International Journal of Semantic Computing, Vol. 13, No. 3 (2019), pp. 289–310.
World Scientific Publishing Company
DOI: 10.1142/S1793351X19400129

# User Perception of Situated Product Recommendations in Augmented Reality

Brandon Huynh*, Adam Ibrahim†, Yun Suk Chang‡,
Tobias Höllerer§ and John O'Donovan¶

*Department of Computer Science, University of California*
*Santa Barbara, California, USA*
**bhuynh@cs.ucsb.edu*
†*ai@cs.ucsb.edu*
‡*ychang@cs.ucsb.edu*
§*holl@cs.ucsb.edu*
¶*jod@cs.ucsb.edu*

Augmented reality (AR) interfaces increasingly utilize artificial intelligence systems to tailor content and experiences to the user. We explore the effects of one such system — a recommender system for online shopping — which allows customers to view personalized product recommendations in the physical spaces where they might be used. We describe results of a $2 \times 3$ condition exploratory study in which recommendation quality was varied across three user interface types. Our results highlight potential differences in user perception of the recommended objects in an AR environment. Specifically, users rate product recommendations significantly higher in AR and in a 3D browser interface, and show a significant increase in trust in the recommender system, compared to a web interface with 2D product images. Through semi-structured interviews, we gather participant feedback which suggests AR interfaces perform better due to their ability to view products within the physical context where they will be used.

*Keywords*: Augmented reality; recommender systems; user interfaces; recommendation quality; recommender trust.

## 1. Introduction

Recommender systems first emerged over two decades ago and have since become standard tools for dealing with information overload [1– 3]. Major retail stores such as Amazon.com have a heavy focus on data-driven marketing, of which collaborative and content-based recommender systems are a core part. About 35% of sales on Amazon, and 75% of movies watched on Netflix are derived from recommendations [4]. The vast majority of recommendations for online retailers are delivered through email or in the traditional web browser interface. Interface technology, however, is developing rapidly: global revenues of Augmented Reality (AR) and Virtual Reality (VR) markets are expected to grow to over $162 billion in 2020 [5]. Heavy investment in AR and VR by major companies such as Apple,

Alphabet, Facebook, and Microsoft will mean that smaller, higher quality devices will become available at lower cost to consumers. Large retailers such as Amazon and IKEA are exploring and introducing new AR driven shopping experiences.

While there has been progress on in-store AR technology to improve shopping experiences, e.g. [6], less work has been done on the concept of in-home shoppers taking advantage of what we call "situated recommendations," whereby *personalized* recommendations of products are placed virtually where the real product will be used. In particular, we are interested in how people *perceive* recommendations that are situated in AR, and how this perception *differs* from that of traditional recommender system interfaces. We attempt to address the following specific research questions:

(1) **RQ1:** Do users perceive product recommendations in AR differently than in a browser-based UI?
(2) **RQ2:** Are there differences in recommender system trust when presented in AR versus a browser-based UI?
(3) **RQ3:** What is the general sentiment towards an AR recommender system for in-home shopping?

To answer these questions, we conducted a $3 \times 2$ within-subjects lab study ($N = 31$). The study examined the effects of three different interaction modalities: an AR interface, a web browser interface with 3D view controls, and a web browser interface with 2D view controls. We also looked at how users respond to differences in recommendation algorithm quality (either high or low quality recommendations). We measured two key metrics, user ratings of each recommended object (also called perceived accuracy), and user trust in the recommender system. We collected subjective feedback on user perception of the modalities through a post study questionnaire and verbal interviews.

