In Pursuit of Satisfaction and the Prevention of Embarrassment: Affective State in Group Recommender Systems

Judith Masthoff and Albert Gatt

University of Aberdeen
Scotland, UK
jmasthoff@csd.abdn.ac.uk

Abstract. This paper deals in depth with some of the emotions that play a role in a group recommender system, which recommends sequences of items to a group of users. Firstly, it describes algorithms to model and predict the satisfaction experienced by individuals. Satisfaction is treated as an affective state. In particular, we model the decay of emotion over time and assimilation effects, where the affective state produced by previous items influences the impact on satisfaction of the next item. We compare the algorithms with each other, and investigate the effect of parameter values by comparing the algorithms' predictions with the results of an earlier empirical study. We discuss the difficulty of evaluating affective models, and present an experiment in a learning domain to show how some empirical evaluation can be done. Secondly, this paper proposes modifications to the algorithms to deal with the effect on an individual's satisfaction of that of others in the group. In particular, we model emotional contagion and conformity, and consider the impact of different relationship types. Thirdly, this paper explores the issue of privacy (feeling safe, not accidentally disclosing private tastes to others in the group) which is related to the emotion of embarrassment. It investigates the effect on privacy of different group aggregation strategies and proposes to add a virtual member to the group to further improve privacy.

Key words. Group modelling, affective state, satisfaction, recommender systems, privacy

1 Introduction

Inspired by Interactive TV, we are interested in recommending *sequences* of items (e.g. news stories, music clips) to *groups* of users. In (Masthoff, 2004a), we have discussed ten group aggregation strategies to combine individual ratings into a single group rating. For example, consider the set of individual ratings of items in Table 1, ranging from 1 (really hate) to 10 (really like). Assuming time to see six items, an *Average* strategy would recommend sequence EFHDJA by averaging over ratings. A *Least Misery* strategy would recommend sequence FEHJDG, taking the minimum of ratings. Empirical evaluation showed that the *Multiplicative* strategy, which multiplies ratings, was most successful in satisfying all individuals.

This paper (or a similar version) is not currently under review by a journal or conference, nor will it be submitted to such within the next three months.

Table 1. Example of individual ratings for ten items (A to J) for a group of three

	A	B	C	D	E	F	G	H	I	J
John	10	4	3	6	10	9	6	8	10	8
Adam	1	9	8	9	7	9	6	9	3	8
Mary	10	5	2	7	9	8	5	6	7	6
Average	7	6	4.3	7.3	8.7	8.7	5.7	7.7	6.7	7.3

This paper investigates how to predict the satisfaction of individuals during a sequence of items. For instance, how satisfied will Adam be at each point during the sequence EFHDJA? A recommender system that adapts to individuals can focus exclusively on maximizing *individual* satisfaction and for this it suffices to always recommend the item with the highest rating. So, there is no need to predict satisfaction accurately. However, if we are interested in keeping a *group* satisfied, accurate prediction of individual satisfaction becomes crucial. To keep the rest of the group happy, an individual might need to be confronted occasionally with items they do not like. An accurate prediction of satisfaction would help to (a) ensure no individual gets too dissatisfied, by presenting unliked items at appropriate times (e.g. when they are in a good mood), (b) evaluate group aggregation strategies under various conditions without the need for real users, (c) inspire a new optimal group aggregation strategy.

This paper extends the modelling of satisfaction in (Masthoff, 2004a) by treating satisfaction as an affective state and incorporating (a) emotional decay and assimilation and (b) the influence of the group on individual satisfaction. We also explore the issue of privacy.

The paper is organised as follows.[1] Section 2 discusses existing literature on affective state. Section 3 proposes three satisfaction functions based on this literature. Section 4 compares the predictions of these functions with the results from an earlier empirical study, and investigates which parameter values are most appropriate. Section 5 discusses the difficulties of empirically evaluating affective models. Section 6 presents an empirical study of our satisfaction functions. Section 7 discusses literature on how the satisfaction of others may influence an individual's satisfaction (focusing on emotional contagion and conformity), and proposes modifications to the satisfaction functions. Section 8 presents an empirical study into the effect of different relationship types on emotional contagion. Section 9 discusses the privacy issues of group recommender systems. Section 10 presents an empirical study into the effect on privacy of different group aggregation strategies. Finally, Section 11 presents our conclusions.

2 Individual Satisfaction

The satisfaction functions proposed in (Masthoff, 2004a) were based on the summation of satisfaction on preceding items and the impact of a new item. In the simplest satisfaction function, the impact of an item was taken to be its rating. Three factors were found to improve the functions: (1) inclusion of low ratings (noting dissatisfaction); (2) normalization, which takes into account ratings of not chosen items; and (3)

[1] An earlier version of Sections 1 to 4 has appeared as (Masthoff, 2005).

a quadratic, rather than a linear estimate of impact, which makes the difference between ratings of say 9 and 10 more significant than that between 5 and 6.

A limitation of these functions is that satisfaction is modeled as increasing with sequence length, and being independent of item order. Nevertheless, several reasons were given that suggest that order may be important (some related to advertising research, others to comments of subjects in experiments like "it is better to end on a high"). In this paper, we will refine the satisfaction function to take this into account.

It seems reasonable to regard satisfaction as an affective state or mood. Since the seventies, psychologists have researched the cognitive effects of mood, finding that it significantly impacts perception, attention, memory, information processing, and judgement (Oatley & Jenkins, 1996). More recently, psychologists and economists have started to research Affective Forecasting: how accurately people can predict what will make them happy, by how much, and for how long. Many studies have found that while people tend to be good at predicting whether they will be happy or unhappy, they are bad at predicting by how much and for how long (Wilson & Gilbert, 2003). Additionally, our own field has evinced a lot of interest into Affective Computing (e.g. Picard, 1997): how computers can recognize and respond to users' emotions, and how computers can simulate having emotions and portray these. We will briefly discuss the results out of these areas that seem most relevant to this paper.

2.1 Mood impacts judgment

Researchers have consistently found an impact of mood on evaluative judgement (for reviews see Oatley & Jenkins, 1996; Schwarz & Clore, 1996). For instance, Isen et al. (1978) found that people in a shopping mall who received a small gift that pleased them, reported in an unrelated survey a few minutes later that their cars and television sets were working better than did people who had not received that gift.

Much research has been done in the context of persuasion: people in a happy mood are easier to persuade than those in a neutral or sad mood (see e.g., Mackie & Worth, 1989). Mood effects are studied and used in commercial and political advertising (e.g., Aylesworth & MacKenzie, 1998; Isbell et al., 2003) with a view to influencing people's mood to make them judge a subsequent ad more favourably. In particular, a viewer's mood (as induced by watching the preceding program) has a significant effect on brand evaluations (Gardner, 1985; Meloy, 2000), with the viewer responding more positively if they were in a more positive mood. The *liking* of a television program has a similar significant effect (Murry et al., 1992; Schumann & Thorson, 1990). So, when a recommender system presents a sequence of items, viewing the first items could induce a mood, which could impact opinions on the next items.

Film clips are often used to elicit emotions in a laboratory setting, and a number of these have been empirically tested and found to consistently induce an emotion (Gross & Levenson, 1995; Rottenberg et al., in press). Researchers comment on the difficulty of finding clips that strongly elicit one emotion rather than a blend of several emotions. We will need to differentiate between the emotions elicited by the content of a clip, and the satisfaction with having seen it. For instance, one might be revolted by the content of a news item (e.g. on happenings in Iraq), but still be satisfied with having seen it. For our modelling, we will for now ignore the emotion elicited by the content, and concentrate on satisfaction in isolation. However, this issue will need to be addressed in future, and it will also complicate evaluation.

2.2 Retrospective feelings can differ from feelings experienced

A series of studies by Kahneman and colleagues (as reported in Wilson & Gilbert, 2003) found that the actual feelings experienced (e.g., self-reported pain during a colonoscopy) differed from those reported retrospectively. For instance, in one colonoscopy experiment, half the subjects were given a small additional period of pain after the end of the normal procedure, but surprisingly reported significantly less pain than the other subjects (Redelmeier et al., 2003). The retrospective reports were heavily influenced by the intensity of the emotional experience when it ended and the peak intensity of the experience. This is important to take into account when we try to measure satisfaction in an empirical study (to evaluate our models), and poses questions regarding what kind of satisfaction we ought to model (for instance, satisfaction at the end of a sequence of items versus at each point during the sequence). Note that we do not intend to make a recommender system that purposely "leaves the needle in longer" (gives additional small misery) to make the experience *feel* better while objectively being worse. However, we could change the order of items in the sequence, to take these kinds of effects into account.

2.3 Expectation can influence emotion

People's affective forecasting can change their actual emotional experience (e.g., Wilson & Klaaren, 1992). Several studies have shown that if you expect to like something, then you might end up liking it more than if you did not have any expectations. For instance, subjects who were told that jokes were going to be funny, found the jokes funnier than subjects who had not been told this. This is called *assimilation*. So, if our recommender system has presented several items you liked, than you might expect to like the next item as well, and therefore your perceived satisfaction with that item might be higher than its actual rating merits. It has been suggested that the opposite can also happen (Wilson & Klaaren, 1992): if you expect to like something, you might end up liking it less. This is called *contrast*. Since, according to Wilson and Gilbert (2003), hardly any studies support contrast, while several support assimilation, we will focus exclusively on the latter.

2.4 Emotions wear off over time

People's emotional reactions become less intense with time, a phenomenon called Emotional Evanescence (Wilson et al., 2003). When something has made you happy, this happiness does not last, but wears off. This is explained in (Wilson & Gilbert, 2003) as needed to save energy and to protect the emotions' role as a signalling device. Another explanation of Emotional Evanescence is given by adaptation-level theory, which claims that an emotional reaction depends on how the current experience compares to similar past experiences. The example given in (Wilson & Gilbert, 2003) is that of temperature: if you have been used to low temperatures, arriving in Florida to a temperature of 15° C might please you. After staying for a while, 30° C becomes your standard, and a day with the same temperature of 15° C might displease you. Note that this is subtly different from the idea of *contrast*. Adaptation level theory does not assume you will *expect* the weather to be good, and compare the actual weather to your expectation, but rather that you will compare it to your standard of what constitutes good weather, a standard which has been influenced by your experiences in the past. So, if you were watching a sequence of items, the items you have watched so far would give you a standard to compare the next item to.

As noted by Wilson & Gilbert (2003), a limitation of adaptation-level theory is that it does not specify what the comparison point should be. Should we only look at the last item or take the average over the last, say five, items? It also does not explain the decrease in emotions over time evoked by a single isolated event. For this paper, it is sufficient to note that a decrease of emotions over time happens; we will not try to model adaptation level theory yet.

2.5 Ideas from affective computing

Picard (1997) discusses the use of emotions and moods in computer systems. She describes emotions as being regulated by moods: for instance, a good mood increases the impact of relatively small positive events. She proposes to model mood as a weighted summation of the impact of positive events and subtraction of the impact of negative events, with the impact of recent events receiving extra weight. She argues that mood can not be of unbounded intensity, so a limit should be used (or a saturation function). Elliot & Siegle (1993) also mention that prior mood should contribute to emotion intensity. They further mention that the relationship between users can influence emotions (we will return to this issue in Section 7).

3 Satisfaction Functions Chosen

In (Masthoff, 2004a), satisfaction with a (fixed-length) sequence of items was calculated as the sum of the impact on satisfaction of the individual items of the sequence:

$$\mathrm{Sat}(items + < i >) = \mathrm{Sat}(items) + \mathrm{Impact}(i), \quad \text{for item } i \text{ and item sequence}^{2}\ items.$$
$$\mathrm{Sat}(<>) = 0$$

Based on the literature review above, we will henceforth assume that Satisfaction will decrease in intensity over time. Also, we give more weight to recent items[3].

Variant 0 $\mathrm{Sat}(items + < i >) = \delta \times \mathrm{Sat}(items) + \mathrm{Impact}(i), \quad \text{with } 0 \leq \delta \leq 1$

[2] Whenever we talk about an item sequence, we mean a sequence of distinct items.

[3] This deviates from the proposal in (Picard, 1997) of a linear decrease in weight, and only consideration of the last four events.

With δ=1 no decrease of satisfaction over time occurs, and with δ=0 no memory of past items would be used. The value of δ is likely to depend on personality, as advocated in (Picard, 1997; de Rosis et al., 2003), and item duration.

The impact of an item depends on the individual's liking of that item, as captured by a preference rating[4]. Rating(*i*) will denote the preference rating for item *i*. To calculate the impact, we perform three steps.

1. *Normalization*. In (Masthoff, 2004a), satisfaction produced was normalized, as opposed to individual rating, using for item sequence *items*:

$$\text{NormSat}(\textit{items}) = \frac{\text{Sat}(\textit{items})}{\text{PossSat}(\textit{items})}$$

$$\text{with}\quad \text{PossSat}(\textit{items}) = \underset{\text{item sequence } s\ \wedge\ \text{length}(s)=\text{length}(\textit{items})}{\text{Max}} \text{Sat}(s)$$

   Our introduction of δ into the Satisfaction function means that normalization now needs to be handled differently. We apply normalization to the rating, not to satisfaction as a whole, to make it independent of the selections so far. For rating *r*,

$$\text{Normalized}(r) = r \times \frac{\text{TotalRatingsExpected}}{\text{TotalRatingsPossible}},$$

$$\text{with}\quad \text{TotalRatingsExpected} = \sum_{\text{item } j} \text{AverageRating}$$

$$\text{TotalRatingsPossible} = \sum_{\text{item } j} \text{Rating}\ (j)$$

   Rather than taking AverageRating to be the midpoint of the scale (5.5), we use the average rating over all individuals in the group over all items (6.93 in Table 1).

   Note that normalization results in ratings 10-10-10-10 being treated as the same as 5-5-5-5. This was also the case in the normalization in (Masthoff, 2004a). In a sense, this is also a form of assimilation.

2. *Rebalancing*. We rebalance the rating scale by subtracting its midpoint (5.5 for our scale), so that negative ratings imply dissatisfaction.

$$\text{Rebalanced}(r) = r - \text{midpoint}, \qquad \text{for rating } r$$

3. *Making the impact quadratic*. We will use the following formula[5]:

$$\text{Quadratic}(r) = \quad r^2, \quad \text{if } r \geq 0; \qquad -r^2, \quad \text{if } r < 0, \quad \text{for rating } r$$

Combining these three steps, we obtain:

$$\text{Impact}(i) = \text{Quadratic}(\text{Rebalanced}(\text{Normalized}(\text{Rating}(i))))\ \text{, for item } i$$

[4] How ratings have been obtained is beyond the scope of this paper, but they are likely to have been inferred rather than explicitly given (see Masthoff, 2004a). Here we assume ratings to be accurate.

