Improving Evaluation Honesty and User Experience in E-learning by Increasing Evaluation Cost and Social Presence

Juha Leino, Tomi Heimonen

Improving evaluation honesty and user experience in e-learning by increasing evaluation cost and social presence

Juha Leino and Tomi Heimonen

University of Tampere, School of Information Sciences,
Kalevantie 4, 33014 University of Tampere, Finland

{juha.leino, tomi.heimonen}@uta.fi

Abstract. While various recommender approaches are increasingly considered in e-learning, lack of studies of actual use is hindering the development. For several years, we have used non-algorithmic recommender features on an undergraduate course website to help students find pertinent study materials. As students earn credit from adding and evaluating materials, some have chosen to evaluate materials dishonesty, i.e. without actually reading them. To improve honesty, in 2012 we coupled 5-star ratings with commenting (previously uncoupled) to increase the cost and complexity of evaluating and gave students individual presence with nicknames (previously anonymous) to increase social presence and enable reputation formation. Our results show that high enough cost of evaluating together with high enough social presence can lead to complete honesty in evaluations and enhance both user experience and student involvement. In effect, designing such e-learning systems includes not only designing the features but also their use, as the two are intertwined.
Keywords: e-learning, recommenders, ratings, social presence, honesty, design.

1 Introduction

Today, plenty of potential study materials can be found on the Internet, ranging from expert columns and scientific papers to video presentations. However, the material quality varies greatly. Also, even if a material is of high quality, it may be too advanced or too elementary for a student at a specific stage of his or her studies [1].

The authors teach an undergraduate-level course on user-centered design (UCD) at a local university. Given that a large majority of college/university students already use online resources to augment course materials and profess readiness to share them [2], the first author (course lecturer) programmed in 2007 the first version of LSRM (Lecture Slides and Reading Materials) for the course website (requires login) and has continued to develop it based on system use and student feedback.

LSRM harnesses the collective intelligence and efforts of the course student community by allowing them to add additional reading materials to complement the materials added by the instructors (the lecturer and the teaching assistant (TA), the second

author). In addition, over the years LSRM has provided students with various non-algorithmic recommending features for evaluating the materials as a community of peers to allow the community to guide its members to the most pertinent materials. In a sense, the system functions as a repository of high-quality material links that are annotated with recommending features (tags, ratings, and comments).

The purpose of LSRM is to encourage students to read more widely on UCD to augment learning and to develop a habit of following the field. UCD is a large topic, and it is practically impossible to cover all of its facets exhaustively on a course. LSRM gives students an opportunity to read more on the facets that they find interesting instead of forcing all to read the same materials. Reading these additional materials is voluntary in the sense that students are not examined on them but since 2009, online activity has affected the course grade. Since 2010, students have been required to add two materials and evaluate five to earn full credit for online activity.

A further goal is to help students develop information literacy, i.e. the skills to locate, select, evaluate, and use information from various sources [3]. Information literacy is considered a survival skill in today's information intensive working life and a foundation for life-long learning [3]. By selecting items from the Internet and then getting feedback on them from the evaluations by others, and by evaluating materials others add, students get to hone their information literacy skills.

The design of LSRM has from the start been driven by such contextual factors as short period of use (one semester) and low number of students (below 60 students) and items (below 150 materials). This has rendered such algorithmic approaches as collaborative filtering (CF) impractical [4]. Instead, LSRM has used recommender features that provide value to the community from the first contribution and that allow students to see their contribution to the community immediately. This is grounded on the notion that making visible to users the value of their contributions to the group has positive social outcomes [5]. In 2011, LSRM employed tagging, 5-star rating and commenting, all of which make the contribution immediately visible and useful/usable.

In the past, the system has been plagued by dishonest ratings, i.e. students evaluating materials without even viewing them. While not common enough to cripple the system, as discussed in [6], perceptions of other students not doing their work properly has harmed the user experience. In 2011, the previously used binary rating feature (*Yes* and *No* buttons to respond to the question *Did you find this material useful?*) was replaced with a 5-star rating feature to increase evaluation cost in order to reduce dishonesty. The change almost halved the percentage of dishonest ratings but at the cost of the overall number of ratings falling.

In 2012, evaluation cost was further increased by coupling rating and commenting into a unified evaluation (evaluation title, 5-star rating, and text justification for the rating). The goal was to further increase honesty and to improve the user experience by increasing trust on evaluations, in part by making the thinking behind the rating transparent. Moreover, while in 2011 the system use had been anonymous, in 2012 nicknames were employed to give students individual presence in the system in order to increase social presence—and thus social pressure to add pertinent materials and evaluations—and to enhance the user experience through sociality.

Based on log data, student questionnaire replies, and student interviews, both measures were successful. There were no dishonest evaluations in 2012 and the perceived social presence increased significantly, resulting in positive behavioral changes and enhanced user experience. Also, students reported having had learning and information literacy benefits from the system to a much larger extent than previously.

In addition to discussing how increasing social presence and cost of evaluations can be used to enhance user experience and significantly reduce dishonesty in e-learning, our study contributes to the field by providing a view of actual use by authentic users of a system employing recommender features. While various recommender approaches are today widely considered in e-learning, there is currently a dire need of actual experiences of such systems to guide the development efforts [7].

2 Related Work

Recommender systems (RS) help us deal with large numbers of items in two somewhat overlapping ways, by helping us find salient items (e.g. books we might be interested in) and by helping us make decisions (e.g. which book to buy) [4,8]. RS consist of one or more recommending features, varying from such non-algorithmic approaches as reviews and ratings to heavily algorithmic prediction-computing [4,6].

Recently, the potential of RS has attracted increasing interest in the e-learning research community (e.g. [2,8,9]). However, various domain-specific differences make transferring RS from one domain to another challenging [10], e.g. user interest is not the only determining factor in e-learning, as pedagogical aspects are also an important consideration [1,8]. Significantly, learners recognize this and are also ready to read uninteresting materials that are important for learning [1].