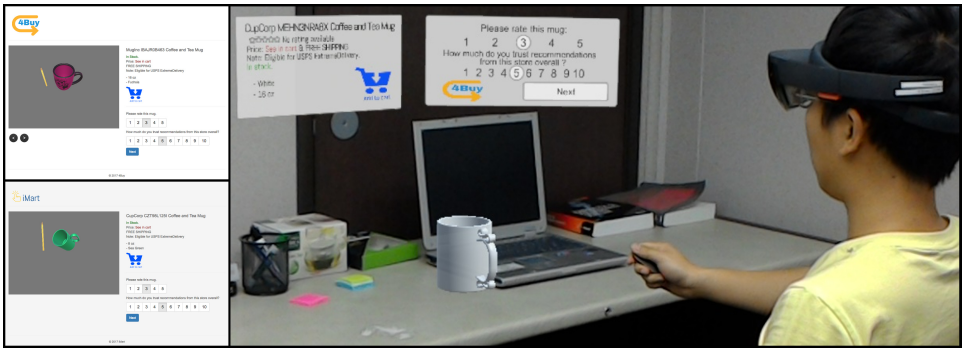


Fig. 1. *Left*: Screen captures of the browser interfaces. For 2D browser, users can look through photos of the mug taken from different predefined angles. For 3D browser, users can freely rotate the mug and view it in any direction. *Right*: Mixed reality screen capture of a user providing feedback to a recommended item.

For the purposes of this study, we implemented a common online shopping user interface across all three modalities to allow for meaningful comparison. To avoid potential novelty effects, study participants undergo significant pre-study training sessions for each modality. Figure 1 shows an overview of our shopping interface. The right image shows a user wearing the HoloLens interacting with a virtual model of a recommended item and providing rating feedback to the system. The left images show the two web browser based interfaces that were tested in the study. In the web browser UIs, participants interact either by rotating the object with the mouse (3D), or clicking through static images (2D).

## 2. Background

Our study combines facets from multiple research fields, including human–computer interaction (HCI), recommender systems, and cognitive science. A discussion of the relevant literature in each area is presented here, to frame our contribution in the context of existing research.

### 2.1. *Augmented reality in retail applications*

Currently, there are many consumer applications for visualizing products in augmented reality. For example, IKEA uses a mobile AR app to place virtual models of their furniture in the physical world and Lego uses AR kiosks to visualize assembled Lego sets on the corresponding box [7]. Nike created a custom sneaker designer that uses projective AR to overlay designs onto a customer's physical sneakers [8]. Recent work by Stoyonova *et al.* [9] reports on a cognitive study of purchase intent using AR in a shopping scenario, but in contrast to this work, does not have a focus on personalized recommendations, and is situated in a store as opposed to a home shopping scenario. Lu *et al.* [10] perform a study of AR for home shoppers, where selected products can be tried in AR before purchase. Olsson and colleagues [11] present a study of user experiences with AR in a shopping center context and report mainly positive feedback for mobile AR supported shopping.

While there are many other examples of AR for improving shopping experiences [11, 12], to our knowledge there is no existing research that explores how users perceive *personalized recommendations* in this modality. We believe that our results can provide useful insight about this rapidly developing technology and its suitability as a channel for delivering personalized recommendations.

### 2.2. *Augmented reality and recommendation*

Many applications that integrate AR and recommendation use mobile platforms to perform location-based content recommendations. The Yelp monocle[a] for service recommendations is probably the most well-known example of this integration with

[a] https://www.yelp-support.com/article/What-is-Yelp-s-Monocle-feature?l=en_US

AR. Balduini *et al.*'s Bottari system [13] provides personalized, location-based AR recommendations of social media content based on the Twitter network and evaluated the system in an urban area. While these approaches integrate AR and recommendation, they contrast with our approach in that they do not focus on evaluating perception of recommendations in AR compared to traditional UIs.

### 2.3. *Interfaces and decision–making*

Prior research in recommender systems has a strong focus on algorithm performance. Recently however, more research attention is being paid to so called *user-aware* recommendation systems that attempt to improve the user's experience with the recommender system by mechanisms that go beyond predictive accuracy, such as conversation [14], explanations [2, 15, 16], and various different flavors of user interfaces [17–19], interaction designs [20] and evaluations [21–23].

In this study, we are interested in a novel user interfaces aspect — that of the impact of placement of recommended content in physical contexts with augmented reality, on the metrics of accuracy and trust. We are also interested in how the interplay of AI performance (quality of the recommendation) with the choice of user interface influences these metrics. It is likely that user specific factors such as experience with visualizations, recommender systems, or multimodal display technology will impact the observed results. Nilashi *et al.* [24] performed a mixed-model evaluation of recommender system users on two real world e-commerce sites and analyzed the impact of many observed and latent factors on trust and purchase intention. Similar mixed model evaluations for recommenders were performed on a hybrid music prediction system by Knijnenburg *et al.* in [23] and in a system for analysis of commuter traffic data from microblogs by Schaffer *et al.* [22]. In this paper we also apply a mixed-model evaluation, designed to capture user-specific characteristics that impact our performance metrics.