[5] This deviates from (Masthoff, 2004a), where no rebalancing occurred and Quadratic(r) was $(r-5)^2$, if r ≥6, and $-(r-6)^2$ if r<6. This simplifies dealing with the new normalization of ratings.

The satisfaction function discussed above sums the (weighted) satisfaction so far with the impact of the new item. The upshot is that satisfaction is predicted to keep increasing if a series of items is consistently pleasing. The question arises whether summation is indeed the right operation, given that some of the literature discussed above says that mood cannot be of unbounded intensity. An alternative would be to take the average. We therefore propose the following variant:

Variant 1 $\mathrm{Sat}(\mathit{items} + < i >) \ = \ \dfrac{\delta \times \mathrm{Sat}(\mathit{items}) \ + \mathrm{Impact}(i)}{1+\delta}$

We divide by 1+δ rather than 2, to get Sat(*items*+<*i*>) = Impact(*i*) when δ=0.

Although the satisfaction function takes decay of emotion into account, it does not yet take into account that mood can alter judgement, and that expectation can influence emotion. Since the literature suggests that impact of an item depends on mood, showing an assimilation effect, we propose the following variant:

Variant 2 $\mathrm{Sat}(\mathit{items} + < i >) \ = \ \delta \times \mathrm{Sat}(\mathit{items}) + \mathrm{Impact}(i, \delta \times \mathrm{Sat}(\mathit{items}))$

with $\mathrm{Impact}(i, s) = \mathrm{Impact}(i) + (s - \mathrm{Impact}(i)) \times \varepsilon$, for all s and $0 \leq \varepsilon \leq 1$

This variant ensures for instance that the impact of an item is higher than it would normally be, if the user's mood (in terms of satisfaction with the sequence so far) is high. Parameter ε models the extent to which the user's mood influences that user's evaluative judgement: with ε = 0 there is no such influence (and we have the same satisfaction function as Variant 0), with ε = 1 the influence is so immense that the new item cannot have any impact on the user's mood.

It is easy to come up with more variants (e.g. combining Variants 1 and 2), but we would first like to understand the effects of the satisfaction functions proposed so far.

4 Simulations

To gain insight into how the variants differ in their predictions, we ran simulations: taking Table 1's ratings for John, Adam and Mary, and eight different item sequences (e.g. FEAHD), each variant predicted how satisfied each individual would be at each point during each sequence. Varying values of δ and ε were used. Figure 1 shows example results for Variant 0 (for full results see Masthoff, 2005). As can be expected, a lower value of δ tends to result in a lower predicted satisfaction. Comparison of the results for sequences AIEFD and AEFID shows the impact item order can have, particularly for small δ. Figure 2 shows example results for Variant 1 (for full results see Masthoff, 2005). The 'averaging' used in Variant 1 has resulted in a nice limitation of satisfaction, rather than an infinite increase. The value of δ clearly has less impact on satisfaction for Variant 1 than for Variant 0. Figure 3 shows example results for Variant 2. There is a large impact of δ: with δ=0.9, all individuals are about equally satisfied and the impact of ε is low. With δ=0.2, the individual satisfaction profiles are quite different, and the impact of ε is high.

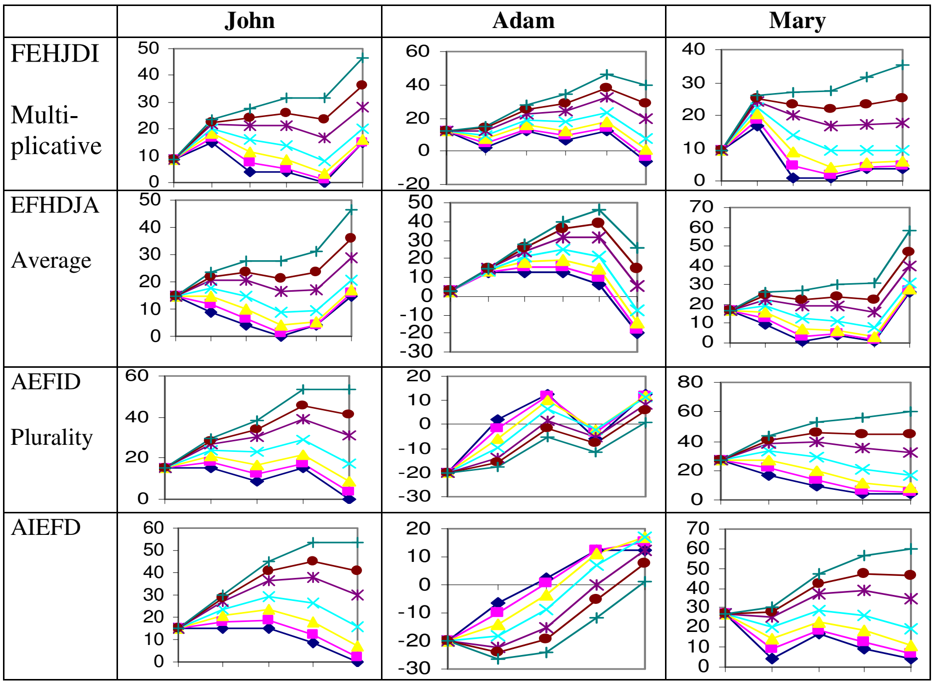


Fig. 1. Predicted satisfaction for Variant 0 per sequence per individual, for several values of δ: —◆— 0 —■— 0.2 —▲— 0.4 —✕— 0.6 —✱— 0.8 —●— 0.9 —+— 1. Y-axis shows satisfaction, X-axis time (first item shown, second shown, etc)

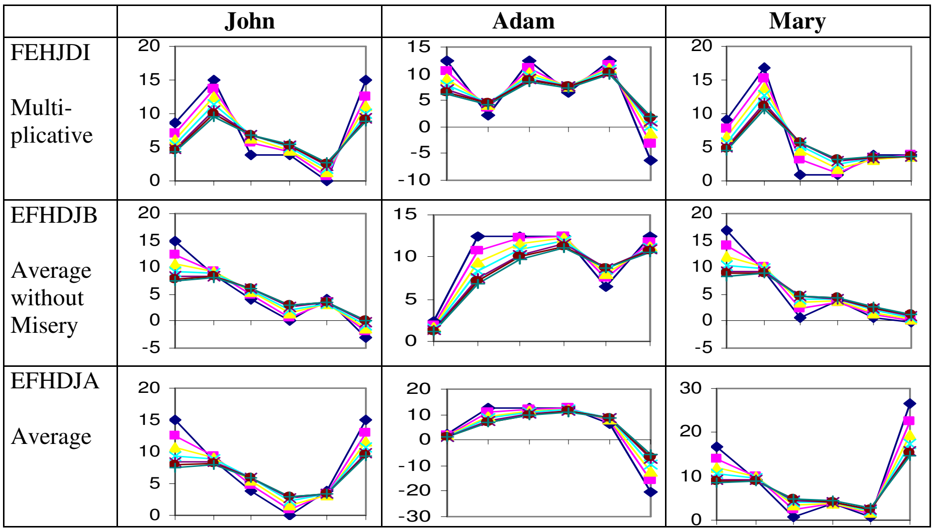


Fig. 2. Predicted satisfaction for Variant 1 per sequence per individual, for several values of δ: —◆— 0 —■— 0.2 —▲— 0.4 —✕— 0.6 —✱— 0.8 —●— 0.9 —+— 1. Y-axis shows satisfaction, X-axis time (first item shown, second shown, etc)

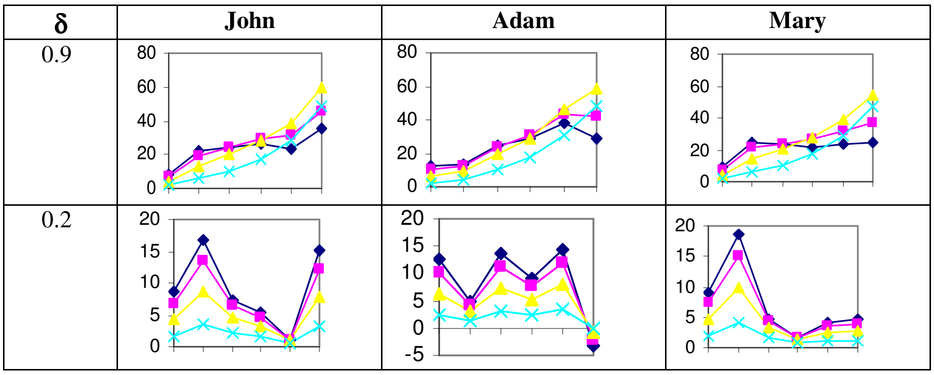


Fig. 3. Predicted satisfaction of Variant 2 for sequence FEHJDI per individual, for several values of ε: —◆— 0 —■— 0.2 —▲— 0.5 —×— 0.8, and two values of δ. Y-axis shows satisfaction, X-axis time (first item shown, second shown, etc)

The same ratings and sequences had been shown to participants of an experiment published in (Masthoff, 2004a), who were asked to predict how satisfied each individual would be with the sequence as a whole[6]. We compared the predictions of the variants with those by the participants (as given in Masthoff, 2004a) in order to determine what values of δ and ε best matched the human predictions. As participants judged the sequences as a whole, we compared their opinions with the complete satisfaction curves produced by the Variants, rather than just the end point. So, we regard Variant 0's prediction for AEFID to be that it dissatisfies Adam, as he is predicted to be dissatisfied after the presentation of the first and fourth item (for any δ).

For low values of δ, Variant 0 predicts sequences FEHJDI and EFHDJA to dissatisfy Adam, and EFHDJB to dissatisfy John. This contradicts the opinion of participants, who predicted these sequences to satisfy Adam and John. We therefore conclude that a higher value of δ (0.8, 0.9, 1) is more in accordance with the participants' behaviour. This is also supported by the predictions of Variant 2: with δ=0.2, Adam is predicted to be dissatisfied at the end, independent of ε. As participants predicted Adam to be satisfied, again a higher value of δ seems better. It is harder to see which value of ε produces the best match with participants' behaviour. Participants predicted Adam and Mary to be equally satisfied with FEHJDI, and John to be slightly more satisfied. When looking at the end of the sequence, a lower value of ε seems to match participants' opinions slightly better. When looking at the curves as a whole (e.g., averaging satisfaction over the sequence), John would be more dissatisfied than Adam, independent of ε.

The eight sequences used are those recommended by various group aggregation strategies (as described in Masthoff, 2004a). We investigated whether the simulation results could provide insight into the suitability of these strategies. Variants 0 and 1 both predict Adam to be dissatisfied with sequences FEAHD, AEFID and AEIBDF, independent of the value of δ. Hence, the strategies that recommend these sequences (Borda,Plurality and Most Pleasure as described in Masthoff, 2004a) can be excluded.

[6] Reasons for the indirectness of that experiment are given in (Masthoff, 2004a) and Section 5.

This is in-line with the conclusion in (Masthoff, 2004a), which was based on participants' opinion. Restricting ourselves to higher values of δ, if we take the satisfaction of the group to be the minimum of the satisfaction of the individuals, then in Variant 0 FEHJDI clearly performs best: for each moment in the sequence, it gives the highest satisfaction (except after the second item for δ is 0.8 and 0.9). This sequence was recommended by the Multiplicative strategy, which was also the strategy preferred by the participants. However, if we take the satisfaction of the group to be the average of the satisfaction of the individuals, then the Average strategy performs best. Taking the minimum corresponds better to the predictions of our subjects.

Variant 1 also predicted Adam to be dissatisfied with EFHDJA independent of δ. This contradicts the opinions of participants, who predicted this sequence to satisfy Adam. There are three possible explanations. Firstly, the 'averaging' in Variant 1 may be wrong. Secondly, participants' predictions may be wrong. Thirdly, both the 'averaging' and subjects' predictions may be correct, but something else in the satisfaction function is wrong, like the normalization (or the constant used in it), or the use of quadratic ratings.

5 Inherent Complexity of Empirical Evaluation

We have shown how we can use satisfaction functions to reason about group aggregation strategies, even when these functions are not yet completely validated. For instance, several strategies performed badly, independently of δ, and of whether 'averaging' was used. Comparing the satisfaction functions with predictions by subjects, allowed us to conclude that δ should have a high value (for this particular task), and given such a high value, the Multiplicative strategy performed best. However, there is a problem: the empirical data used comes from an indirect experiment, where subjects were not shown items, but only ratings, and had to predict satisfaction. As discussed above, the Affective Forecasting literature shows that people are very poor at predicting the intensity of emotions[7]. So, a next step should be to empirically compare the real satisfaction experienced by subjects with that predicted by the satisfaction functions. However, this is particularly challenging; indeed it was the topic of an "evaluation challenge" for the 2005 UM workshop on the Evaluation of Adaptive Systems.

Firstly, as pointed out by David Chin's entry to the competition, items can evoke a wide range of emotional responses (Chin, 2005). It may be hard to distil satisfaction from these. Secondly, accurate ratings for the individuals are needed. We could ask subjects to rate items, but if our assumption is right, then their rating would depend on presentation order. Although this could be avoided by introducing long intervals between items, this would increase the likelihood that subjects' moods vary. Attempts to iron out these issues through items that elicit identical ratings for all subjects failed (Masthoff, 2004a). This individual variation in the emotive value of items is also noted by Chin (2005), who points out that a song that instils happiness in the majority may evoke sadness in someone whom it reminds of a deceased friend. A further complication is that such associations vary over time: the appraisal of a song that reminds

[7] People are, however, good at predicting valence, so this does not invalidate our conclusions

someone of their spouse will depend on the state of the marriage, making it hard to ask subjects to rate items on one day, and then to do the experiment on a later date.

Thirdly, items should be unrelated, as topical relatedness can influence judgement (e.g. an item about an earthquake in Bulgaria getting a higher rating after an item on their team playing Bulgaria) (Masthoff, 2004a). In the case of music recommendations the tempo and genre of the songs is likely to have such an impact (Chin, 2005).