The ability of a recommender system to establish trust with its users is recognized as crucial to the system's success [11]. For this reason, RS have enjoyed more success in low-risk domains; users lack confidence to act on recommendations in high-risk domains [12]. Speculations and examples of dishonest users skewing recommendations have not helped in establish trust in RS [12,13]. In e-commerce the goal for dishonesty is typically to distort recommendations favorably or unfavorably for an item, often for financial profit, whereas in e-learning dishonesty has other motives, such as getting credit without earning it [6,14].

While having users explicitly rate items is a common approach to gathering user preferences, there are few guidelines to selecting rating scales, despite the fact that some scales tend to produce higher and some lower ratings on the same item [9,15]. In e-commerce, contributions appear to come mainly from highly opinionated users, resulting in an unrepresentative sample of user views [13]. Hu et al. [16] suggest that when benefits are not clear, user motivations for contributing can be explained with a *brag-and-moan* model. Similarly, it has been suggested that strongly negative or positive consumption experiences may lead to *expressing positive emotions* or *venting negative feelings* [17]. Consequently, while ratings and reviews have become an important information source in e-commerce, their ability to reflect actual item quality

has been questioned [13,16]. Still, altruism and concern for others also appear to be important motivations for contributing in e-commerce [17].

Social presence has been shown to have a positive impact on number of contributions and user behavior [18,19]. However, while RS can provide social texture that can lead to perceptions of others being present, different users require different interface cues to perceive a system as having social presence [18].

While accuracy metrics are important, there is a growing recognition that user experience is, in fact, more decisive [12,20]. RS need to be not only useful and accurate but also pleasurable to use [20]. Consequently, when evaluating recommenders, evaluating user experience is essential [12]. In e-learning, measuring satisfaction is important, as it is closely related to motivation [10]. The problem for researchers is that measuring user experience requires "field studies with long-term users of the system … measuring behavior in a natural context" [12]. Finding or building up and maintaining user communities for research is not easy while commercial systems tend to guard their trade secrets jealously [12]. This has resulted in a dire need for case studies of actual users in a real use context to guide employing RS in e-learning [7].

3 Study setting and data collecting

The 2012 UCD course was lectured in fall semester, and consisted of seven 2-hour lectures (Sept.) and fourteen 2-hour practice sessions (Sept.–early Dec.). The grade was based on design assignment (70%), ten smaller assignments (20%), and online work (full credit required adding two materials and evaluating five) (10%) plus extra 10% for high attendance. In 2012, the student community consisted of 36 students (18 female) while in 2011 there were 37 students (8 female).

In 2012, twenty students (56%), of whom 12 were females, filled out a questionnaire on LSRM. Movie tickets were raffled among the respondents. The questionnaire consisted of five sections: 1) Materials and adding them, 2) Evaluating materials, 3) Tagging materials, 4) Tools, and 5) Other aspects (e.g. social presence). Each section contained evaluative statements using a 7-point interval scale (1= strongly disagree; 7=strongly agree) and open-ended questions. Virtually all evaluative statements allowed commenting, which students frequently did. In 2011, 19 students (51%) filled out the questionnaire that consisted of six sections, as 5-star ratings and commenting had not been coupled and so the questions about them were in separate sections.

The 2012 students who filled out the questionnaire were on average more diligent and motivated than their peers (Table 1). However, the two groups were not that different when it came to viewing materials. Importantly, one student who added no materials or evaluations filled out the questionnaire, thus giving us a view into the thinking of these students, too. Students who filled out the questionnaire represent well the majority of students (81%) who made at least the required number of contributions. In fact, removing respondents who made no contributions (7) eliminates the differences between the groups. The trends for 2011 are similar.

Table 1. Students who filled out the questionnaire vs. students who did not (2012)

	Grade	Motivation	Added materials	Added evaluations	Materials viewed
Students who filled in the survey (avg.)	4.3	4.55	2.6	4.95	22.7
Students who did not (avg.)	3	3.81	1.25	3.13	12.44
Statistical significance of difference	Yes	Yes	Yes	Yes	No
p-value (unpaired t-test)	.0001	.0041	.0012	.0071	.0756

In addition, we interviewed three students (8%) in 2012 to gain deeper understanding of student motivations and views concerning the system and its use. The second author conducted the semi-structured interviews that lasted 30–45 minutes per student, as the first author had designed and built LSRM, something that the students were aware of. When a quote is from an interview, we mention it; otherwise, quotes are from questionnaire replies. In 2011, no interviews were conducted.

As the system collected virtually click-by-clack data of individualized student activity on the LSRM page, we are able to contrast *saying* (questionnaire and interview data) with *doing* (actual activity data), thus reducing the potential of say-do problems.

4 LSRM system description

The LSRM system was implemented with PHP, JavaScript, and HTML. As most interactive parts were implemented with AJAX (Asynchronous JavaScript and XML), most interactions took place within the current use context. Changing views and navigating by tags, however, reloaded the page. Clicking a material link opened the link in a new browser window.

The LSRM interface in 2012 consisted of a web page that gave different views to the material available (Figure 1). The interface was originally in Finnish, and has been translated into English for relevant parts for this paper.

Because we often compare the student behavior and perceptions in 2012 to those in 2011 to see how the changes in the system affected them, we also discuss to some extent the 2011 design. Moreover, the 2011 design and experiences are relevant because the changes in the system are largely based on student feedback on it.

In 2011, there were no separate views except for navigating with tags. All materials were added to and listed under lectures (most recent lecture on top). Also, while in 2011 all actions had been anonymous, in 2012 all material additions, evaluations, and comments were identified by student (or instructor) nickname.

In 2011, the system allowed students to add materials, rate (5-star scale) materials others had added, comment (*Title* and *Text* fields) all materials (also their own to enable discussing), and tag all items. Rating and commenting were decoupled. In 2012, rating and commenting were coupled into an evaluation (evaluation title, star rating on 5-star scale, and text explanation for the rating). Stars were given at the precision of full stars while star averages were displayed at the precision of half-stars. Evaluat-

ing one's own material addition was impossible. To enable discussing, evaluations could be commented (voluntary in the sense that commenting did not affect grade).

