recommendations: Satisfaction vs. promotion. In: Proceedings of the Workshop Beyond Personalization, in conjunction with the International Conference on Intelligent User Interfaces, San Diego, CA, pp 13–18

Billsus D, Pazzani MJ (1999) A personal news agent that talks, learns, and explains. In: Proceedings of the Third International Conference on Autonomous Agents, Seattle, WA, pp 268–275

Carenini G, Moore DJ (2001) An empirical study of the influence of user tailoring on evaluative argument effectiveness. In: Proceedings of the 17th International Joint Conference on Artificial Intelligence, Seattle, WA, pp 1307–1314

Chen L, Pu P (2007) Hybrid critiquing-based recommender systems. In: International Conference on Intelligent User Interfaces, Honolulu, HI, USA, pp 22–31

Cho Y, Im I, Hiltz JFSR (2003) The impact of product category on customer dissatisfaction in cyberspace. Business Process Managment Journal 9(5):635–651

Cramer H, Evers V, Someren MV, Ramlal S, Rutledge L, Stash N, Aroyo L, Wielinga B (2008a) The effects of transparency on perceived and actual competence of a content-based recommender. In: Semantic Web User Interaction Workshop in conjuction with the international conference on Human Factors in Computing Systems, Florence, Italy, pp 455–496

Cramer HSM, Evers V, Ramlal S, van Someren M, Rutledge L, Stash N, Aroyo L, Wielinga BJ (2008b) The effects of transparency on trust in and acceptance of a content-based art recommender. User Model User-Adapt Interact 18(5):455–496

Czarkowski M (2006) A scrutable adaptive hypertext. PhD thesis, University of Sydney

Dale R (1998) Dynamic document delivery: Generating natural language texts on demand. In: Proceedings of the 9th International Workshop on Database and Expert Systems Applications, IEEE Computer Society, Vienna, Austria, DEXA '98, pp 131–136

Felfernig A, Gula B, Teppan E (2007) User acceptance of knowledge-based recommenders. Machine Perception and Artificial Intelligence 70:249–276

Felfernig A, Gula B, Letiner G, Maier M, Melcher R, Schippel S, Teppan E (2008) A dominance model for the calculation of decoy products in recommendation environments. In: Symposium on Persuasive Technology in conjuction with Artificial Intelligence and the Simulation of Behavior Con-

vention, Aberdeen, Scotland, pp 43–50

Guy I, Ronen I, Wilcox E (2009a) Do you know? recommending people to invite into your social network. In: International Conference on Intelligent User Interfaces, Sanibel Island, FL, USA, pp 77–86

Guy I, Zwerdling N, Carmel D, Ronen I, Uziel E, Yogev S, Ofek-Koifman S (2009b) Personalized recommendation of social software items based on social relations. In: ACM conference on Recommender systems, New York City, NY, USA, pp 53–60

Häubl G, Trifts V (2000) Consumer decision making in online shopping environments: The effects of interactive decision aids. Marketing Science 19:4–21

Herlocker JL, Konstan JA, Riedl J (2000) Explaining collaborative filtering recommendations. In: ACM conference on Computer supported cooperative work, Philadelphia, PA, USA, pp 241–250

Hingston M (2006) User friendly recommender systems. Master's thesis, Sydney University, Australia

Laband DN (1991) An objective measure of search versus experience goods. Economic Inquiry 29(3):497–509

Masthoff J (2002) The evaluation of adaptive systems. In: Patel N (ed) Adaptive evolutionary information systems, Idea group publishing, pp 329–347

Masthoff J (2004) Group modeling: Selecting a sequence of television items to suit a group of viewers. User Modeling and User Adapted Interaction 14:37–85

McCarthy K, Reilly J, McGinty L, Smyth B (2004) Thinking positively - explanatory feedback for conversational recommender systems. In: Explanation Workshop in conjunction with the European Conference on Case-Based Reasoning, Madrid, Spain, pp 115–124

McCarthy K, Reilly J, McGinty L, Smyth B (2005a) Experiments in dynamic critiquing. In: International Conference on Intelligent User Interfaces, San Diego, CA, USA, pp 175–182

McCarthy K, Reilly J, Smyth B, Mcginty L (2005b) Generating diverse compound critiques. Artificial Intelligence Review 24:339–357

McNee SM, Riedl J, Konstan JA (2006a) Being accurate is not enough: How accuracy metrics have hurt recommender systems. In: International Conference on Human Factors in Computing Systems, Montreal, Canada, pp 1097–1101

McNee SM, Riedl J, Konstan JA (2006b) Making recommendations better: An analytic model for human-recommender interaction. In: Extended Abstracts of the 2006 ACM Conference on Human Factors in Computing Systems (CHI 2006), Montreal, Canada, pp 1103–1108