Finally, we need to know how satisfied each subject is after each item has been presented. Though self-reporting has been suggested as a convenient and relatively accurate way of measuring affective state (de Vicente, 2003), it poses two potential problems. Firstly, if subjects are asked at the end of an experiment to judge their satisfaction at preceding time points, reports are likely to be inaccurate, given the difference between retrospective and experienced emotions. Secondly, asking subjects to report satisfaction during the experiment takes time and may influence the results (given the decrease in emotions over time). Using sensor data has been suggested as a more accurate way to determine affective state as it avoids introspection (for an overview see Picard et al., 2004). However, if we determine satisfaction on the basis of sensor data (e.g. heart rate as suggested by Chin, 2005), the emotional content of items can influence results. Also, as indicated by Chin (2005) sensor data seems better at spotting the absence or presence of an emotion (e.g, like or dislike) than the actual strength of the emotion, which is most relevant for evaluating our models. Finally, sensors tend to be both costly and intrusive.

6 Experiment 1: Empirical Evaluation on a Learning Task

As discussed above, empirical evaluation of our satisfaction functions is inherently complex. In this section, we nevertheless present a first empirical study[8], the purpose of which is (a) to yield insight into our satisfaction functions and (b) to explore how studies of this kind can be conducted.

To overcome some of the problems discussed above, the study was conducted in a learning domain, rather than a recommender domain. A learner's affective state is likely to be multidimensional: Kort et al. (2001) distinguishes between the dimensions of anxiety-confidence, boredom-fascination, frustration-euphoria, dispirited-encouraged, terror-enchantment. Heylen et al. (2003) adds social emotions, like embarrassment and pride. We focused on satisfaction with performance, using our satisfaction functions to model learners' satisfaction with their performance on a sequence of tasks. Järvenoja & Järvelä (2005) found that performance is an important source of emotion in secondary school students.

For this experiment, a learning domain was chosen over a recommender domain for a number of reasons. The setup of the experiment required items for which we could accurately predict satisfaction. As discussed above, such items are hard to obtain in a recommender domain. In contrast, we hypothesised that we could accurately modify task difficulty, and that task difficulty would correlate with satisfaction with

[8] This experiment happened before the Evaluation of Adaptive System challenge.

performance. It also seemed easier in a learning domain to avoid the topical relatedness problems discussed above, and to keep the tasks free of emotional content.

Weiner (1995) distinguishes between two affects relating to achievement: 'outcome-dependent' (how well you did) and 'attribution-linked' (was it because of what you did or not). So, it may be that a learner's satisfaction with their performance depends not only on how well they performed, but also on how difficult they perceived the task to be. If they perceive the task to be very difficult, then they could be satisfied with a lower score than if they perceived the task to be easy. This is not covered in the models above, and it will be part of the evaluation to see if such an effect does occur.

6.1 Adjustment and addition to the models

Before performing the experiment, two changes were made to the modelling. First, we defined a learner's satisfaction with a task *t* as:

$$\mathrm{Sat}(<t>) = \mathrm{Impact}(t).$$

This is a slight variation on the models given in Section 3, as it results in a different value of Sat(<*t*>) for Variants 1 and 2. After all, in Section 3, we had defined Sat(<>) = 0 rather than defining Sat(<*t*>). This resulted in

$$\mathrm{Sat}(<t>) = \mathrm{Sat}(<>+<t>) =$$

$$= \begin{cases} (\delta\times\mathrm{Sat}(<>)+\mathrm{Impact}(t)) / (1+\delta) = \mathrm{Impact}(t) / (1+\delta) & \text{for Variant 1} \\ \delta\times\mathrm{Sat}(<>) + \mathrm{Impact}(t,\delta\times\mathrm{Sat}(<>)) = & \\ \quad \mathrm{Impact}(t,0) = \mathrm{Impact}(t) + (0-\mathrm{Impact}(t))\times\varepsilon = (1-\varepsilon)\times\mathrm{Impact}(t) & \text{for Variant 2} \end{cases}$$

This seems counter intuitive, hence we now define Sat(<*t*>) instead. Secondly, a variant of the models was added, which basically combines Variants 1 and 2:

Variant 3 $$\mathrm{Sat}(items + < i >) \;=\; \frac{\delta\times\mathrm{Sat}(items) + \mathrm{Impact}(i, \delta\times\mathrm{Sat}(items))}{1+\delta}$$

6.2 Experimental Design

6.2.1 Method and Procedure

For the purpose of this experiment, it was essential to have a carefully controlled task setting, allowing an accurate a priori estimate of task difficulty. For this reason, a lexical decision task (a standard paradigm in psycholinguistic research, see e.g. Harley, 2001) was administered, using DMDX, a software package for running reaction time experiments (Forster & Forster, 2003). In each task, subjects were exposed to twenty letter strings flashing for short intervals in succession on a computer screen, and asked to decide whether the string was a real English word or not by pressing a designated "yes" or "no" key. In each task, ten word and ten non-word strings were used in a randomized order. Error rates and reaction times were recorded; however, we will only discuss error data below. Before the start of the experiment, a practice task, using numbers instead of words, was administered to familiarise them with the software.

Task difficulty depends on at least two factors. The first is the time window in which subjects are exposed to a string and have to respond. Our experiment kept this constant across tasks: strings were flashed on screen for 800ms and a response had to be made within 2000ms from the onset of each string. An inter-stimulus interval of 750ms was also provided, during which the screen was blank. The second factor affecting task difficulty, and the one manipulated in our experiment, is the familiarity of the words presented. Familiarity of words was systematically varied across tasks, such that one task was “easy”, one “medium” and one “difficult”.

Subjects were divided into two groups (A and B), differing only in the order in which the difficult and easy tasks were presented. Both groups did the medium task (M) last. Group A subjects saw the other tasks in the order *difficult (D) – easy (E)*, while Group B had the opposite order (E – D). Our satisfaction functions imply that the order of tasks impacts satisfaction. Our between subject design allows us to investigate this. Best values for δ and ε will be investigated across the sample as a whole.

In order to assess their affective state at the start of the experiment, subjects were first asked to rate their current mood, using a slider going from a sad face to a happy face. The slider position returned a value in the interval (1,100). This method makes the “mood” variable continuous (i.e. it is measured on a ratio scale). Thus, it permits explicit comparison of different measures, which is not the case with nominal or ordinal scales (such as “Likert-type” measurements). The latter are rather coarse-grained in that (a) they restrict subjects to predefined measures; (b) the distance between adjacent points on the scale is unknown.

After each of the three tasks, subjects received feedback on the number of correct answers (e.g. “you answered 15 from 20 correctly”), followed by two questions to assess their affective state, namely:

- “How satisfied are you with your performance on the last task?”
- “How satisfied are you with your performance so far?”

The questions were answered using the same type of slider as per the initial mood rating. There was an inter-task interval of 2.5 minutes. Each task took about 1 minute.

6.2.2 Subjects

Twenty-two students of the University of Aberdeen participated in the experiment in exchange for entry in a prize draw. Eleven subjects were randomly assigned to each experimental condition. Subjects self-rated their fluency in English, choosing from “native”, “non-native but fluent” and “non-fluent”.

6.2.3 Materials

Each task consisted of 10 words and 10 non-word strings, all of five characters in length and composed of two syllables. For each real word, there was a non-word string with an identical initial character. Non-words were selected from the database of the *English Lexicon Project* (Balota et al., 2002). Real words were selected from the MRC psycholinguistic database, a machine-readable dictionary containing psycholinguistic norms for English words (Wilson, 1988), including norms for subjective familiarity ratings obtained from native speakers. The normalised scale for subjective word familiarity in MRC ranges from 100 to 700. Under the assumption that word recognition would be more difficult for less familiar words, task difficulty was operationalised by identifying the following three intervals on this scale:

- easy (E): 450- 600; medium (M): 300 – 450; difficult (D): 200 - 300

The frequency of the words selected, obtained from the British National Corpus (BNC; Aston & Burnard, 1998), correlated significantly with subjective familiarity. In fact, the log-transformed frequency and the subjective familiarity ratings of the selected words presented a near-perfect correlation ($r = .946$; $p < .01$).

6.2.5 Research Questions

The aim of the experiment was to answer the following research questions:

- Is it possible to reliably manipulate satisfaction with individual tasks by manipulating task difficulty? In particular, we need to establish
  - whether task difficulty can be reliably manipulated, that is, whether D is more difficult than E, and M of intermediate difficulty
  - that a higher score results in higher satisfaction, and that task difficulty does not influence subjects' satisfaction with their score (to exclude the 'attribution-linked' sense of achievement discussed above). Are subjects indeed more satisfied with their (higher) score on E than their (lower) score on D?
- Which variants of the satisfaction function perform best? Emotion wearing off predicts that having the easier task first would result in lower satisfaction after two tasks (given that satisfaction with task 1 will have decayed). Assimilation predicts that having the easier task first would result in higher satisfaction after two tasks, as satisfaction with good performance on the first task would increase expectations of good performance on the second, resulting in higher satisfaction overall. Variants 0 and 1 do not model assimilation, so if they are correct then we should find that group A (who had the easier task last) is more satisfied after two tasks than group B. Variants 2 and 3 model emotions wearing off as well as assimilation, so they may be true regardless of whether we find a difference or not.
- What values for δ and ε produce the best fit? Do we indeed need values per user?

6.3 Results and discussion

6.3.1 Task difficulty

Table 2 shows the average performance over all subjects on tasks D, E, and M. In this section, we report p-values on pairwise t-tests unless otherwise stated. The manipulation of task difficulty proved reliable, with D more difficult than E (p<.00001), and M at an intermediate level significantly below D (p<.00001) and above E (p<.05). Independent samples t-tests confirmed that Group A and B differed reliably on their first (p<.005) and second (p<.005) tasks.

Table 2. Correct responses over all subjects.

	Task D	Task E	Task M
Average	12.2	17.1	16
Standard Deviation	2.8	2.6	2.7

6.3.2 Comparability of groups: performance and initial mood

To ensure that the groups were homogeneous, we compared initial mood and task performance. Table 3 shows the correct responses per group. No significant differences were observed between groups on D (p>.7), E (p>.1), M (p>.1), and all tasks combined (p>.3). Thus, group performance was homogeneous and independent of task order.

Table 4 shows the self-reported initial mood of subjects in both groups. Slider results were transformed by deducting the midpoint of the scale (50), dividing by ten (obtaining a scale from -5 to 5), and squaring while maintaining the sign (obtaining results between -25 and 25). All other slider results will be transformed in the same way. One subject in group B was an outlier in the initial mood rating (-19.22) and was excluded from all subsequent analyses regarding satisfaction. No difference was found between the groups' initial mood (t-test, p>.7), suggesting homogeneity.

Table 3. Correct responses per group for each task presented.

	Group A		Group B	
	Average	STDEV	Average	STDEV
First task	12	2.9	18	1.3
Second task	16.3	3.2	12.5	2.9
Third task	15.7	2.6	16.3	2.9
Over all tasks	44	7.7	46.7	5.5

Table 4. Self-reported initial mood

	Before removing outliers		After removing outliers	
	Average	STDEV	Average	STDEV
Mood of group A	5.01	5.03	5.01	5.03
Mood of group B	3.50	8.82	5.77	4.84

6.3.3 Satisfaction with performance on individual tasks

To test whether task difficulty indeed predicts satisfaction with performance on individual tasks, we compared subjects' satisfaction with their performance on individual tasks for the three tasks, shown in Table 5.

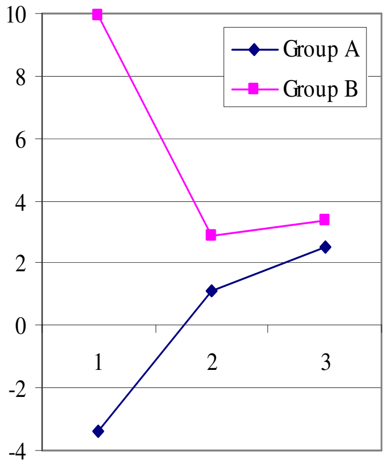
Table 5. Self-reported satisfaction of subjects with individual task performance.

	All subjects		Group A		Group B	
	Average	STDEV	Average	STDEV	Average	STDEV
Task D	-3.14	4.86	-3.40	4.55	-2.86	5.40
Task E	8.66	9.97	7.50	10.21	9.94	10.09
Task M	5.63	7.24	6.38	8.44	4.81	5.98
Over all tasks	11.15	16.33	10.48	18.42	11.90	14.64

Satisfaction with performance on E and M was significantly greater than for D (p<.0001), but no significant difference was found between satisfaction for E and M (p>.06), despite a significant (but small) difference in their performance between those two tasks. The trend is however in the right direction, suggesting that task difficulty can indeed predict individual satisfaction, in that increased difficulty results in lower individual satisfaction. There were no significant differences between groups on satisfaction ratings for individual tasks or in the overall satisfaction ratings (p>.5 for all comparisons).

6.3.4 Satisfaction with overall performance

Figure 4 shows subjects' satisfaction with overall performance. We focus on subjects' satisfaction after the second and third tasks, as both groups have had the same two (or three) tasks by then. Variants 0 and 1 of the satisfaction function would predict that Group B subjects, who had the easy task first, would have lower overall satisfaction after the second task (as a consequence of emotions wearing off). However, the trend in the graph is in the opposite direction, though the difference is not significant (p>.2). This may be due to assimilation, as modelled in Variants 2 and 3. Another feature of Figure 4 is the apparently steeper slope in change of satisfaction for Group B after the first task, compared to that for Group A. This difference also failed to reach significance (p>.2).



Overall satisfaction after		Average	STDEV
First task	Group A	-3.40	4.55
	Group B	9.94	10.09
First two tasks	Group A	1.12	5.08
	Group B	2.87	4.29
All three tasks	Group A	2.53	4.82
	Group B	3.36	5.47

Fig. 4. Satisfaction of subjects with overall performance after each task.

Comparing subjects' satisfaction with their performance on the third task on its own (Table 5) to the increase in their overall satisfaction due to the third task, it is clear that the increase is a lot smaller than it would have been if a simple addition was taking place. Similarly, for group B, the decrease in overall satisfaction due to the second task is a lot larger than subjects' dissatisfaction with their performance on this task[9]. The most obvious explanation of both these observations is that rather than summation, some kind of averaging is taking place, suggesting that Variants 1 and 3 are to be preferred to Variants 0 and 2. Combined with our earlier observation that the

[9] In contrast to the first case, this can not have been caused by emotions wearing off over time.

results conflict with the predictions of Variants 0 and 1, and favour the assimilation modelled in Variants 2 and 3, this may mean that Variant 3 models satisfaction best.

6.3.5 Comparison with Variants 1 and 3 and parameter values

To analyse this further, we investigated how well Variants 1 and 3 (the two variants that used a form of averaging) can fit the data, for varying values of δ and ε.