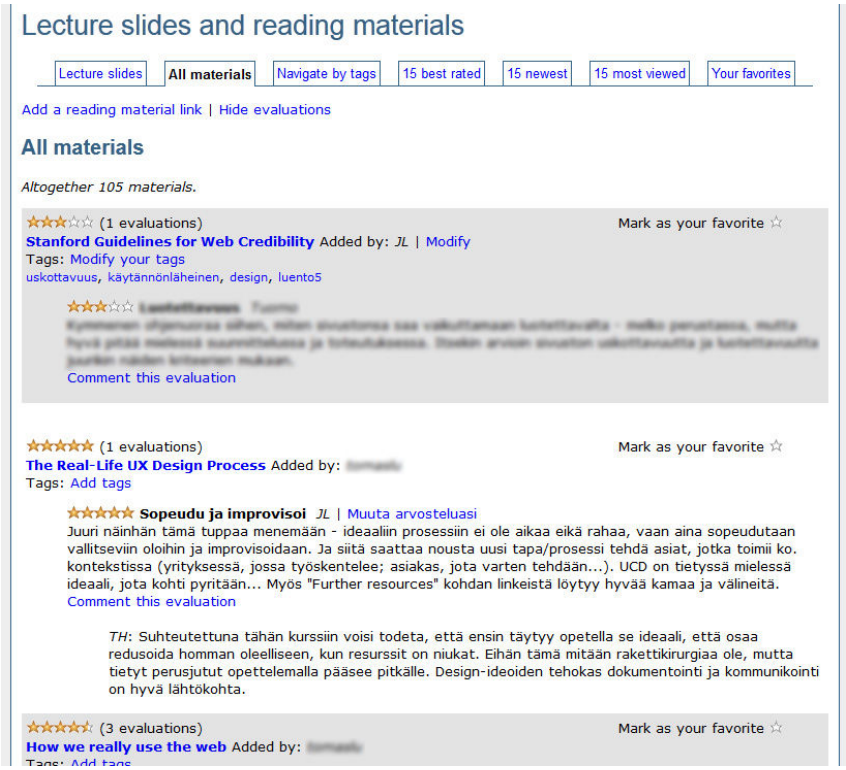

Fig. 1. LSRM interface in 2012—the content is authentic (student nicknames and texts blurred) but the order of the materials has been adjusted for illustrative purposes

While in 2011 comments were hidden by default and had to be opened for viewing for each material to keep the page length at bay, in 2012 evaluations and comments were displayed by default as a result of student feedback. Tagging materials was also possible, and unlike in 2011, students were asked to tag the materials they added.

4.1 Working hypothesis behind the 2012 LSRM design

We based the current design of LSRM on working hypothesis grounded on our experiences and student feedback from the earlier versions of the system. Accordingly, we decided to 1) use nicknames instead of anonymity to increase sociality and trust on both evaluations and materials, and 2) increase the cost and complexity of evaluating materials by coupling five-star rating with a title and comment to increase trust on evaluations and perceived quality of evaluations. Furthermore, with materials getting evaluated more comprehensively, we expected the pressure on students to add good materials to increase and result, in turn, in increased perceived quality of materials.

4.2 LSRM tool usability

According to questionnaire responses, the tools LSRM offered were considered easy to use (M=5.60; SD=.82). In 2011, the ease-of-use was evaluated slightly lower (M=5.06; SD=1.16) but the difference is not statistically significant. Consequently, we conclude that usability problems were few and did not significantly color the student perceptions. Also, there were no year-on-year differences in tool usability.

5 Results

The number of material additions and evaluations followed closely what was required for the full credit. Seven students evaluated no materials, 25 students evaluated the required five materials, and four evaluated six (M=4.14; SD=2.09). The lecturer evaluated three. Likewise, eight students added no materials, 17 added the required two, seven added three, three added four, and one added five (M=2.00; SD=1.31). The instructors added 36 materials. The number of materials added and evaluations made correlated strongly, $r(103)=.71$, $p=.01$, meaning that students who contributed in one way also contributed in the other. Only two students commented evaluations made by others (each once) while the instructors commented five.

The average number of honest evaluations per student almost doubled from 2.09 (SD=2.65) in 2011 to 4.14 (SD=2.86) in 2012. The difference is statistically significant, $t(71)=2.28$, $p=.026$. However, this did not come at the cost of activity; the average numbers of evaluations were not statistically different from 2011. Thus, increasing evaluating complexity did not reduce the number of evaluations in 2012.

In 2012, students viewed on average 18.1 materials (range=0–70, SD=17.25). This represents a clear and statistically significant increase from 2011 at $t(71)=3.61$, $p=.001$ when students on average viewed only 7.1 materials (range=0–29; SD=7.06).

5.1 Additional reading materials

Students perceived being able to add additional reading materials positively both in 2012 (M=5.70; SD=1.17) and 2011 (M=5.11, SD=.99). Student comments also indicate that the majority of students saw the feature very positively. They especially appreciated how the materials complemented and added to the lectures, giving more information on the topics that interested them and covering a wider spectrum of topics than possible in the lectures: "*You found materials that had not been discussed in the lectures. ...you had an opportunity to read on topics you were interested. Also, you could share with others interesting and useful materials that had inspired you.*" Students also emphasized that nobody can find all the good materials alone. Several students also mentioned that having to add materials instilled a good habit of following the field.