McSherry D (2005) Explanation in recommender systems. Artificial Intelligence Review 24(2):179 – 197

Murphy PE, Enis BM (1986) Classifying products strategically. Journal of Marketing 50:24–42

Oberlander J, Mellish C (1998) Final report on the ILEX project. online: http://www.hcrc.ed.ac.uk/ilex/final.html

Paramythis A, Weibelzahl S, Masthoff J (2010) Layered evaluation of interactive adaptive systems: Framework and formative methods. User Modeling and User-Adapted Interaction 20:383–453

Pommeranz A, Broekens J, Wiggers P, Brinkman WP, Jonker CM (2012) Designing interfaces for explicit preference elicitation: a user-centered investigation of preference representation and elicitation process. User Modeling and User-Adapted Interaction 22 (this issue)

Pu P, Chen L (2006) Trust building with explanation interfaces. In: International Conference on Intelligent User Interfaces, Sydney, Australia, pp 93–100

Pu P, Chen L (2007) Trust-inspiring explanation interfaces for recommender systems. Knowledge-based Systems 20:542–556

Pu P, Chen L, 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 (this issue)

Rashid AM, Albert I, Cosley D, Lam SK, McNee SM, Konstan JA, Riedl J (2002) Getting to know you: learning new user preferences in recommender systems. In: International Conference on Intelligent User Interfaces, San Francisco, CA, USA, pp 127–134

Reilly J, McCarthy K, McGinty L, Smyth B (2004) Incremental critiquing. In: SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence, Cambridge, UK, pp 143–151

Ricci F, Rokach L, Shapira B, Kantor P (eds) (2010) Recommender Systems Handbook. Springer

Roth-Berghofer T, Schulz S, Leake DB, Bahls D (2008) Workshop on explanation-aware computing. In: European Conference on Artificial Intelligence, Patras, Greece

Roth-Berghofer T, Tintarev N, Leake DB (2009) Workshop on explanation-aware computing. In: International Joint Conference on Artificial Intelligence, Pasadena, CA, USA

Roth-Berghofer T, Tintarev N, Leake DB (2010) Workshop on explanation-aware computing. In: European Conference on Artificial Intelligence, Lisbon, Portugal

Shapiro C (1983) Optimal pricing of experience goods. The Bell Journal of Economics 14 (2):497–507

Sinha R, Swearingen K (2002) The role of transparency in recommender systems. In: Conference on Human Factors in Computing Systems, Minneapolis, MN, USA, pp 830–831

Symeonidis P, Nanopoulos A, Manolopoulos Y (2008) Justified recommendations based on content and rating data. In: Workshop on Web Mining and Web Usage Analysis in conjunction with the International Conference on Knowledge Discovery and Data Mining, Las Vegas, NV, USA

Thompson CA, Göker MH, Langley P (2004) A personalized system for conversational recommendations. J Artif Intell Res 21:393–428

Tintarev N, Masthoff J (2007a) Effective explanations of recommendations: User-centered design. In: Recommender Systems, Minneapolis, MN, USA,

pp 153–156

Tintarev N, Masthoff J (2007b) A survey of explanations in recommender systems. In: WPRSIUI associated with ICDE'07, Istanbul, Turkey, pp 801–810

Tintarev N, Masthoff J (2008a) Over- and underestimation in different product domains. In: Workshop on Recommender Systems in conjunction with the European Conference on Artificial Intelligence, Patras, Greece, pp 14–19

Tintarev N, Masthoff J (2008b) Personalizing movie explanations using commercial meta-data. In: International Conference on Adaptive Hypermedia, Hannover, Germany, pp 204–213

Tintarev N, Masthoff J (2009) Evaluating recommender explanations: Problems experienced and lessons learned for evaluation of adaptive systems. In: UCDEAS workshop in conjuction with UMAP, Trento, Italy, pp 54–63

Tintarev N, Masthoff J (2010) Designing and evaluating explanations for recommender systems. In: Kantor PB, Ricci F, Rokach L, Shapira B (eds) Recommender Systems Handbook, Springer, pp 479–510

Vig J, Sen S, Riedl J (2009) Tagsplanations: Explaining recommendations using tags. In: International Conference on Intelligent User Interfaces, Sanibel Island, FL, USA, pp 47–56

Wang W, Benbasat I (2007) Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. Journal of Managment Information Systems 23:217–246

Wärnestål P (2005) User evaluation of a conversational recommender system. In: Workshop on Knowledge and Reasoning in Practical Dialogue Systems in conjunction with the International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, pp 32–39