First, we investigated the value of δ that yields the optimal fit of Variant 1 to the experimental data. As we suspect δ to be different per subject, we compared predicted and actual satisfaction for each subject for varying values of δ. Since after one task the predicted satisfaction equals per definition the reported satisfaction, we made comparisons after two and three tasks, calculating the error rate as the sum of squares:

$$\begin{aligned}\text{ErrorRate} = \; & (\text{PredictedSat}(<\text{T}_1\text{T}_2>) - \text{ReportedSat}(<\text{T}_1\text{T}_2>))^2 + \\ & (\text{PredictedSat}(<\text{T}_1\text{T}_2\text{T}_3>) - \text{ReportedSat}(<\text{T}_1\text{T}_2\text{T}_3>))^2\end{aligned}$$

We considered calculating predicted satisfactions based on subjects' reported satisfaction with each individual task only. According to Variant 1, with Sat being the satisfaction *predicted* by the model:

$$\text{Sat}(< T_1 T_2 >) = \frac{\delta \times \text{Sat}(< T_1 >) + \text{Impact}(T_2)}{1+\delta} = \frac{\delta \times \text{Impact}(T_1) + \text{Impact}(T_2)}{1+\delta}$$

$$\text{Sat}(< T_1 T_2 T_3 >) = \frac{\delta \times \text{Sat}(< T_1 T_2 >) + \text{Impact}(T_3)}{1+\delta}$$

Assuming Impact(T) = ReportedSat(<T>) for all tasks T, this will allow us to calculate Sat(<T_1T_2>) and subsequently Sat(<$T_1T_2T_3$>).

A problem is that Sat(<T_1T_2 T_3>) and Sat(<T_1T_2>) are not independent, the former is calculated using the latter. So, any errors in the modelling early on will be propagated. Therefore, we have used an alternative method. To calculate Sat(<T_1T_2 T_3>), we use ReportedSat(<T_1T_2>) instead of Sat(<T_1T_2>). Similarly, to calculate Sat(<T_1T_2>), we use ReportedSat(<T_1>)[10]. This means that we are focusing on the extent to which the models correctly predict *changes* in satisfaction. So, we calculate predicted satisfaction as:

$$\text{PredictedSat}(< T_1 T_2 >) = \frac{\delta \times \text{ReportedSat}(< T_1 >) + \text{ReportedSat}(< T_2 >)}{1+\delta}$$

$$\text{PredictedSat}(< T_1 T_2 T_3 >) = \frac{\delta \times \text{ReportedSat}(< T_1 T_2 >) + \text{ReportedSat}(< T_3 >)}{1+\delta}$$

Figure 5 shows the error rate for subjects in both groups for varying values of δ. Variant 1 seems very bad at predicting the satisfaction of some subjects. For those subjects a higher value of δ produces a lower error rate, with δ=1 producing the best results for eleven subjects (six in group A, five in group B). What may not be clearly visible in the graphs is that for eight subjects (four in each group), δ of 0.4 or below produces the best results, confirming the suspicion that δ is user dependent.

[10] Though this does not make any difference, as PredictedSat(<T_1>)=ReportedSat(<T_1>).

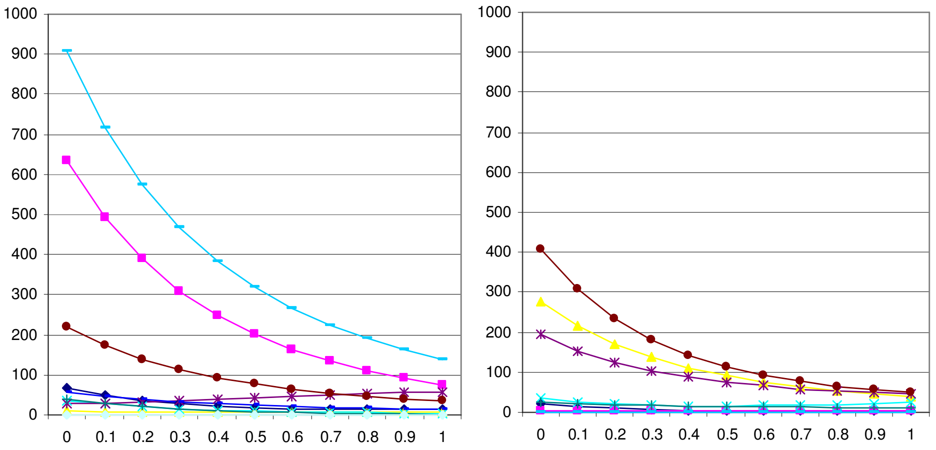


Fig. 5. Error rate as a function of δ, for δ between 0 and 1. Each line represents a subject. Graph on the left shows subjects in group A, graph on the right shows subjects in group B.

To bring the error rate down further for many subjects a value of δ greater than 1 would be required. However, this conflicts with the idea that δ models the wearing off of emotions over time. Hence, Variant 1 seems implausible. Once again, the suspicion arises that we are seeing an assimilation effect. We will investigate this below, by seeing whether ε in Variant 3 can substantially improve the fit.

Next, we will investigate which values of δ and ε produce the best fit of Variant 3 to the experimental data, using the same procedure as for Variant 1. Using the formula from Variant 3, we calculate predicted satisfaction as:

$$\text{PredictedSat}(<T_1T_2>) = \frac{\text{Mood} + \text{ReportedSat}(<T_2>) + (\text{Mood} - \text{ReportedSat}(<T_2>)) \times \varepsilon}{1+\delta}$$

$$\text{with} \quad \text{Mood} = \delta \times \text{ReportedSat}(<T_1>)$$

$$\text{PredictedSat}(<T_1T_2T_3>) = \frac{\text{Mood} + \text{ReportedSat}(<T_3>) + (\text{Mood} - \text{ReportedSat}(<T_3>)) \times \varepsilon}{1+\delta}$$

$$\text{with} \quad \text{Mood} = \delta \times \text{ReportedSat}(<T_1T_2>)$$

Table 6 shows the values of δ and ε that produce the best fit of Variant 3 to the data, and the error rate for those values. There are still some subjects with a rather high error rate, but overall Variant 3 has improved the error rate substantially. For instance, compare the error rate of subject A10 from Variant 1 (138.46) with that from Variant 3 (0.14). The assimilation parameter ε seems to be useful. So, overall Variant 3 is best. The results also confirm our suspicion that both δ and ε are user dependant.

Table 6. Values of δ and ε that produce the best fit of Variant 3 to the data, and the error rate for those values.

Group A				Group B			
Subject	**δ**	**ε**	**Error Rate**	**Subject**	**δ**	**ε**	**Error Rate**
A1	0.9	0	12.56	B1	0	1	0.06
A2	1	0.5	1.32	B2	0	0.7	1.09
A3	0.4	0	7.56	B3	0.8	0.7	0.01
A4	0	1	0.07	B5	1	0.5	13.36
A5	0	0	26.74	B6	0	0.9	11.57
A6	0.1	0.9	0.32	B7	1	0.3	34.37
A7	0.5	0.4	0.02	B8	0	0.9	1.32
A8	0.6	1	0.93	B9	0	0.3	6.29
A9	0	0.4	4.82	B10	0.4	0	0.01
A10	0.3	0.8	0.14	B11	0.1	0	0.02
A11	0.1	0.6	0.00				

6.3.6 Limitations of this experiment and future work

Showing that values of δ and ε can be found that produce a decent fit for existing data does not necessarily imply that the values generalise to new data. In a follow-up experiment, we will use longer task sequences, for instance with six tasks, determine values of δ and ε based on the first three tasks, and then investigate how well Variant 3 fits the data of the last three tasks. The best value of δ may well depend on the time duration of the tasks: if tasks are longer, emotions could wear off more. Since this experiment had a short duration, lasting approximately 15 minutes in total, we need to investigate the effect of longer task durations. We note, however, that whilst the limited task duration may have reduced emotional decay, we found δ<1 in Variant 3 to provide the best fit to the experimental data for most subjects, suggesting that emotional decay did occur. The inter-task interval will also have contributed to this. Moreover, this work is inspired by interactive TV, and the length of news items and music clips is rather limited as well, so, it is interesting for us to investigate how good our satisfaction functions perform under those conditions.

7 Group influences on satisfaction

Until now, we have assumed that an individual's satisfaction only depends upon the experiences of that individual (their liking of items and the item order). It is, however, likely that the *feelings* of others in the group may influence an individual's satisfaction. Research in organizational behavior and social psychology has highlighted the roles of *emotional contagion*, the influence of an individual's affective state on that of others in the group (Barsade, 2002; Hatfield et al., 1994), as well as *conformity*, whereby *judgments* are influenced by those of others. This potentially impacts satisfaction, an issue we turn to below.

7.1. Emotional Contagion

Effects of emotional contagion have been found in both field and laboratory studies. Totterdell et al. (1998) and Bartel and Saavedra (2000) found evidence of mood convergence within groups, and Barsade (2002) showed that emotional contagion happens for negative as well as positive emotions.

Emotional contagion does not affect everyone to the same degree (Doherty, 1997; Laird et al., 1994). Doherty proposed and validated a 15-item scale for measuring this susceptibility. Emotional contagion seems also to depend on the attention that is being paid to others (Hatfield et al., 1994), and non-verbal cues seem particularly important in "catching" others' emotions (Mehrabian, 1972). This may mean that emotional contagion is more likely to happen in a music recommendation system, than a TV news recommendation system, as people may be more aware of others in the group when their eyes are not fixed on a television screen.

Most evidence on emotional contagion points to its being a sub-conscious, automatic process (e.g. Hatfield et al., 1994). In addition, there is some evidence (as reported in Barsade, 2002) that comparison of one's emotions to those of others may motivate conscious emotion adaptation, changing the real emotion felt rather than just the display of it.

It is rather difficult to model emotional contagion, as the satisfactions of users are interdependent: the satisfaction of a user and the group she is in will be mutually reinforcing. For simplification, we will assume that contagion happens simultaneously, and that the satisfaction of each user is only influenced by the "independent" satisfaction of the other users, i.e. their satisfaction before contagion happened. So, we define the satisfaction of an individual user *u* with a sequence *items* in the context of a group *g* as the independent satisfaction of that user (Sat(*u, items*)[11]) summed with the contagion from other users' independent satisfaction:

$$\mathrm{Sat}(u, \mathit{items}, g) = \mathrm{Sat}(u, \mathit{items}) + \sum_{v \in g} \mathrm{Contagion}(\, \mathrm{Sat}(u, \mathit{items}),\ \mathrm{Sat}(v, \mathit{items})\,)$$

There seem to be two obvious ways of modelling the contagion of the satisfaction of a user *u* by the satisfaction of a user *v*. Either we assume that the contagion only depends on the other person's satisfaction, and define:

$$\mathrm{Contagion}(\mathrm{Sat}(u, \mathit{items}), \mathrm{Sat}(v, \mathit{items})) = \zeta * \mathrm{Sat}(v, \mathit{items})$$

Or we assume that it depends on a comparison between an individual's satisfaction and another user's, defining:

$$\mathrm{Contagion}(\mathrm{Sat}(u, \mathit{items}), \mathrm{Sat}(v, \mathit{items})) = \zeta * (\mathrm{Sat}(v, \mathit{items}) - \mathrm{Sat}(u, \mathit{items}))$$

[11] We will use Sat(*u*, *items*), though we only defined Sat(*items*) before. This enables us to distinguish between multiple users' satisfaction, and to have parameters (like δ and ε) dependent on the user.

The value of ζ depends on the susceptibility of the user to emotional contagion. Until now, we have assumed the user will "catch" the other users' emotions, becoming happier if the group is happier, and sadder if they are sadder. This would mean $\zeta \geq 0$. However, resentment and gloating (e.g., Ortony et al., 1988) may play a role, in which case ζ could be negative (for instance, a child may feel devastated when its siblings get a larger piece of cake)[12]. The true nature and extent of contagion is in any case likely to depend on the relationship of an individual and her cohorts.

7.2 Types of relationships

Anthropologists and social psychologists have found substantial evidence for the existence of four basic types of social relationships: communal sharing, authority ranking, equality matching, and market pricing (Fiske, 1992; Haslam, 1994). In communal sharing relationships, all group members share in the group's resources as needed and depend on one another for mutual support. Authority ranking relationships are asymmetric: one person has precedence over the other (e.g., because they are your boss, parent, older, or somebody you respect). In equality matching relationships, everyone gets the same, takes turns, independent of needs or status. In market pricing relationships a kind of bartering takes place, and tradeoffs are made (for instance, you can watch television, if you have done the washing-up first). The question is how these kinds of relationships might impact the influence others' emotions have on your own.

In a communal sharing relationship, you are likely to feel empathy with others. This empathy would mean that others' satisfaction affects your own satisfaction positively, and others' dissatisfaction affects it negatively. Therefore, for communal sharing relationships, a positive value of ζ seems most likely.

In an authority ranking relationship, you are likely to be influenced a lot by the people above you due to respect (or fear). However, depending on personality, it may also be possible to feel pity for the people below you. Therefore, for authority ranking relationships, a higher positive value of ζ seems likely for users ranking above you, and a low or close to zero positive value of ζ for users ranking below you.

In a market pricing relationship, you are likely to mainly care about yourself, whether you are getting a "good deal". This could either lead to indifference regarding others (as you only care about yourself) or to jealousy (as you compare the deal you got to the deal others got). If somebody else seems happier, you might feel that they got a better deal and feel resentment, reducing your satisfaction. Similarly, if somebody else seems less happy, your satisfaction may go up. Therefore, for market pricing relationships, a value of ζ close to zero (to model indifference) or a negative value (to model jealousy) seems most likely[13].

In an equality matching relationship, you are unlikely to have strong views on the other person. Emotional contagion could occur as it happens with strangers. Therefore, a low positive value of ζ seems most likely.

[12] Contagion may not be the best term for negative ζ, but we kept this for the sake of simplicity.

[13] Above we have modeled the contagion of an individual by a group as the summation of the contagion by others in the group. For market pricing relationships, an alternative way of modeling would be to count the number of others that "beat" us.

Given that we believe ζ to depend on both the susceptibility of the user and their relationship with each other user, we will replace ζ in the models above with ζ_{uv}. To distinguish between the two factors that influence the value of ζ_{uv}, we define:

$\zeta_{uv} = \sigma_u \times \rho_{uv}$,

where

- σ_u ($0 \leq \sigma_u \leq 1$) models the susceptibility of the user to emotional contagion, independent of the other user involved, and
- $\rho_{\mathrm{Relation}(u,v)}$ ($-1 \leq \rho_{\mathrm{Relation}(u,v)} \leq 1$) models the contribution to contagion of the relationship between users *u* and *v*.