### 2.4. *Trust dynamics in recommender systems*

Understanding and building user trust in predictions is an important goal of most recommender systems. Prior research has studied this from a computational model perspective to improve automated recommendations for collaborative filtering [25] and matrix factor approaches [26]. Others, such as [27, 28], have leveraged network information to build and propagate trust. In contrast to those relatively static approaches, we are interested in real-time human judgements of trust in both the system and its individual item predictions.

Recent work by Harman *et al.* [29] examines trust dynamics in a fictional and controlled online dating scenario under a repeated choice experiment with 200 trials. They found that users quickly learn to identify when poor recommendations are being made and lower their trust accordingly. An interesting aspect of their study looked at a personalized treatment against a non-personalized treatment and found that failures (poor recommendations) in the personalized condition had a more

damaging impact on trust than in the non-personalized treatment. In our experiment design, we evaluated trust dynamics in a similar repeated choice and personalized scenario, but based on a simple home shopping task. A similar study by Yu *et al.* [30] also explores trust dynamics for an automated system under a variety of performance quality conditions. They find that increasing user familiarity with the system decreases the rate of change of trust after successes or failures of the automated system.

Building on the work from [29] and [30], we aim to explore the dynamics of trust for situated product recommendations in AR under conditions of high or low quality recommendations. Our study includes repeated interactions with the system to explore differences in trust dynamics and we hope to see trends similar to those found in the previous two approaches, with poor recommendation conditions showing less trust with each interaction.

## 3. System Architecture

To test our hypothesis, we implemented online shopping interfaces for each modality as well as a content recommendation system which generates a set of distinct high and low quality recommendations based on user profile data. These recommendations are distributed evenly across the three modalities, where they are rendered using the Unity game engine. During the study, users interact with each modality and give ratings which are sent back to the server to be recorded.

### 3.1. *Browser interface*

We implemented a simple e-commerce graphical interface in a web browser (see Fig. 1). The interface shows the recommended object, the store logo, and generic text descriptions of the object. The browser interface is broken up into two presentation modalities: 2D browser and 3D browser. In the 2D browser modality, item recommendations are presented as a set of 2D pictures of the product taken from different angles. Users cycle through these pictures by clicking on the arrow buttons below the image. A pencil is shown in the images to provide a point of reference for scale. Users can rate the items by clicking on the radio buttons provided on the right side of the image. The 3D browser modality displays a 3D model of item recommendations that users can interact with. In this modality, users click and drag within the display window to rotate the object about its central $X$ and $Y$ axes. Users can provide ratings in the same way as the 2D browser. Note that both kinds of browser interactions take place on a traditional computer and monitor.

### 3.2. *Augmented reality interface*

For the AR interface, we use a Microsoft HoloLens device. Our application uses the devices' Spatial Mapping API to map the environment and situate virtual products

and UI elements within the environment. We use HoloLens' World Anchor system to fix the recommended item and UI elements in the same position throughout the study. Users are able to walk around and look at the virtual items from different directions and provide feedback via the rating interface, presented through two panels as shown in Fig. 1. The graphical interface is bare-bones, only displaying the store logo and generic text descriptions similar to the browser implementations.

For interacting with the interface, we implemented a 3D cursor using a raycast formed by the user's head gaze direction. We define head gaze direction as the forward direction of the headset. Using the 3D cursor, users can aim and click on the rating panel. Although the HoloLens device supports hand gestures for clicking, we opted to use the HoloLens bluetooth clicker. This provides a fairer comparison to the browser interface, as there may be additional effects introduced by gesture-based interaction.

### 3.3. *Content-based recommendation*

In order to generate personalized recommendations, we use an algorithm based on attribute Preference Elicitation (PE) and Multi-Attribute Utility Theory (MAUT). The item attributes considered by the recommender are *color*, *shape*, and *size*. We provide a validation for this choice of attributes in the Experimental Design section. For each attribute, we compute the error between the recommender choice for that attribute, denoted as *recAttChoice*, and the user's preference, *userAttPref*.