When viewed numerically, students perceived the added materials in 2012 much as in 2011. In 2012, students rated the quality of the materials at 5.40 (SD=.60) and in 2011, at 5.21 (SD=.79). The difference is not statistically significant. However, other differences support the idea that students viewed the materials more positively in

2012. While in 2011, some students suspected that others had added materials without selecting them carefully, in 2012 only one student questioned the quality of materials. Most comments saw material quality in positive terms: "*The materials covered a very wide spectrum … It felt that the people who added materials had really wanted to add the articles and not just find quickly 2 articles….*"

In fact, many students reported having read numerous articles to find links that they felt were worth adding in the sense of being "*genuinely interesting and useful*" to other students: "*I wanted to find as high-quality materials as possible, materials that I felt had taught me something so that others also could learn from them.*" Many students professed altruistic motives and wanted others to benefit from their work: "*Even though, as far as I know, the evaluations that the added links got didn't affect the grade, I didn't feel it proper to add a material that I would not have reviewed positively myself.*" At the same time, at least some students were acutely aware that others would evaluate the links they add: "*Of course, there also was something of a 'social pressure' since I knew that others would later evaluate my materials….*"

When selecting links to add, students considered what was useful and relevant for oneself—"*But the good thing was that I had to learn to find articles that are pertinent for me*"—and others: "*It taught me to search for information on a topic and be critical towards it.*" Students felt that selecting items "*forced me to read the materials with care and at the same time think about their good and bad aspects. …it helped learning.*" As one student put it nicely, "*…adding materials in and of itself was beneficial because looking for a suitable article made you reflect on the concepts taught on the course and appraise the relevance of the articles according to them.*" In effect, many student comments show that they had to work on their information literacy skills, learning to locate, select and evaluate materials [3], and that they were aware of learning these skills. These aspects were much less discussed in 2011 comments.

However, not all students went through the trouble of finding pertinent links to add. Two students admitted having taken the easy way out: "*…adding materials was 'compulsory chore'…. I thought that all I've got to do is find some relatively sensible piece of text that has something to do with the course….*" While getting all the students to work hard on finding good materials is probably impossible, using social presence and high-cost evaluations appears to have resulted in most students taking adding materials seriously and thus reaping benefits from it.

5.2 Selecting materials to view in LSRM

Students viewed on average 18.1 materials in 2012, slightly more than 2.5 times the average in 2011 (M=7.1). Instructor-added materials were on average viewed 5.56 times (SD=2.63) and student-added ones 6.67 times (SD=3.57). The difference is not statistically significant. The most important selecting criterion was the content of the link, the "*interestingness*" of the article. There were two sources for judging this, 1) the title of the link—"*Interesting topics, interesting title in particular. The title should tell what it's all about.*"—and 2) the evaluations: "*I browsed materials and their evaluations to form an impression … and decided based on the impressions of the interestingness of the topic and the quality with which the topic was covered.*"

Also, both the existence and number of evaluations functioned as important heuristics "*A material with evaluations stood out from the mass and I was more likely to check them out.*" The number of evaluations was a heuristic for interestingness of the link for many students quite independent of their valence. Being part of the communal activity appears to be part of this: "*If the evaluations by others repeatedly referred to some aspect [in the article] or brought up an interesting thing, you also wanted to read it. It wasn't just the good evaluation that affected; I also read 'worse' ones.*" In fact, selecting already evaluated materials had social aspects: "*Evaluated materials appeared more attractive…. The good thing about them was that you get to compare your viewpoint to that of the evaluator.*" This is something that making commenting a compulsory part of evaluations enabled. Interestingly, some students also mentioned having viewed unevaluated materials on purpose so that their evaluations would be useful to others, bringing an altruistic aspect to sociality.

In 2012, students judged the impact of evaluations on which materials they read to be on average 5.05 (SD=1.73). While the 2011 average was lower (M=4.00; SD=1.97), the difference is not statistically significant. Still, in 2012, students tended more towards feeling that evaluations had an impact, as 75% rated the effect at 5-7 while in 2011 only 53% did so. In 2012, the question concerned evaluations while in 2011 only ratings. Since there was no statistical difference between the averages of star ratings given in 2012 and 2011 at $t(294)=1.17$, $p=.24$, we conjecture that coupling the star-rating with a comment is the main reason for proportionally higher ratings of effect.

We identified three facets of evaluations that had an appreciable impact on the process of selecting a material for further investigation: The star rating averages, evaluation valence, and the number of evaluations. While star rating averages clearly affected the equation—rating averages correlated positively with the number of viewers, $r(103)=.52$, $p=.01$—some students felt that star ratings did not necessarily tell much while the textual reviews did: *"I don't trust the star ratings, as everybody has their own rating scales in their heads. Comments and reasonings, on the other hand, give you hints about why the article might be worth reading.*" Perhaps consequently, star averages became more important as a selection heuristic towards the end of the course: "*At the beginning there weren't so many materials, so you could read them all or at least glance at them. … Towards the end I browsed the ones with most stars and opened the link if the evaluations aroused my interest sufficiently.*"

The valence of the evaluation also affected the equation for many: "*I selected for reading materials when the title aroused my interest. The ratings and reviews affected if I opened an article or not. If somebody had commented that rather superficial and circumspect, then no need to think twice if I opened it or not :)*" In turn, praising evaluations attracted viewings: "*I selected materials based on topic… The evaluations also influenced what I selected; if the article was highly praised, I took a look.*"

However, according to student comments, the existence and the number of evaluations was at least as significant a selection criterion as valence. In fact, the correlation between the number of ratings and the number of students viewing the item is higher at $r(103)=.64$, $p=.01$ than between the average of evaluations for the materials and the number of students viewing it.

That a material ended up being opened did not, naturally enough, mean that it was carefully read. Opening the material was only a step in the process. The next step was to glance at the material, to decide if it really was worth reading. If the article passed this impression-forming glance and the students started to read it, the judging process continued: "*I only bothered to read the article to the end if it gave me some new information or a new viewpoint, i.e. I felt it to be useful.*"

5.3 Evaluating materials

In 2012, evaluating a material involved an in-depth evaluation (title, star-rating, and text justification) while in 2011, compulsory evaluating consisted of simply giving a star rating while commenting was voluntary. Students viewed the possibility to evaluate materials positively, on average at 5.55 (SD=1.15) in 2012 and at 5.11 (SD=1.59) in 2011. The difference is not statistically significant; increasing the cost of evaluating did not reduce the positivity of student views. As reasons for liking evaluations, students mentioned social aspects and possibility for expressing opinions in addition to evaluations helping in selecting materials for reading.