In an ideal world, four values of ρ would suffice for all users, one for each type of relationship. A complication, however, is that relationships are not often clear-cut; they can have elements of multiple types in them and can change per day.

7.3 Conformity

Many experiments have provided evidence that humans adjust their opinions to conform with those of a group when the *majority* or all of the group expresses a different opinion than the individual originally had. For instance, Asch (1951, 1956) showed that subjects confronted with a very easy judgment task (the control group had only 0.7% errors) were influenced by the ‘obviously’ incorrect judgments of other (fake) subjects, producing 37% errors. There have been many studies showing that adding more members to the majority does not lead to a linear increase in conformity. While Asch (1951) found that a majority of three (i.e. one individual with three others who all disagreed) gave maximal conformity, later studies have shown that adding more members to the majority increases conformity, but with diminishing increments per added member (Latané & Wolf, 1981). Latané & Wolf proposed to model this effect as a power function:

$$\mathrm{Influence}(g) = \mu \times |g|^{\lambda} \qquad 0 \leq \lambda < 1$$

with μ a scaling constant that reflects the influence of a single person, and $|g|$ the number of people in group *g*.

A series of studies have found that having a social supporter, somebody whose judgment matches your own, decreases conformity substantially (Allan, 1975). It has therefore been proposed to divide the influence of the group by the number of individuals being influenced (Latané & Wolf, 1981).

Two main reasons have been given for conformity (Deutsch & Gerard, 1955). The first is *informational influence*: individuals may conform because they trust other people’s judgment more than their own. The second is *normative influence*: individuals may conform because of group pressure, wanting others to like them. Informational influence may cause individuals to change both their public and private opinions. Normative influence is more likely to change only the individuals’ public opinions, keeping their private opinions the same.

When predicting satisfaction, we are mainly interested in individuals’ real satisfaction, so using their private judgments. Therefore, informational influence is the one that is most likely to affect our predictions. Above, we defined the impact on user *u*'s satisfaction of a new item *i* given this user's pre-existing satisfaction *s* as:

$$\text{Impact}(u, i, s) = \text{Impact}(u, i) + (s\text{-Impact}(u, i)) \times \varepsilon, \qquad \text{for } 0 \leq \varepsilon \leq 1$$

To make impact reflect the informational influence of a group *g*, we define:

$$\text{Impact}(u, i, s, g) = \text{Impact}(u, i, s) + \text{InformationalInfluence}(u, g, i)$$[14]

To calculate the informational influence, we will use Latané & Wolf's (1981) proposal. Since group *g* may be composed of multiple subgroups with different opinions, the informational influence of all subgroups is summed. The informational influence of a particular subgroup is a function of (1) its influence factor (as defined by Latané & Wolf), multiplied by (2) the difference in opinion of the subgroup and the user, and divided by (3) the number of people in the group outside the subgroup. This last accounts for the social supporter effect proposed by Latané & Wolf. Hence, we define the informational influence on a user *u* of a group of others *g* for item *i* as:

$$\text{InformationalInfluence}(u, g, i) = \sum_{sg \subseteq g \,\wedge\, \text{Faction}(sg,g,i)} \frac{\mu_{u,sg} \times |sg|^{\lambda} \times (\text{Impact}(sg,i) - \text{Impact}(u,i))}{|g - sg|}$$

with Impact(*sg*, *i*) = Impact(*v*, *i*) for $v \in sg$, and

$$\text{Faction}(sg, g, i) \equiv \underset{opinion}{\exists} \left(\underset{v \in sg}{\forall}\ opinion = \text{Impact}(v,i) \wedge \underset{v \in g\text{-}sg}{\forall}\ opinion \neq \text{Impact}(v,i) \right)$$

Informational influence is likely to depend on the personality of the individual, for instance, their level of susceptibility to others' opinions (Asch, 1951; 1956). Mausner (1954) also found self-confidence and perceived competence relative to others to have an impact on informational influence in a judgment task. Thus, a would-be expert on classical music is less likely to be influenced on classical music items than somebody who sees himself as a novice. This is why we have made $\mu_{u\,sg}$ dependent on *u*. Informational influence is also likely to depend on the relationship between an individual and their group, particularly on how much the others' opinion is trusted. This is why we have made $\mu_{u\,sg}$ dependent on *sg* as well.

Although we have suggested that normative influences may not change individuals' private judgments, fake emotions expressed due to normative influences could exert contagion on the group. So, others may feel better because they wrongly believe that somebody else is happy. Above, we defined

$$\text{Sat}(u, items, g) = \text{Sat}(u, items) + \sum_{v \in g} \text{Contagion}(\,\text{Sat}(u, items),\ \text{Sat}(v, items)\,)$$

[14] We are assuming that subjects can somehow deduct other people's judgments from their reactions, separated from the mood they are in (so independent of *s*). This is a simplification, as the formulas become too complex if we have to take *s* into consideration as well.

To deal with normative influences, we change this to:

$$\mathrm{Sat}(u, \mathit{items}, g) = \mathrm{Sat}(u, \mathit{items}) + \sum_{v \in g} \mathrm{Contagion}(\, \mathrm{Sat}(u, \mathit{items}),\ \mathrm{PortrayedSat}(v, (g \cup \{u\}) - \{v\}, \mathit{items})\,)$$

As in the case of emotional contagion, it is rather difficult to model portrayed satisfaction, as the portrayed satisfactions of users are interdependent. For simplification, we will assume that the normative influence happens simultaneously, and that the portrayed satisfaction of each user is only influenced by the real satisfaction of the other users, i.e. their satisfaction before normative influence happened. We define the portrayed satisfaction of a user *u* with sequence *items* given a group of others *g* as

$$\mathrm{PortrayedSat}(u, g, \mathit{items}) = \mathrm{Sat}(u, \mathit{items}) + \mathrm{NormativeInfluence}(u, g, \mathit{items})$$

with (similar to the definition of InformationalInfluence above):

$$\mathrm{NormativeInfluence}(u, g, \mathit{items}) = \sum_{sg \subseteq g \,\wedge\, \mathrm{Faction}(sg, g, \mathit{items})} \frac{\theta_{u,sg} \times |sg|^{\kappa} \times (\mathrm{Sat}(sg, \mathit{items}) - Sat(u, \mathit{items}))}{|g - sg|}$$

with

$$\mathrm{Faction}(sg, g, \mathit{items}) \equiv \exists_{opinion} \left(\forall_{v \in sg}\ opinion = \mathrm{S}at(v, \mathit{items}) \wedge \forall_{v \in g\text{-}sg}\ opinion \neq \mathrm{S}at(v, \mathit{items}) \right)$$

and Sat(*sg*, *i*) = Sat(*v*, *i*) for $v \in sg$.

8 Experiment 2: Emotional Contagion

The goal of this experiment is to shed some light on how emotional contagion might influence the emotions of individuals in a group, and what effect the type of relationship might have.

8.1 Experimental Design

8.1.1 Method

The experiment consisted of a number of fictional situations. In each one, subjects were asked to imagine that they would be watching television with someone, and that they liked the program a little. Two within-subjects variables were manipulated: (1) the emotions of the other person, which were either *happier* or *unhappier*; (2) the type of relationship the subject had with their imaginary partner, one of (a) Authority Ranking, with the other person higher in rank, (b) Communal Sharing, (c) Market Pricing and (d) Equality Matching. The combination yielded eight fictional situations, in each of which subjects were asked to rate how the other person's emotion would influence their own. Before the tasks, subjects' susceptibility to emotional contagion was measured.

The reasons why a scenario of watching TV was selected were twofold: (a) the measure of satisfaction should be independent of abilities (so that e.g. real world knowledge about their partner's relative intelligence would not interfere), (b) the causes of satisfaction should not be transferable, that is, subjects should not be able to profit from another person's gain (as they would in a lottery). Since it seems awkward to express level of satisfaction with TV programs using numbers, vague descriptions such as "you enjoy it a little" were used.

8.1.2 Subjects

Twenty-four undergraduate students of the University of Aberdeen participated voluntarily in the experiment, which took place in a classroom setting. Subjects were predominantly male (2 female, 18 male, 4 did not disclose their gender).

8.1.3 Materials

We used an adaptation of Doherty's (1997) validated scale for measuring the subjects' susceptibility to emotional contagion. This scale consists of 15 items (three questions each related to the emotions of fear, anger, sadness, happiness and love). This was modified to exclude all items related to love, which were deemed unacceptable in a classroom setting. Each item had five answer categories ("never", "rarely", "usually", "often", "always").

To measure subjects' believed emotional contagion, we asked eight questions of the form:

> *"Think of somebody* [who meets some criterion]. *Assume you and this person are watching television together. You are enjoying the program a little. How would it make you feel to know that the other person is* [other's emotion]*?*
> *My enjoyment would"*

Each question had five answer categories ("decrease a lot", "decrease slightly", "remaining the same", "increase slightly", and "increase a lot"). Negative and positive emotional contagion was simulated by stating that the other person was *"enjoying it greatly"* or *"really hating it"*. The different relationships were distinguished using the following criteria:

- Authority Ranking: *"you respect highly (maybe your grandfather, your boss, ..)"*
- Communal Sharing: *"you share everything with (maybe your best friend)"*
- Market Pricing: *"you do deals with (like, if you do the cooking, I will do the washing up)"*
- Equality Matching: *"you are on equal footing with, you tend to get the same treatment (maybe a cousin or a class mate)"*

The questionnaire used is shown in Appendix 1.

8.1.4 Research questions

The aim of the experiment is to answer the following research questions:

- Do subjects predict emotional contagion to happen when two people are watching television together? And if so, is there a difference between positive and negative emotional contagion, or do both happen to the same extent?
- Does emotional contagion differ depending upon the relationship type? In particular, do Authority Ranking and Communal Sharing relationships indeed lead to more emotional contagion than Equality Matching and Market Pricing relationships?
- Does a Market Pricing relationship lead to a small or to an inverse emotional contagion? So, if the other person's enjoyment is higher than yours, does yours stay about the same or decrease further (and similarly, if the other person's enjoyment is lower than yours, does yours stay the same or increase further)?
- Does emotional contagion differ depending upon the susceptibility to emotional contagion, and is there an interaction with relationship type?

8.2 Results and discussion

One subject was omitted as they failed to complete the experiment, leaving 23.

8.2.1 Susceptibility to emotional contagion

We transformed the scale into numbers, using "never"=0, "rarely"=1, "usually=2", "often"=3 and "always"=4. Analysis was carried out by averaging over these numbers. Despite its not being an interval scale, this was the way Doherty used the scale, and as he has validated his scale, we used it in a similar way. Subjects' susceptibility to negative emotional contagion, was interpreted as the average score of the sadness, fear and anger questions, while the average over the happiness questions indicates susceptibility to positive contagion. To determine subjects' overall susceptibility, we averaged their susceptibility to positive and negative emotional contagion.[15] One subject did not answer the last of the fear related questions, and one subject did not answer the last of the happiness related questions. One subject answered "never" to all questions, resulting in susceptibility scores of 0. The results suggest that subjects are more susceptible to positive than to negative emotions (two-tailed t-test, p<.0001).

8.2.2 Emotional Contagion

Figure 6 shows positive emotional contagion, that is, subjects' reported changes in enjoyment when the other person was happier than they, and negative emotional contagion. To be able to analyze the data using parametric tests, we treated each answer category of question Q1-Q8 as a separate variable coded 1 or 0 depending on whether the answer was chosen or not, performing pairwise two-tailed T-tests.

[15] We could have averaged over all questions, but decided not to do this, as this would give a lot more weight to negative emotions, as we had to remove the love related questions.

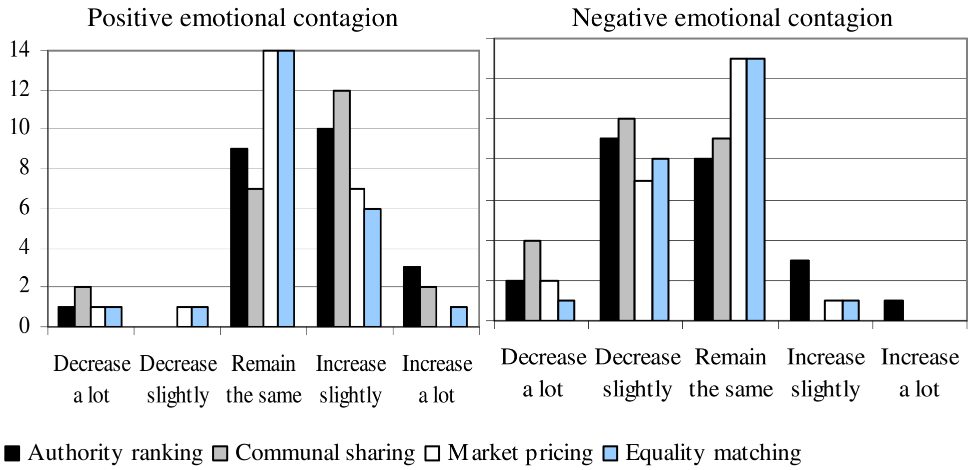


Fig. 6. Results of Experiment 2 on questions Q1-4 (Positive emotional contagion), and Q5-8 (negative emotional contagion). Y-axis shows number of subjects who replied as indicated.

The distinction between different relationships influences positive emotional contagion. Significantly more subjects responded that their enjoyment would remain the same in the Market Pricing scenario compared to Communal Sharing and Authority Ranking ($p<0.05$ for both tests). Similarly, more subjects responded that their enjoyment would remain the same in Equality Matching compared to Communal Sharing ($p<0.05$). The difference between Equality Matching and Authority ranking was not significant. Reliably more subjects reported that their enjoyment would increase in the Authority Ranking and Communal Sharing Scenarios, compared to Market Pricing and Equality Matching ($p<0.05$ for all four tests). The differences between Authority Ranking and Communal sharing were not significant. These results indicate that the positive emotions of people you respect (Authority Ranking) or love (Communal Sharing) impact your emotion more than those of people you are in a more indifferent (Equality Matching) or competitive (Market Pricing) relationship with.

The distinction between different relationships also influences negative emotional contagion. Significantly more subjects responded that their enjoyment would remain the same in the Market Pricing and Equality Matching scenarios, than Communal Sharing ($p<0.05$ for both tests). No significant differences were found with Authority ranking. Moreover, more subjects' responded that their enjoyment would decrease in the Communal Sharing scenario, compared to Market Pricing and Equality Matching ($p<0.05$ for both tests). The differences between Authority Ranking and the other scenarios were, however, not significant.