Many students especially liked the text part of the evaluation: "*...the possibility of textual evaluation was very good, and it was also good that you could comment evaluations! It even led to some discussing.*" However, one student mentioned a negative aspect concerning the text comment: "*You saw what others had praised so you read it with interest, too, but I myself didn't like to evaluate. Especially since I'm not a professional or somebody who knows a lot, so my comments may have appeared pretty bad for somebody who knew more.*" Two interviewed students touched the same theme, noting that it was easier to comment in LSRM than on the Internet because there are so much more knowledgeable people on the Internet. For them, a smaller community with the members more or less at the same level of knowledge made it easier to comment. Consequently, a small, closed community of peers can create a safer environment to encourage participation. This finding is in line with the idea that the sense of community is connected to a feeling of *membership* that includes boundaries that provide members with emotional safety [21].

Star-ratings were seen in a more problematic light: "*Comments enrich and give new viewpoints, stimulate discussion. Star-ratings I found somewhat unnecessary.*" One reason for disliking star-ratings was the difficulty in deciding the appropriate ratings: "*Occasionally it was hard to give stars because even if the material was really useful for me, it's not necessarily that for everybody so you don't want to rate it too highly, either.*" The problem was exacerbated by the fact that students were not told which specific aspect to rate. This was a clear design mistake; the interface should have made the rated aspect clear.

Although students in 2012 viewed on average slightly over 2.5 times more materials than in 2011, the number of evaluations on average was statistically the same in 2012 (M=4.14, SD=2.09) as in 2011 (M=3.89, SD=2.75). Also, on average materials had statistically as many ratings/evaluations in 2012 (M=1.45, SD=1.54) as in 2011 (M=1.8, SD=1.96). In effect, in 2012, students therefore read more materials but evaluated fewer in relation to the number read than in 2011. However, significantly,

on average students made more *honest* evaluations, i.e. viewed the material before evaluating it, in 2012 (M=4.14, SD=2.09) than in 2011 (2.86, SD=2.65). The difference is statistically significant, $t(71)=2.28$, $p=.026$.

In fact, in 2012, not a single dishonest evaluation was made: The use log data shows unequivocally that on each occasion, the student had opened the link before evaluating it. Increasing social presence and the cost and complexity of evaluations removed dishonest attempts to get points without earning them entirely. Also, since the number of evaluations per student did not decrease, increasing the cost of evaluating did not reduce the number of evaluations. In effect, more students were motivated to do the required work when the 2012 design was used.

Not only did increasing cost and complexity of evaluations in comparison to ratings result in complete honesty but it also resulted in perceptions of honesty. Not a single student mentioned suspecting dishonesty in the 2012 feedback while such suspicion was entertained in the 2011 feedback (when dishonest rating in fact took place). While social presence likely played a significant role in this, too, given that changing the rating scale from binary (2010) to 5-star (2011) reduced dishonest rating almost by half (from 43% in 2010 to 26% in 2011), we conjecture that needing to write a textual justification for the rating made cheating simply too difficult: "*You had to read the materials to be able to evaluate it.*" Since student perceived others as doing the required work, they also ended up reciprocating by doing their own share.

The care with which materials were read before they were evaluated increased clearly. When we look at the time periods that passed between opening the link and adding the evaluation (when evaluating took place within the same session, as it typically did) for honest ratings/evaluations, we notice that the reading times almost doubled in 2012. If we examine the reading times that were shorter than 15 minutes (to filter out sessions that may have included other activities), there is still a clear statistical difference between 2012 (M=409 seconds, SD=242.33) and 2011 (M=231 seconds, SD=203.94), $t(164)=5.133$, $p < .001$. The materials were clearly read longer in 2012 before they were evaluated. In fact, in 2012, only in two cases (1.8%) did a student evaluate a material after reading it for less than a full minute while in 2011, there were 21 (23.3%) such cases. Still, having to write a textual evaluation in addition to clicking a rating must also have been a partial reason for the increased time between open a link and evaluating it.

In 2012, only four students (21%) said that they had rated all the materials they had read while in 2011, eight students (42%) said the same. In both years, the main reason for not rating a viewed material was the same; students felt that they had not read the material carefully enough to rate it.

While in 2011 some students did refer to social factors as a rationale for not evaluating, it was in 2012 that social aspects were mentioned repeatedly in this context. Sociality inherent to the system clearly affected evaluating behavior: "*There were social aspects to evaluating, so I did not want to write an evaluation that just said 'nice one' or 'interesting article' but something more. For this reason I wanted to evaluate only articles on which I had a clear opinion and something a bit deeper to say—something that might inspire others to comments and something that others could comment.*" Besides a certain social pressure, there also was a sense of moral

duty towards others: *"...I would have felt wrong about evaluating a material that I had not read entirely.*" In effect, as with material additions, certain altruistic motivations were evident in many student comments concerning evaluations.

Also, a few students mentioned that they did not evaluate some materials they had read "*because I had nothing new/significant to add to the comments by others.*" While no student in 2011 mentioned thinking twice about rating a rated item, coupling ratings and comments made students feel that they had to have something significant to say, something that had not already been said: "*Somebody else might have already said what was essential in his or her comment.*"

Another reason for students not to evaluate a material they had read in 2012 was that "*the materials had not aroused any big emotion*." Mediocre, bland articles simply did not garner evaluations: "*I evaluated materials based on whether they stirred up thoughts or not. I selected for evaluating only materials on which I had some kind of an opinion. In the evaluating phase I simply skipped lackluster articles altogether.*" If students did not have something to say about the material, they did not evaluate it.

Consequently, students read articles based on personal interests and need. If they read the whole article and felt that it was "*useful*" and "*interesting*" and they had something to say about it (that somebody else had not already said), they probably evaluated the material. As a result, students largely ended up evaluating good articles: "*I didn't really bother to read materials that I found worthless with the first glance, so I ended up choosing for evaluation only good materials.*"

Usefulness was the most important criterion for students when they evaluated materials. However, how clearly written and presented and how illustrative the article was also affected the evaluation. Moreover, students appreciated learning something new from the material. "*Usefulness and practicality, can I use the material in future in studies and at work. Also if I learned something new and if the materials was relevant to the course and its content.*"

The above factors largely explain why over 50% of the star-ratings were four stars and 70% 4–5 stars in 2012. There is no statistical difference between the average star-ratings in 2012 (M=3.82, SD=.83) and 2011 (M=3.70, SD=.86), $t(294)=1.17$, $p=.244$ (Table 2).

Table 2. Distributions of star ratings in 2012 and 2011

	1 star	**2 stars**	**3 stars**	**4 stars**	**5 stars**
2011	0 (0%)	13 (9%)	42 (29%)	64 (44%)	25 (17%)
2012	4 (3%)	1 (1%)	41 (27%)	79 (52%)	27 (18%)

In 2012, three students gave 1-star evaluations (two once and one twice). Two materials ended up having one star as the average of its ratings (both had two 1-star evaluations). Based on the evaluation texts, it appears that the students giving the 1-star evaluations felt that the materials should not have been added to the system. Interestingly, it was only when nicknames were used that 1-star ratings were made; in 2011, no 1-star ratings were made. We conjecture that a heightened sense of social presence/sociality resulted in people showing disapproval for substandard materials.

Most appeared to have communal, even altruistic motivations, and they probably expected the same in return from the other members of the community. Thus, while students mostly ended up selecting good materials to evaluate, when the experience was strongly negative, they were ready to *vent negative feeling* [17], even if the comments connected to 1-star evaluations were still quite polite and matter-of-fact in tone.

Several student comments underline that students were aware of the benefit they got from reading materials and that they understood that evaluating materials did make them to read them more carefully: "*evaluating materials involved reading a lot of materials when 5 evaluations were required. 5 felt a lot but afterward I felt it was useful that it made me read so many articles.*" Students were also clearly aware of the information literacy benefits that evaluating materials and reading evaluations by others brought: *"Evaluating materials increased the teaching value of the articles. By reading the evaluations by others we got good feedback on how to apply scientific texts in studying. Also, finding and reading articles in our field is very important, especially for working life.*" This contrast with the 2011 comments where students focused on the ability of rating to guide them to better materials and to warn them against bad ones but did not discuss much the benefits of reading and evaluating.

Another social aspect of evaluations was curiosity of how others evaluated the materials one had added: "*Of course I checked out what kinds of evaluations the materials I had added had gotten.*" One interviewed student in fact mentioned she had just before coming to the interview (interviews took place after the course) checked if there had been any new evaluations on her materials. Another interviewed student said that the system gave him a feeling that the materials he had added were useful to others and that one of them had had "*a fair amount*" of evaluations. What others said clearly mattered to and interested students.

5.4 Perceived social presence and its impact

While contributions had previously been anonymous, in 2012 students were asked to choose a nickname for LSRM when registering to the course. The purpose was to increase perceived social presence and to allow reputation formation by giving students individual presence in the system. In effect, there was a significant effect for social presence, $t(36)=2.06$, $p=.047$, with students reporting on average higher social presence in 2012 (M=3.53, SD=1.22) than in 2011 (M=2.75, SD=1.15).

Given that we were using a 7-point scale, average perceived social presence of 3.53 might not seem very high. However, LSRM is in a sense competing against such popular social systems as Facebook and instant messaging when it comes to how social students perceive it. In that sense, the seemingly low average of 3.53 may in fact actually indicate a relatively high perceived social presence for a system that does not support real-time presence. In fact, some comments from students who professed not to have felt others as present emphasized the slow rhythm of interaction: "*I didn't find much [social presence] because there wasn't that much activity and I didn't get announcements of e.g. that my materials had been evaluated etc. The thought of there being other students affected so that I wondered what others thought about my evaluations….*" One interviewed student encapsulated the general perception by saying that

the feeling of sociality was clear but not very strong and that the level was appropriate to the system, as it was about as much as can be archived "*unnaturally*."

Student comments indicate that the perceived social presence had a clear and positive impact on behavior: "*Sociality in the service affected my actions significantly: It affected so that I wanted to select as suitable articles as possible to add to the page and that I wanted to say something more deep than just 'quite nice' in the evaluation.*" Repeatedly, students mention having tried to find materials that would be useful to others and making evaluations that would help others. While there were similar trends also in 2011, altruistic aspects are more emphasized in the 2012 comments.

Some comments clearly connected nicknames to reputation: "*When adding materials I thought that I can't add just any odd stuff … because all the other students see them. The fact that my name was connected to the materials and evaluations I added also made me think twice what to say.*" Students appear to have felt that through the nickname they had an individual presence and reputation in the system, and that affected their behavior positively. Having an individual presence in the systems also made students consider their self-image in relation to the community, as one interviewed students explained: "*Also, building my self-image influenced it; I didn't feel it satisfactory for myself to put there something that I wouldn't want to read myself.*" The student comments indicate that most wanted to be responsible members of the community, and in this sense, the achieved social presence was high enough.

While some students professed not having felt the presence of others, their comments show that their actions were nevertheless influenced by awareness of others, e.g. "*I didn't really feel others to be present that much. Still, the thought that others see what I add there affected what materials I added and what kind of evaluations I made. I.e., I did my job with care.*" The impact of this awareness on student behavior appears to have been larger than the numeric evaluation of social presence indicates.