We found no real difference between negative and positive contagion. Despite our subjects, according to Doherty's scale, being more susceptible to positive contagion, we did not see any difference in the results. One possible explanation is that the difference between "you enjoy a little" and "the other person is enjoying it greatly" is smaller than the difference between "you enjoy a little" and "the other person really hates it". So, according to our proposed models, we would expect more contagion in the latter case, and this could counterbalance the smaller susceptibility. Another possible explanation is that subjects are bad at predicting the size of a change in emotions (as confirmed by the affective forecasting literature discussed above), though they are good at predicting the direction of a change in emotions (so, our analysis of the relationship impact remains valid). In addition, the scale might not have given enough options to make small distinctions in the size of the change.

8.2.3 Effect of susceptibility on emotional contagion

We next investigated whether the subjects' susceptibility to emotional contagion (as determined using Doherty's scale) affected the reported emotional contagion. Figure 7 shows the results of negative emotional contagion in two groups: one of subjects highly susceptible to negative emotional contagion, and one with the remaining subjects. The highly susceptible group contained the 11 subjects whose susceptibility to negative emotions was above 1.5, meaning that on average they replied "usually" (which had a value of 2). At a glance, a higher percentage of subjects who were highly susceptible indeed showed negative contagion happening. The only substantial difference, however, is for the Communal Sharing relationship.

A similar division into High and Low susceptibility groups for positive emotional contagion proved harder, since almost all subjects were highly susceptible, if we were to use the same cut-off point as before. We would be left with only 5 subjects in the Low group, which is too few to reason sensibly with percentages. We would have to raise the cut-off point to just above 2, in order to get sensible group sizes. However, this seems rather arbitrary, particularly as six subjects have a score of exactly 2, and it seems hard to defend an average of "usually" as indicating low susceptibility.

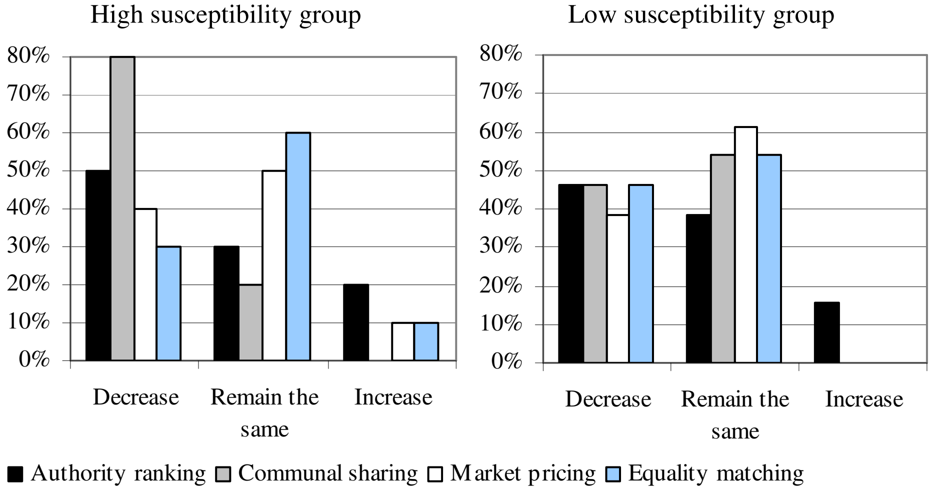


Fig. 7. Percentage of subjects who answered that their enjoyment would decrease, remain the same, or increase for questions Q5-Q8 for High and Low susceptibility groups.

8.2.4 Answers to research questions

From the analysis above, we find the following answers to our research questions:

- *Occurrence of emotional contagion*. Most subjects do think that emotional contagion will happen, but only for certain situations (see below).
- *Difference between positive and negative contagion*. We did not find a difference, but this may be due to the experimental setup (as explained above).
- *Effect of relationship type*. There is a clear effect of relationship type on emotional contagion. We did indeed find that Authority Ranking and Communal Sharing relationships lead to more emotional contagion than Equality Matching and Market Pricing relationships.
- *Market pricing*. Hardly any subjects showed inverse contagion, and the data seems to point more to a very small positive value of $\rho_{\text{Market pricing}}$ than a negative one.
- *Effect of susceptibility*. There is some indication that susceptibility to emotional contagion has an effect for Communal Sharing relationships. Given that $\rho_{\text{Market pricing}}$ and $\rho_{\text{Equality matching}}$ seem close to zero, it would have been hard to find an effect for those (as susceptibility is multiplied with ρ in our models). We do not have enough data to draw strong conclusions here, but it seems wise to leave the susceptibility parameter in the model.

9 Embarrassment and the issue of privacy

Lorrie Cranor, in her invited talk at UM2005, mentioned the case of a user whose Tivo caused him considerable embarrassment when it wrongly assumed he was gay (having observed him watching gardening and cooking programs), and started recording gay programs for him (Zaslow, 2002). In a group recommendation situation, this would be worse. Imagine your embarrassment if the TV started showing erotic items to you and your friends, because it has wrongly assumed you are interested in them. Particularly if, for the sake of transparency, it explained that it was showing them, because of *your* interests. This is not merely a question of the system making wrong inferences, since it is possible that a user is interested in such items, but would rather their family or friends did not know. So, while the issue of privacy is regarded as increasingly important in the user modeling community (Kobsa & Cranor, 2005), it becomes additionally important when recommendations are made to groups.

To improve privacy in a group recommender, each item could have a privacy measure associated with it, in addition to a rating. This measure could be modeled as a number between 0 and 1, with 1 meaning complete disclosure (the user does not care whether others know this rating) and 0 meaning complete privacy (the rating can only be used for personal recommendations). The group aggregation strategies could take this measure into account (e.g. attaching less weight to ratings the user wants to keep private).

A complication is that privacy is not only important to hide facts that are inherently embarrassing (like somebody wanting to hide their interest in building explosives). When discussing conformity above, we mentioned normative influences: people changing their judgment to fit in with the group, to be liked. So, teenagers might want to hide far more innocent things, such as a preference for a particular band, when it is clear that all their friends detest it. When a recommendation is made to a group, we have to ensure that individuals still have this option to conform, to hide the fact that they are different. One way to achieve this may be to let users have separate ratings depending on the circumstances. For instance, the user could have a rating for private use, one for their family, another for their friends, etc, allowing them to have different personas in different settings as advocated by various privacy researchers (e.g. Kay et al., 2003). This is backed up by a world-wide-web poll which showed that almost 60% of users would value being able to assume different aliases/roles on the Internet (GVU 1998, as reported in Kobsa & Schreck, 2003). An advantage of the multiple personas idea is that it extends to implicit modeling: when observing user actions, only the user model of the active persona could be updated.

A limitation of the different personas approach is that humans are fallible, and it is quite possible that a user would occasionally forget to switch from his family persona to his private persona, when starting to watch some late night programs. Some seemingly obvious ways to prevent this unfortunately do not work:

- Always using a private persona when the user is alone would not give the group recommender any implicit data about individual user preferences to work from. A possible workaround may be to have the user give explicit permission to transfer data between user profiles, but this would require quite some effort from the user as it needs doing regularly in order to keep the profile up to date.
- It is cumbersome to ask users which profile they would like to use whenever the group changes (assuming this is automatically detectable). Moreover, switching personas may provoke suspicion ("my husband is hiding his tv preferences from me, so he must be a pervert or a maniac").

Another limitation of using different personas is that an individual might not know beforehand what the opinion of the others will be, making it harder to conform.

Because of these two limitations, we will assume that even when users are using multiple personas, they may occasionally want to hide the preferences of their persona. This means that scrutability in a group recommender with varying groups should never go as far as divulging the details of the individuals' ratings to the group. In the movie group recommender system PolyLens, where inferred ratings are optionally visible to the group, 93% of users preferred to share their ratings with the group, suggesting that utility outbalances privacy (O'Connor et al., 2001). However, PolyLens is a special case, as users would normally be members of only one group, and that group tends to be based on a similar taste in movies or a close friendship.

We will use anonymity: when telling the group why a certain recommendation was made, individual ratings and identities are not divulged, avoiding such explanations as "This was recommended because Pete really likes it". Instead, we could use explanations like "This was recommended because one of you really likes it". However, in smaller groups, and particularly if groups vary regularly, it may still be quite easy to figure out who is meant (e.g. one might notice that whenever Pete is present, children's movies are recommended, or the Rolling Stones are not played). Knowing that somebody really likes something would allow a group to apply social pressure or to use deduction to find out who the culprit is.

It therefore seems wise to hide the existence of strong individual preferences that go against the majority of the group. At least we should avoid explanations like "This was recommended because one of you really likes it". We assume that users will know what group aggregation strategy is being used, to make the system more transparent. The question then arises whether the aggregation strategies differ on how much they protect privacy. For instance, the Least Misery Strategy selects items nobody really hates. So, if you notice that an item you really like is not selected, and perhaps note that many other people in the group seem to regret its omission, you may get suspicious that somebody else hates it. The Average Strategy, by its nature, would only give the information that most people probably like or dislike an item. So, we assume that the Average Strategy will protect privacy better. In theory, it seems that Multiplicative Utilitarianism (the strategy which performed best keeping users satisfied in Masthoff, 2004a) and Average Without Misery would protect privacy somewhere between the Least Misery and the Average strategy. In Experiment 3 below, we will investigate users' views on this. The protection of privacy could then be an important aspect of deciding which aggregation strategy to use.

In order to disguise user preferences further, we could add an additional, virtual member to the group, clearly a different one for each group. This could be an embodied agent presenting the recommendations. Users would be told that the virtual member has tastes of its own, and that these tastes will influence the recommendations as well, making it harder to attribute tastes to individuals, as they could always blame the virtual character (who could be quite vocal about some of its tastes). It would also add a bit of serendipity to the system. Rather than always showing the group what they clearly will like, items can be thrown in for which it is unknown whether they will be liked. The virtual member could also impact conformity, since normative influences diminish when another member of the group is shown to share your opinion. An individual with a taste deviating from the group might feel less lonely, and more likely to express an opinion, though this depends on them trusting the virtual member.

10 Experiment 3: Privacy of Group Aggregation Strategies

In this experiment, we investigate the impact different group aggregation strategies may have on privacy. We are mainly interested in the extent to which users may think a group aggregation strategy could betray their tastes to others, assuming they will be most anxious to hide strong opinions (that they really like or really hate an item) and opinions that deviate from those of others in the group.

We assume that a group recommender potentially betrays someone's taste if from a recommendation given (in particular, from the inclusion or absence of an item in the recommendation) others in the group are quite sure that somebody in the group likes or hates the item. Note that this does not necessarily betray *who* likes or hates the item (unless the group has only two members). That information could be obtained through social pressure, or knowledge gained from different (sub)groups. In this experiment, we will investigate to what extent users *feel* sure about the existence of somebody in the group with a taste opposite to their own. Note that "feel sure" is different from "are sure". We could of course theoretically analyze the aggregation strategies, to determine which strategy *provides* most privacy, but it seems equally important to determine which strategy is *regarded* as providing most privacy.

10.1 Experimental Design

Twelve graduate students and staff of the University of Aberdeen's computing science department participated in the experiment. Subjects were predominantly male (11 male, 1 female). We used a within-subjects design. Subjects were told to assume that they were going to listen to music together with a group of other people, that the DJ had everyone's opinions on songs as ratings between 1 (really hate) to 10 (really like), and that the DJ would use the individual people's ratings to calculate a rating of each song for the group. As a within subject variable, four hypothetical situations were sketched, each corresponding to a group aggregation strategy. All situations (with a small exception, see below) followed the format:
"The DJ has calculated ratings for the group by [group aggregation strategy].
1. A song your really like has not been played.
- *How sure are you that somebody in the group hates it?*
- *How sure are you that most of the group hates it?*

2. A song you really hate has been played.
- *How sure are you that somebody in the group likes it?*
- *How sure are you that most of the group likes it?"*

We used a 7-point Likert scale, ranging from "not sure at all" to "extremely sure". From the group aggregation strategies described in (Masthoff, 2004a), the four we used were:

- Average: *"averaging all individual ratings."*
- Least Misery: *"taking the minimum of individual ratings."*
- Multiplicative Utilitarian: *"multiplying the individual ratings."*
- Average Without Misery: *"removing all songs anybody really hated and averaging the individual ratings for the remaining songs."*

A slight modification to the question was made for Average Without Misery. As this strategy makes it impossible for an item you really hate to be selected, we removed the word "really" from the sentence "A song you really hate has been played." See Appendix 2 for the exact wording of the questions.

10.2 Results and Discussion

It should be noted that none of the aggregation strategies should really produce certainty on any of the questions. We left the description of the situation purposely vague: subjects did not know anything about the ratings given by others (we had no choice on this, given the questions we were going to ask). Any of the strategies could result in a song an individual really liked (or hated) not being played (or being played) *without* anybody in the group having the opposite opinion. For instance, if everybody in the group liked all the songs, then a lowest rating of 7 (somebody still quite liking it) might result in it not being played. Therefore, theoretically subjects could have answered 1 (not sure at all) to all questions. However, as would happen in a real life situation, subjects seemed to assume that everybody would have some songs they liked and some they hated. Also, their answers may have been relative: indicating that they thought a statement more likely to be true for one strategy than for another.

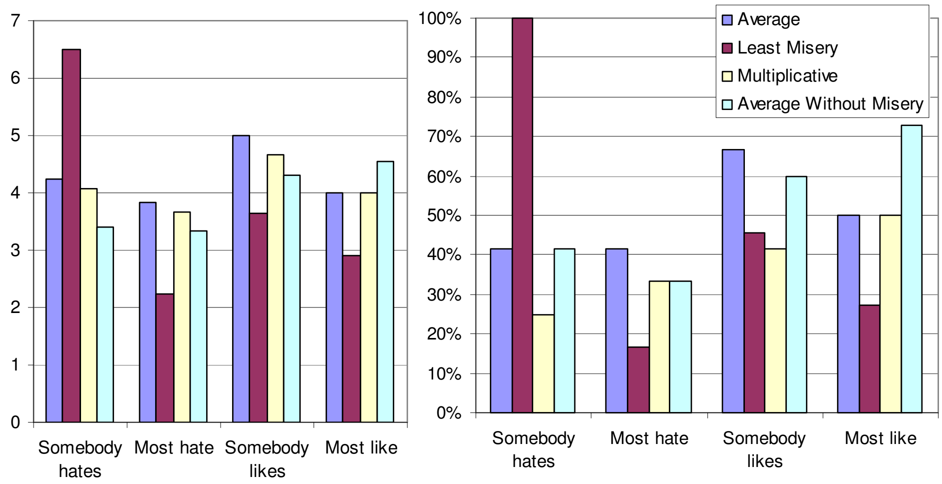


Fig. 8. Left-hand side: Average over subjects of the certainty scores (1=not sure at all, 7=extremely sure) Right-hand side: Percentage of subjects being certain (score of 5 or above).