Social presence had both activating and experience-enhancing impact: "*…the presence of others there activated me, too, and it was great to see that others actually read and evaluated materials.*" In particular, student comments underline that perceiving others as present improved the user experience, e.g. "*The presence of other students affected positively because [that way] you knew that somebody else is also reading these comments and not just the teacher alone.*"

The positive effects of increased social presence appear at least partially attributable to increased social pressure that drove students to do more than just bare minimum for earning points: "*[I read the materials I added] very carefully indeed exactly because of the sociality connected to evaluating materials. There was some 'social pressure' involved in evaluating articles because other students could read your evaluations, respond to them, disagree and comment, respond to the evaluation…*"

The effects of social presence are likely intertwined with the effects that the emerging sense of community had. Besides membership, using LSRM also had many other elements that McMillan and Chives [21] suggest as contributing to the sense of community, including *personal investment* (added evaluations and materials), bidirectional *influence* (students affecting the community and vice versa) and *integration and fulfillment of needs*.

5.5 Social presence of instructors and its impact

Students felt that the lecturer's presence had a positive impact in two ways. First, he maintained a feeling of activity in the system, e.g. "*A lot of added materials and evaluations on materials [by the lecturer]. I feel that it was good that the lecturer kept the page active when it occasionally got silent.*" Second, the lecturer's presence brought positive social pressure: "*The presence of the lecturer did encourage investing in the materials. I didn't have the nerve to add just any old dude's blog there and added instead content by recognized sites or known experts.*"

Some students mentioned having been nervous about evaluating materials added by the lecturer. In fact, student-added materials (M=1.88, SD=1.54) did get on average more evaluations that instructor-added materials (M=.61, SD=1.18), $t(103)=4.34$, $p < .001$. Interestingly, there was no statistically significant difference between the average star ratings given to instructor-added materials (M=3.95, SD=.72) and student-added materials (M=3.79, SD=.84). The scale, in a sense, was the same, meaning that the 16 students who evaluated instructor-added materials (in contrast, 30 students evaluated student-added materials) did not give the materials special treatment.

Comments and evaluations from the lecturer were warmly welcomed: "*I especially liked how the lecturer commented on some evaluations and gave his own examples (e.g. on a material that I had added).*" The evaluations and comments that the instructors added were mainly positive or, in one case, only mildly challenging.

6 Discussion

Using nicknames increases sociality and trust on materials and evaluations. Using nicknames clearly increased social presence, as evidenced by the statistically significant increase in student evaluations and student questionnaire replies. Even the students who did not feel others as present described how the idea of other users affected their behavior positively. Many students reported altruistic motivations, and their comments show that they felt a certain sense of duty towards others. Also, students had a sense of individual presence in the system, which created social pressure that also affected their actions positively. Overall, students perceived others as taking online activity seriously, and this motivated them to approach it with due diligence.

Increasing evaluation cost engenders trust on evaluations. Students clearly trusted better that student evaluations were properly made than in 2011. How much of this is attributable to nicknames and how much to increased evaluating cost is an open question, but considering that in 2011, increasing the rating cost resulted in a significant improvement in honesty, we conjecture that coupling commenting and rating affected it significantly and also encouraged reading materials more carefully.

Requiring more thorough evaluations increases pressure to add good materials. There was less questioning of the motives of the students adding materials in 2012, indicating that at least some aspects of the perceived quality had improved. Also, many student comments show that students did approach finding materials to add very seriously, in part because they knew that they would be evaluated. We conclude that the more complex evaluations made students more careful about the links

they added but that the effect is again intertwined with the effects of the nickname use.

Overall, our working hypothesis concerning nicknames and evaluation cost worked out well. Using nicknames and a more complex evaluating approach removed dishonesty entirely from evaluations. Importantly, this was accomplished without the number of contributions falling; in fact, the number of honest contributions increased. This is a significant improvement to the system and gives other e-learning practitioners practical tools and approaches to root out dishonesty

6.1 A word of caution: Use and interface design are intertwined

When applying our results in e-learning, and particularly in other contexts, it should be noted that our results are subject to specific contextual factors. Our system is designed for formal e-learning where compulsoriness can be used to encourage contributions, students form a small, closed community of peers, and the use period is short. In informal e-learning, for instance, using compulsoriness may not be possible or even advisable. Also, designing an e-learning space must go hand in hand with designing its use (compulsoriness, regulations etc.), as the two are intertwined. For example, our system used compulsoriness and high evaluation cost to improve evaluation/rating honesty. However, if there had been no compulsoriness, this design would likely have failed. With voluntary evaluating, it would have been advisable to lower the evaluation cost to encourage contributing; after all, there would have been little motivation for dishonesty. In 2011, while there were dishonest ratings (compulsory), there was no dishonest commenting (voluntary).

6.2 Enhanced user experience

The 2012 use and interface designs led to students approaching their work more honestly and with more gusto. The space was significantly more social due to the tools bringing sociality and providing social texture and the nicknames providing sociality through individual presence. Student trust on materials and especially evaluations increased; students saw others as doing their work properly, which led to altruistic motivations and a sense of duty towards others, leading to deeper involvement. In effect, this perception was justified by actual changes in honesty and due diligence, as most students did work hard to add links that were meaningful and tried to make evaluations that would be useful. Also, in 2012, many students were more aware of the benefits—in particular, improved learning and information literacy skills—they accrued from using the system with due diligence. Seeing benefits in turn encouraged using the system, leading to a virtuous circle: Positive behavior led to positive experience and perception that in turn encouraged positive behavior.

6.3 How to develop LSRM further

Further increases in evaluation cost do not appear necessary, as 100% honesty was already reached with the current approach. Nevertheless, the aspect to be rated needs

to be made clear in the interface so that all students are rating the same aspect. The most important criterion for students, *usefulness* of the materials, is the obvious candidate. While increasing the cost of adding materials by requiring a description might seem a logical step to induce further trust on materials, it may prove problematic; in an interview, one student said that this would be a "*miserable feature*" that would only lead to marketing one's materials instead of objectively describing them.