Two subjects were puzzled by the "a song you hate has been played" questions for Average Without Misery. They did not answer these, stating "This can't happen" and "DJ cheated". One of these subjects also did not answer the "a song you really hate has been played" questions for the Least Misery strategy. They are right that these situations would be unlikely to happen, though both situations can theoretically occur.

Figure 8 shows the average over subjects of the certainty scores, and the percentage of subjects who were sure (i.e., had a score ≥ 5). The only strategy that clearly seems to evoke different responses is Least Misery: *all* subjects were sure that when a song they really liked had not been played somebody hated it, with an average score of 6.5. They also were very unsure that most hated it (average of 2.25). This is exactly the kind of situation where tensions in a group might arise: an individual being upset because one of his favorites has been skipped, being extremely sure that somebody else hated it (and therefore caused the skipping), and not believing that everybody else hated it (so, feeling free to prosecute the "sinner"). Thus, Least Misery seems bad at protecting privacy. Although the inferences being made may well be wrong, they can cause distress when they are right (as they often will be).

We assumed in the previous section that the Average Strategy would protect privacy better than the Multiplicative and the Average Without Misery strategies, as the latter both prevent misery. This was not confirmed by the results. The trend was for the Multiplicative Strategy to perform slightly (but not significantly) better than the Average Strategy. This shows that users are not very good at grasping the implications of the strategies. As one subject commented "I have no intuitive feeling for the difference between averaging and multiplying when non-zero numbers are involved". This corresponds with findings by a.o. Tversky and Kahneman (as discussed in Carofiglio & de Rosis, 2001) on how bad people are in probabilitistic reasoning. Though one may expect our subjects to be better at this than the average population, given their computing science background, Barwise (as discussed in Carofiglio & de Rosis, 2001) found that even mathematicians are not better at such tasks.

Overall, we conclude that Least Misery should be avoided for privacy reasons, but that it is fine to use the Multiplicative strategy.

11 Conclusions

There has been little work so far on computational models of affective state that can predict (rather than measure) how the user of a recommender system will be feeling. González et al (2004) have worked on the modelling of affective state for a recommender system, but their model is mainly based on an emotional intelligence test combined with updating of emotional values based on feedback to recommendations. There has been more research on the modelling of the affective state of artificial agents, where emotions are often related to plan completion and goal attainment, and modelling often happens using belief networks (e.g. de Rosis et al., 2003; Grath, 2000). Users in a recommender system do not really have plans (unless it is to be entertained or informed), so this research does not tend to generalize to recommender systems. This paper tries to fill the gap. Modelling affective state is particularly important when adapting to groups of users, as a group adaptation system will not be able to please all of its users all of the time. In such a case, an accurate prediction of the affective state of individual users can be helpful to prevent any user becoming too dissatisfied. Models of affective state can also be used to evaluate group aggregation strategies (Masthoff, 2004a).

We have proposed four different functions for modelling user satisfaction in a group recommender system that recommends sequences of items, incorporating assimilation and decay of emotions with time.

The evaluation of models of affective state is particularly difficult. In this paper, we have explained the issues involved, and have shown how evaluation is possible via indirect experiments, simulations, and experiments in another domain. In the process, we have found that Variant 3 of our satisfaction functions seems to perform best, and confirmed that time decay and assimilation parameters (δ and ε) depend on the user.

Unfortunately, modelling affective state becomes additionally complex when dealing with groups, because of members of the group influencing each other's emotions, via emotional contagion and conformity. We have shown how this can be incorporated into the models. We have also shown that the type of relationship the user has with others is important to take into account. In particular, Authority Ranking and Communal Sharing relationships seem to evoke more emotional contagion then Market Pricing and Equality Matching relationships.

Models of individual satisfaction are not only useful when adapting to groups. As argued in (Masthoff, 2003; 2004b), adaptation to individuals can sometimes also benefit from group aggregation strategies, for instance, when ratings on multiple criteria need to be combined, or when virtual group members are added, e.g. representing a teacher.

Models of affective state are not only useful for recommender systems, but also for other adaptive systems such as intelligent tutoring systems. While initially most of the focus in intelligent tutoring systems was on modelling of and adapting to cognitive aspects of the learner, gradually interest has increased in modelling of and adapting to a learner's affective state (see for instance Picard et al., 2004; del Soldato, 1994; de Vincent, 2003). Wosnitza & Volet (2005) discuss the importance of human teachers having access to their students' affective states. Understanding a student's affective state can allow an intelligent tutoring system to adapt its instruction, material selection and feedback to keep the learner sufficiently motivated and challenged; it can also be used to teach learners how to deal with failure and frustration (Burleson & Picard, 2004). While existing work tends to *measure* affective state, rather than calculating it, and this can be done theory-neutrally (Picard et al., 2004), a theory *is* required to *predict* changes in affective state. For instance, if you want to select the next exercise for a student to do, it would be good if you could predict how the student would be feeling after each possible exercise. We believe that the work presented in this paper can be adapted to do this. It would also avoid the need for continuous self-reporting which interrupts the learning experience and the need for intrusive and costly sensors.

A lot of work on group modelling has been on modelling common knowledge between group members (e.g. Introne & Alterman, 2006; Suebnukarn & Haddawy, 2006), modelling how a group interacts (e.g. Read et al, 2006; McLaren et al., 2006) and group formation based on individual models (e.g. Read et al., 2006; Alfonseca et al., 2006). The affective state of the group members is normally not taken into account. It would be interesting to explore how affective state can be used as part of group formation (or group break up), and to model not just how individuals are contributing and how their knowledge deviates from the group's, but also how they are likely to feel because of this.

One might think that accurate predictions of individual satisfaction can also be used to improve the transparency of adaptive systems: showing how satisfied others in your group are, or how satisfied criteria are, could improve the users' understanding of the working of the system and perhaps make it easier to accept items they do not like. This may be a good idea for a system that adapts to individual users. However, in a group adaptation system, users' need for privacy is likely to conflict with their need for transparency. An important task of a group adaptation system is to avoid embarrassment. Users often like to conform with the group to avoid being disliked. We modelled this normative conformity as part of our modelling of how others in the group can influence individual affective state. We may be able to use this to predict how embarrassed a user would be with disclosure of their judgement, and base our explanation of system recommendations on this. We have also investigated how different group aggregation strategies may affect privacy.

This paper has discussed many of the issues involved in modelling affective state in recommender systems. We hope it will inspire a lot more research into this area.

Acknowledgements

Judith Masthoff's research is partly supported by Nuffield Foundation Grant No. NAL/00258/G. Albert Gatt's research is supported by EPSRC Grant No. GR/S13330/01. We would like to thank the anonymous reviewers for their constructive comments.

References

Alfonseca, E., Carro, R.M., Martín, E., Ortigosa, A. and P. Paredes: 2006, 'The Impact of Learning Styles on Student Grouping for Collaborative Learning: A Case Study', in this issue.

Allan, V. L.: 1975, 'Social Support for Non-Conformity'. In: L. Berkowitz (ed.), *Advances in Experimental Social Psychology*. Vol. 8. New York: Academic Press, pp. 1-43.

Asch, S.E.: 1951, 'Effects of Group Pressure on the Modification and Distortion of Judgements'. In: H. Guetzkow (ed.): *Groups, Leadership and Men*. Pittsburgh, PA: Carnegie Press, pp. 177-190.

Asch, S.E.: 1956, 'Studies of Independence and Conformity: A Minority of one against a Unanimous Majority'. *Psychological Monographs* **70** (Whole no. 416).

Aston, G. and L. Burnard: 1998, 'The BNC Handbook: Exploring the British National Corpus with SARA'. Edinburgh, UK: Edinburgh University Press.

Aylesworth, A.B. and S.B. MacKenzie: 1998, 'Context is Key: The Effect of Program-Induced Mood on Thoughts about the Ad'. *Journal of Advertising* **27**, 17-33.

Balota, D.A., Cortese, M.J., Hutchison, K.A., Neely, J.H., Nelson, D., Simpson, G.B., and R. Treiman: 2002, 'The English Lexicon Project: A Web-Based Repository of Descriptive and Behavioral Measures for 40,481 English Words and Nonwords'. http://elexicon.wustl.edu/, Washington University.

Barsade, S.G.: 2002, 'The Ripple Effect: Emotional Contagion and its Influence on Group Behavior'. *Administrative Science Quarterly* **47**, 644-675.

Bartel, C.A. and R. Saavedra: 2000, 'The Collective Construction of Workgroup Moods'. *Administrative Science Quarterly* **45**, 197-231.

Burleson, W. and R.W. Picard: 2004. 'Affective Agents: Sustaining Motivation to Learn through Failure and a State of Stuck'. In: C. Frasson and K. Porayska-Pomsta (eds.): *ITS04 Workshop on Social and Emotional Intelligence in Learning Environments*, Maceio - Alagoas, Brasil. Online Proceedings: http://www.cogsci.ed.ac.uk/%7Ekaska/WorkshopSI/

Carofiglio, V. and F. de Rosis: 2001, 'Exploiting Uncertainty and Incomplete Knowledge in Deceptive Argumentation'**.** *International Conference on Computational Science*, San Francisco, CA, pp. 1019–1030.

Chin, D.: 2005, 'Addressing Problems in the first Adaptive System Challenge'. In: S. Weibelzahl, A. Paramythis, and J. Masthoff (eds.): *UM05 Workshop on the Evaluation of Adaptive Systems*, Edinburgh, UK, pp 74-78. Online Proceedings:
http://www.easy-hub.org/hub/workshops/um2005/proceedings.html

Deutsch, M. and H.B. Gerard: 1955, 'A Study of Normative and Informational Social Influences upon Individual Judgment'. *Journal of Abnormal and Social Psychology* **51***,* 629-636.

Doherty, R.W.: 1997, 'The Emotional Contagion Scale: A Measure of Individual Differences'. *Journal of nonverbal Behavior* **21**, 131-154.

Elliot, C. and G. Siegle: 1993, 'Variables Influencing the Intensity of Simulated Affective States'. *AAAI Spring Symposium on Reasoning about Mental States: Formal Theories and Applications*, Menlo Park, CA, pp. 58-67.

Fiske, A.P.: 1992, 'The four Elementary Forms of Sociality: Framework for a Unified Theory of Social Relations'. *Psychological Review* **99**, 689-723.

Forster, K.I. and J.C. Forster: 2003, 'DMDX: A Windows Display Program with Millisecond Accuracy'. *Behavioral Research Methods, Instruments and Computers* **35**, 116-124.

Gardner, M.: 1985, 'Mood States and Consumer Behavior: A Critical Review'. *Journal of Consumer Research* **12**, 281-300.

González, G., López B. and J.LL. de la Rosa: 2004, 'Managing Emotions in Smart User Models for Recommender Systems'. *Sixth International Conference on Enterprise Information Systems*, Porto, Portugal, pp. 187-194.

Grath, J.: 2000, 'Émile: Marshalling Passions in Training and Education'. *Fourth International Conference on Autonomous Agents*, Barcelona, Spain, pp. 325-332.

Gross, J.J. and R.W. Levenson: 1995, 'Emotion Elicitation Using Films'. *Cognition & Emotion* **9**, 87-108.

Harley, T.: 2001, 'The Psychology of Language: From Data to Theory'. Hove, UK: Psychology Press.

Haslam, N.: 1994, 'Categories of Social Relationship'. *Cognition* **53**, 59-90.

Hatfield, E., Cacioppo, J. and R.L. Rapson: 1994, 'Emotional Contagion'. New York: Cambridge University Press.

Heylen, D., Nijholt, A., op den Akker, R. and M. Vissers: 2003, 'Socially Intelligent Tutor Agents'. In: T. Rist, R. Aylett, D. Ballin and J. Rickel (eds.): *The 4th International Working Conference on Intelligent Virtual Agents*, Kloster Irsee: Germany, ,LNAI 2792. Berlin: Springer Verlag, pp341-347.

Introne, J. and R. Alterman: 2006, 'Using Shared Representations to Improve Coordination and Intent Inference', in this issue.

Isbell, L.M., Ottati, V.C. and K.C. Burns: 2003, 'Affect and Politics: Effects on Judgment, Processing, and Information Selection'. As accessed on 16 October 2004.
http://www.cbrss.harvard.edu/events/ppbw/papers/isbell.pdf

Isen, A. M., Shalker, T., Clark, M. and L. Karp: 1978, 'Affect, Accessibility of Material in Memory and Behavior: A cognitive loop?' *Journal of Personality and Social Psychology* **36**, 1-12.

Järvenoja, H. and S. Järvelä: 2005, 'How Students Describe the Sources of their Emotional and Motivational Experiences during the Learning Process: A Qualitative Approach'. *Learning and Instruction* **15**, 465-480.

Kay, J., Kummerfeld, R.J. and P. Lauder: 2003, 'Managing Private User Models and Shared Personas'. In K. Cheverst, B. de Carolis and A. Krüger (eds.): *UM03 Workshop on User Modelling for Ubiquitous Computing*, Johnstown, PA. Online Proceedings:
http://www.di.uniba.it/~ubium03/apapers.html

Kobsa, A. and L. Cranor (eds.): 2005, 'UM05 Workshop on Privacy-Enhanced Personalization', Edinburgh, UK. http://www.isr.uci.edu/pep05/papers/w9-proceedings.pdf

Kobsa, A. and J. Schreck: 2003, 'Privacy through Pseudonymity in User-Adaptive Systems'. *ACM Transactions on Internet Technology* **3**, 149-183.

Kort, B., Reily, R. and R. Picard: 2001, 'External Representation of Learning Process and Domain Knowledge: Affective State as a Determinate of its Structure and Function'. In: S. Ainsworth et al. (eds.): *AI-ED01 Workshop on External Representations in AIED: Multiple Forms and Multiple Roles*, San Antonio, TX. Online Proceedings:
http://www.psychology.nottingham.ac.uk/research/credit/AIED-ER/contributions.html

Laird, J.D., Alibozak, T., Davainis, D., Deignan, K., Fontanella, K., Hong, J., Levy, B. and C. Pacheco: 1994, 'Individual Differences in the Effects of Spontaneous Mimicry on Emotional Contagion'. *Motivation and Emotion* **18**, 231-247.