In fact, the most promising approach to encourage positive behavior and improving the user experience further appears to be enhancing sociality and individual presence in the system. First, the system needs to be more connected to students' everyday lives. LSRM should make it possible for students to subscribe to email notices so that they can maintain awareness of any development in the system that concerns them. In addition, the system should incorporate a private group in Twitter and Facebook, to mention two obvious candidates. This way, the information about new materials and evaluations would reach students without them having to log in the system. Making groups private is important, as the community being small, closed, and consisting of peers (same level of knowledge) were important factors for students.

The second approach to increasing sociality is to enhance the sense of individual presence by allowing viewing material additions, evaluations and comments by individual students. This would also increase social pressure, as one could not hope to be hidden in the mass of materials. Interviews gave indications that students would not find this intrusive. In all likelihood, this would further increase the care with which students add and evaluate materials.

7 Conclusion

We managed to enhance students' user experience and remove dishonesty from additional reading material evaluations by increasing the evaluation cost and by replacing anonymity with nicknames, thus giving students an individual presence. This study also contributes to the field by providing much-needed experiences of using recommender features in e-learning in a genuine use context.

However, when applying our results in other systems and contexts, one should bear in mind that the results were obtained within formal e-learning context and that, consequently, contextual factors may limit their applicability elsewhere.

While we hope to develop LSRM further to the directions outlined here, we also encourage other practitioners to report their experiences of RS in various educational contexts. It is important that practice and theory go hand in hand in employing RS in e-learning instead of theories being developed independent of the ground realities.

References

1. Tang, T., McCalla, G.: Beyond Learners' Interest: Personalized Paper Recommendation Based on Their Pedagogical Features for an E-learning System. In: LNCS. Vol. 3157, pp. 301--310. Springer, Heidelberg (2004).

2. Hage, H., Aïmeur, E.: Harnessing Learner's Collective Intelligence: A Web 2.0 Approach to E-learning. In: B.P. Woolf, E. Aïmeur, R. Nkambou, Lajoie (Eds.). LNCS, vol. 5091, pp. 438--447. Springer, Heidelberg (2008)
3. Kiliç-Çakmak, E.: Learning strategies and motivational factors predicting information literacy self-efficacy of e-learners. AJET 26(2), 192--208 (2010)
4. Schafer, J. B., Konstan, J., Riedl, J.: Recommender Systems in E-commerce. In: Proc. the 1st ACM Conf. on Electronic commerce, pp. 158--166. ACM, New York (1999)
5. Rashid, A.M., Ling, K., Tassone, R., Resnick, P., Kraut, R., Riedl, J.: Motivating Participation by Displaying the Value of Contribution. In: Proc. the SIGCHI conf. on Human Factors in computing systems, pp. 955--958. ACM, New York (2006)
6. Leino, J.: Case Study: Material Additions, Ratings, and Comments in a Course Setting. In: O.C. Santos, J.G. Boticario (Eds.), Educational Recommender Systems and Technologies: Practices and Challenges, IGI Global, pp. 258--280 (2011)
7. Manouselis, N., Drachsler, H., Vuorikari, R., Hummel, H. Koper, R.; Recommender Systems in Technology Enhanced Learning. In: P. Kantor, F. Ricci, L. Rokach, B. Shapira (Eds.), Recommender Systems Handbook, pp. 387–415. Springer, Heidelberg (2011)
8. Santos, O.C., Boticario, J.G.: Modeling Recommendations for the Educational Domain. Procedia Computer Science 1(2), pp. 2793--2800 (2010)
9. Ghauth, K.I., Abdullah, N.A.: Learning Materials Recommendation Using Good Learners Ratings and Content-based Filtering. Journal of Educational Technology Research and Development 58(6), 711--727 (2010)
10. Drachsler, H., Hummel, H.G.K., Koper, R.: Identifying the Goal, User model and Conditions of Recommender Systems for Formal and Informal Learning. Journal of Digital Information 10(2), 4--24 (2009)
11. Pu, P., Chen, L.: Trust-inspiring Explanation Interfaces for Recommender Systems. Knowledge-Based Systems 20, 542--556 (2007)
12. Konstan, J., Riedl, J.: Recommender Systems: From Algorithms to User Experience. User Model User-Adap Inter 22, 101--123 (2012)
13. Talwar, A., Jurca, R., Faltings, B.: Understanding User Behavior in Online Feedback Reporting. In: Proc. the 8th ACM conf. on Electronic commerce, pp. 134--142. ACM, New York (2007)
14. Lam, S.K., Riedl, J.: Shilling Recommender Systems for Fun and Profit. In: Proc. the 13th int. Conf. on World Wide Web, pp. 393--402. ACM, New York, (2004)
15. Sparling, E.I., Sen, S.: Rating: How Difficult is It? In: Proc. the 5th ACM conf. on Recommender systems, pp. 149--156. ACM, New York (2011)
16. Hu, N., Pavlou, P., Zhang, J.: Can Online Reviews Reveal a Product's True Quality? In: Proc. the 7th ACM conf. on Electronic Commerce, pp. 324--330 ACM, New York (2006)
17. Hennig-Thurau, T., Gwinner, K.P., Walsh, G., Gremler, D.D.: Electronic Word-of-Mouth via Consumer-opinion Platforms: What Motivates Consumers to Articulate Themselves on the Internet? Journal of Interactive Marketing 18(1), 38--52 (2004)
18. Leino, J., Räihä, K.-J.: User Experiences and Impressions of Recommenders in Complex Information Environments. IEEE Data Engineering Bulletin 31(2), 32--39 (2008)
19. Nov, O., Ye, C.: Why Do People Tag? Communications of ACM 53(7), 128--131 (2010)
20. McNee, S.M., Riedl, J., Konstan, J.: Being Accurate Is Not Enough: How Accuracy Metrics Have Hurt Recommender Systems. In: CHI '06 Extended Abstracts on Human Factors in Computing Systems, pp. 1097--1101. ACM, New York (2006)
21. McMillan, D.W., Chavis, D.M.: Sense of Community: A Definition and Theory. Journal of Community Psychology 14(1), 6--23 (1986)