Mehrabian, A.: 1972, 'Nonverbal Communication'. Chicago, IL: Aldine-Atherton.

Latané, B. and S. Wolf: 1981, 'The Social Impact of Majorities and Minorities'. *Psychological Review* **88**, 438-453.

Mackie, D.M. and L.T. Worth: 1989, 'Processing Deficits and the Mediation of Positive Affect in Persuasion'. *Journal of Personal and Social Psychology* **57**, 1-14.

Masthoff, J.: 2003, 'Modeling the Multiple People that Are Me'. In: P. Brusilovsky, A. Corbett and F. de Rosis (eds.): *User Modeling 2003*: *9th International Conference*, Johnstown, PA, LNAI 2702. Berlin: Springer Verlag, pp. 258-262.

Masthoff, J.: 2004a, 'Group Modeling: Selecting a Sequence of Television Items to Suit a Group of Viewers'. *User Modeling and User Adapted Interaction* **14**, 37-85.

Masthoff, J.: 2004b, 'Selecting News to Suit a Group of Criteria'. In: L. Ardissono and M. Maybury (eds.), *AH04 Workshop on Personalization in Future TV: Methods, Technologies* *and Applications*, Eindhoven, the Netherlands, pp. 252-263. Online Proceedings: http://www.di.unito.it/~liliana/TV04/schedule.html

Masthoff, J.: 2005, 'The Pursuit of Satisfaction: Affective State in Group Recommender Systems'. In: L. Ardissono, P. Brna and A. Mitrovic (eds.): *User Modeling 2005: 10th International Conference*, Edinburgh, UK, LNAI 3538. Berlin: Springer Verlag., pp. 297-306.

Mausner, B.: 1954, 'Prestige and Social Interaction. The Effect of one Partner's Success in a Relevant Task on the Interaction of Observer Pairs'. *Journal of Abnormal and Social Psychology* **49**, 557-560.

McLaren, B.M., Walker, E., Harrer, A., Bollen L. and J. Sewall: 2006, 'Creating Cognitive Tutors for Collaborative Learning: Steps toward Realization', in this issue.

Meloy, M.: 2000, 'Mood-Driven Distortion of Product Information'. *Journal of Consumer Research* **27**, 345-359.

Murry, J., Lastovicka, J. and S. Singh: 1992, 'Feeling and Liking Responses to Television Programs: An Examination of two Explanations for Media-Context Effects'. *Journal of Consumer Research* **18**, 441-451.

Oatley, K. and J.M. Jenkins: 1996, 'Understanding Emotions'. Malden, MA: Blackwell.

O' Conner, M., Cosley, D., Konstan, J.A. and J. Riedl: 2001, 'PolyLens: A Recommender System for Groups of Users'. *Seventh European Conference on Computer Supported Collaborative Work*, Bonn, Germany, pp 199-218.

Ortony, A., Clore, G.L. and A. Collins: 1988, 'The Cognitive Structure of Emotions'. Cambridge, UK: Cambridge University Press.

Picard, R.W.: 1997, 'Affective Computing'. Cambridge, MA: MIT Press.

Picard, R.W., Papert, S., Bender, W., Blumberg, B., Breazeal, C., Cavallo, D., Machover, T., Resnick, M., Roy, D. and C. Strohecker: 2004, 'Affective Learning: a Manifesto'. *BT Technology Journal* **22**, 253-269.

Redelmeier, D. A., Katz, J. and D. Kahneman: 2003, 'Memories of Colonoscopy: A Randomized Trial'. *Pain* **104**, 187-194.

Read, T., Barros, B., Bárcena, E. and J. Pancorbo: 2006, 'Coalescing Individual and Collaborative Learning to Model User Linguistic Competences', in this issue.

de Rosis, F, Pelachaud, C., Poggi, I., Carofiglio, V. and B. de Carolis: 2003, 'From Gretna's Mind to her Face: Modelling the Dynamics of Affective Status in a Conversational Embodied Agent'. *International Journal of Human-Computer Studies* **59**, 81-118.

Rottenberg, J., Ray, R.R. and J.J. Gross: in press, Emotion Elicitation Using Films. In: J.A. Coan and J.J.B Allen (eds.): *The Handbook of Emotion Elicitation and Assessment*. New York: Oxford University Press. As accessed on 16 October 2004,
http://www-psych.stanford.edu/%7Epsyphy/Pdfs/chapter_2_films_final.pdf

Schumann, D. and E. Thorson: 1990, 'The Influence of Viewing Context on Commercial Effectiveness: A Selection-Processing Model'. *Current Issues and Research in Advertising* **12**, 1-24.

Schwarz, N. and G.L. Clore: 1996, 'Feelings and Phenomenal Experiences'. In: E.T. Higgins and A. Kruglanski (eds.): *Social Psychology: Handbook of Basic Principles*. New York: The Guilford Press, pp. 433-465.

del Soldato, T.: 1994, 'Motivation in Tutoring Systems'. *Technical Report CSRP303*, School of Cognitive and Computing Sciences, University of Sussex, UK.

Suebnukarn, S. and P. Haddawy: 2006, 'Modeling Individual and Collaborative Problem-Solving in Medical Problem-Based Learning', in this issue.

Totterdell, P, Kellet, S., Teuchmann, K. and R.B. Briner: 1998, 'Evidence of Mood Linkage in Work Groups'. *Journal of Personality and Social Psychology* **74**, 1504-1515.

de Vicente, A.: 2003, 'Towards Tutoring Systems that Detect Students' Motivation: An Investigation'. *Ph.D. thesis, School of Informatics, University of Edinburgh*, UK.

Weiner, B.: 1995, 'Judgments of Responsibility: a Foundation for a Theory of Social Conduct'. New York: The Guilford Press.

Wilson, M.D.: 1988, 'The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2', *Behavioural Research Methods, Instruments and Computers* **20**, 6-11.

Wilson, T.D. and D.T. Gilbert: 2003, 'Affective Forecasting'. *Advances in Experimental Social Psychology* **35**, 345-411

Wilson, T.D., Gilbert, D.T. and D.B. Centerbar: 2003, 'Making Sense: The Causes of Emotional Evanescence'. In: I. Brocas and J. Carrillo (eds.): *The Psychology of Economic Decisions, Vol. 1: Rationality and Well Being.* New York: Oxford University Press, pp. 209-233.

Wilson, T.D. and K. Klaaren: 1992, 'The Role of Affective Expectations in Affective Experience'. In: M.S. Clark (ed.): *Review of Personality and Social Psychology, Vol. 14: Emotion and Social Behaviour*, Newbury Park, CA: Sage, pp. 1-31.

Wosnitza, M. and S. Volet: 2005, 'Origin, Direction and Impact of Emotions in Social Online Learning'. *Learning and Instruction* **15**, 449-464.

Zaslow, J.: 2002, 'If TiVo Thinks you are Gay, Here's How to Set it Straight'. *Wall Street Journal, 26 November 2002*.

Appendix 1: Questionnaire for Experiment 2

Gender: male / female

If someone I'm talking with begins to cry, I get teary-eyed.
□ never □ rarely □ usually □ often □ always

Being with a happy person picks me up when I'm feeling down.
□ never □ rarely □ usually □ often □ always

When someone smiles warmly at me, I smile back and feel warm inside.
□ never □ rarely □ usually □ often □ always

I get filled with sorrow when people talk about the death of their loved ones.
□ never □ rarely □ usually □ often □ always

I clench my jaws and my shoulders get tight when I see angry faces on the news.
□ never □ rarely □ usually □ often □ always

It irritates me to be around angry people.
□ never □ rarely □ usually □ often □ always

Watching the fearful faces of victims on the news makes me try to imagine how they might be feeling.
□ never □ rarely □ usually □ often □ always

I tense when overhearing an angry quarrel.
□ never □ rarely □ usually □ often □ always

Being around happy people fills my mind with happy thoughts.
□ never □ rarely □ usually □ often □ always

I notice myself getting tense when I'm around people who are stressed out.
□ never □ rarely □ usually □ often □ always

I cry at sad movies.
□ never □ rarely □ usually □ often □ always

Listening to the shrill screams of a terrified child in the dentist's waiting room makes me feel nervous.
□ never □ rarely □ usually □ often □ always

Q1. Think of somebody you respect highly (maybe your grandfather, your boss, ..). Assume you and this person are watching television together. You are enjoying the program a little. How would it make you feel to know that the other person is enjoying it greatly? My enjoyment would

□ decrease	□ decrease	□ remain the same	□ increase	□ increase
a lot	slightly		slightly	a lot

Q2. Think of somebody you share everything with (maybe your best friend). Assume you and this person are watching television together. You are enjoying the program a little. How would it make you feel to know that the other person is enjoying it greatly? My enjoyment would

□ decrease	□ decrease	□ remain the same	□ increase	□ increase
a lot	slightly		slightly	a lot

Q3. Think of somebody you do deals with (like, if you do the cooking, I will do the washing up). Assume you and this person are watching television together. You are enjoying the program a little. How would it make you feel to know that the other person is enjoying it greatly? My enjoyment would

□ decrease	□ decrease	□ remain the same	□ increase	□ increase
a lot	slightly		slightly	a lot

Q4. Think of somebody you are on equal footing with, you tend to get the same treatment (maybe a cousin or a class mate). Assume you and this person are watching television together. You are enjoying the program a little. How would it make you feel to know that the other person is enjoying it greatly? My enjoyment would

□ decrease	□ decrease	□ remain the same	□ increase	□ increase
a lot	slightly		slightly	a lot

Q5. Think again of somebody you respect highly (maybe your grandfather, your boss). Assume you and this person are watching television together. You are enjoying the program a little. How would it make you feel to know that the other person is really hating it? My enjoyment would

□ decrease	□ decrease	□ remain the same	□ increase	□ increase
a lot	slightly		slightly	a lot

Q6. Think again of somebody you share everything with (maybe your best friend). Assume you and this person are watching television together. You are enjoying the program a little. How would it make you feel to know that the other person is really hating it? My enjoyment would

□ decrease	□ decrease	□ remain the same	□ increase	□ increase
a lot	slightly		slightly	a lot

Q7. Think again of somebody you do deals with (like, if you do the cooking, I will do the washing up). Assume you and this person are watching television together. You are enjoying the program a little. How would it make you feel to know that the other person is really hating it? My enjoyment would

□ decrease	□ decrease	□ remain the same	□ increase	□ increase
a lot	slightly		slightly	a lot

Q8. Think again of somebody you are on equal footing with, you tend to get the same treatment (maybe a cousin or a class mate). Assume you and this person are watching television together. You are enjoying the program a little. How would it make you feel to know that the other person is really hating it? My enjoyment would

□ decrease	□ decrease	□ remain the same	□ increase	□ increase
a lot	slightly		slightly	a lot

Appendix 2: Questionnaire for Experiment 3

Assume you are going to listen to music together with a group of other people.
The DJ has asked all of you for your opinions on twenty songs. Each of you has rated each song from 1 (really hate) to 10 (really like). The DJ will use the individual people's ratings to calculate a rating of each song for the group. The DJ will then play the ten songs with the highest rating.

The DJ has calculated ratings for the group by **averaging** all individual ratings.

A song you **really like** has **NOT** been played. How sure are you that
somebody in the group hates it? *Not sure at all* 1 2 3 4 5 6 7 *Extremely sure*
most of the group hates it? *Not sure at all* 1 2 3 4 5 6 7 *Extremely sure*

A song you **really hate** has been played. How sure are you that
somebody in the group likes it? *Not sure at all* 1 2 3 4 5 6 7 *Extremely sure*
most of the group likes it? *Not sure at all* 1 2 3 4 5 6 7 *Extremely sure*

The DJ has calculated ratings for the group by taking the **minimum** of individual ratings.

A song you **really like** has **NOT** been played. How sure are you that
somebody in the group hates it? *Not sure at all* 1 2 3 4 5 6 7 *Extremely sure*
most of the group hates it? *Not sure at all* 1 2 3 4 5 6 7 *Extremely sure*

A song you **really hate** has been played. How sure are you that
somebody in the group likes it? *Not sure at all* 1 2 3 4 5 6 7 *Extremely sure*
most of the group likes it? *Not sure at all* 1 2 3 4 5 6 7 *Extremely sure*

The DJ has calculated ratings for the group by **multiplying** the individual ratings.

A song you **really like** has **NOT** been played. How sure are you that
somebody in the group hates it? *Not sure at all* 1 2 3 4 5 6 7 *Extremely sure*
most of the group hates it? *Not sure at all* 1 2 3 4 5 6 7 *Extremely sure*

A song you **really hate** has been played. How sure are you that
somebody in the group likes it? *Not sure at all* 1 2 3 4 5 6 7 *Extremely sure*
most of the group likes it? *Not sure at all* 1 2 3 4 5 6 7 *Extremely sure*

The DJ has calculated ratings for the group by **removing** all songs anybody really hated and **averaging** the individual ratings for the remaining songs.

A song you **really like** has **NOT** been played. How sure are you that
somebody in the group hates it? *Not sure at all* 1 2 3 4 5 6 7 *Extremely sure*
most of the group hates it? *Not sure at all* 1 2 3 4 5 6 7 *Extremely sure*

A song you **hate** has been played. How sure are you that
somebody in the group likes it? *Not sure at all* 1 2 3 4 5 6 7 *Extremely sure*
most of the group likes it? *Not sure at all* 1 2 3 4 5 6 7 *Extremely sure*

Authors' vitae

Dr. Judith Masthoff is a lecturer at the University of Aberdeen. She received her Ph.D. from Eindhoven University of Technology (the Netherlands) on the topic of an agent-based adaptive instruction system (awarded 1997 SNS bank prize for best applied thesis of the university in that year). Her interests lie in the areas of intelligent user interfaces, group recommender systems, intelligent tutoring systems, the evaluation of adaptive systems, personalized time-based media, and automated diagrammatic reasoning. Previous research has included work on adaptive multi-modal interfaces for the medical imaging domain for Philips Electronics' research laboratory.

Albert Gatt is completing his PhD in Natural Language Generation at the University of Aberdeen, where his research forms part of the EPSRC-funded TUNA Project. His thesis is on the Generation of Referring Expressions, focusing on plurals and sets. He is interested in the use of psycholinguistic and corpus-based methods for knowledge acquisition and implementation of GRE algorithms. He has a background in experimental psychology and linguistics.