If the attribute considered is a binary attribute (here, shape or size), let *recBin Att* $\in \{-1, 1\}$ denote the recommender's choice for that attribute, where each possible value corresponds to a specific high level choice, as indicated in Table 1. The reported user preference for that attribute, *userAttPref*, has values in $[1, 5]$. In the pre-study questionnaire, values of 1 and 5 corresponded to a strong preference toward one of the possible values of the binary attribute, values of 2 and 4 to a slight preference, and 3 to no preference. The error can be computed from those two variables as:

$$Err_{recBinAtt} = L(recBinAtt, sgn(userAttPref - 3)) \cdot W_{recBinAtt}, \tag{1}$$

where $L(\hat{y}, y)$ is the 0–1 binary loss function which equals 1 if $\hat{y} \neq y$ and 0 otherwise, and the weight is defined based on the importance given by the user to that

Table 1. Attributes for item classification.

| Attribute | High-level choice | Value |
|---|---|---|
| Shape | Non-cylindrical | −1 |
| | Cylindrical | 1 |
| Size | Small | −1 |
| | Large | 1 |
| Color | Disliked | −1 |
| | Neutral | 0 |
| | Liked | 1 |

particular attribute as $W_{recBinAtt} = (userAttPref - 3)^2$. Note that 3 is subtracted from the user's rating in order to turn the values ranging from 1 to 5 from the pre-study questionnaire into values in $[[-2, 2]]$. The sign function is then applied to map the user's rating to its corresponding value in Table 1. This essentially decouples the user's preference into a binary choice and a weight.

Color was treated slightly differently: users were asked how much color weighed in their decision, and then asked to choose colors they liked and colors they disliked among 13 colors; colors not selected are considered neutral. The estimation of the error on a given color choice by the recommender *recColChoice* therefore is

$$Err_{recColChoice} = (\mathbb{1}_{DislikedCol}(recColChoice) + 0.5 \cdot \mathbb{1}_{NeutralCol}(recColChoice)) \cdot W_{col}, \quad (2)$$

where $\mathbb{1}_A(x)$ is the indicator function on set $A$ defined as 1 if $x \in A$ and 0 if $x \notin A$. The weight $W_{col}$ has the same range of values as the weights used for the binary attributes, and the 0.5 factor for a neutral color ensures that picking a neutral color will yield an error superior to that of a liked color and inferior to that of a disliked color. The overall error is then obtain by summing the individual attribute errors

$$\text{Error} = Err_{recColChoice} + Err_{recShapeChoice} + Err_{recSizeChoice}. \quad (3)$$

It is worth noting that the errors can easily be turned into utility measurements by replacing $L$ by $(1 - L)$ in Eq. (1) and $\mathbb{1}_{DislikedCol}$ by $\mathbb{1}_{LikedCol}$ in Eq. (2).

The personalized recommender system computes all the possible values of the total error based on the user weights, and stores for each value of the total error the incorrect attributes contributing to that value of the total error. There are $2^{\text{card}(\{W_{BinAtt}: W_{BinAtt} \neq 0\})}$ possible ways to get an error from potentially incorrect binary attributes that the user indicated having a preference for, and $3^{L(W_{col},0)}$ different possible values for the error from the color. Note that the error is degenerate; that is, different choices of incorrect attributes may yield the same value of the total error, which can be used to diversify the recommendations. The recommender system can then use a look-up table to show products in order of increasing error.

We define *high quality* recommendations as ones for which every attribute satisfies the user's preferences, and *low quality* recommendations as ones maximising the error, i.e. none of the attributes satisfy the user's preferences. Extreme values of the error were picked to avoid issues with parameter tuning in the recommender algorithm (e.g. power used in the calculation of the weights), which may depend on the granularity of the preference scales or the different possible interpretations of the scale labels by the users. In a longer sequence of interactions, or real-world deployment of the system, less strict parameters could be adopted to improve diversity and novelty of predicted items. An example of the two classes of recommendations are shown in Fig. 2.


(a) Good recommendations (b) Bad recommendations

Fig. 2. Example of recommendations for a user indicating preference for large, non-cylindrical mugs with navy, lime, cyan as liked colors and indigo, magenta, fuchsia as disliked colors.

## 4. Experiment Design

Our main study had a $3 \times 2$ within subjects design with counterbalancing. The two independent variables were *UI modality* and *algorithm quality*, and the main dependant variables were item ratings (accuracy) and user trust in the recommender system. A preference profile was gathered from each participant in the experiment several days before the in-person study, via a Qualtrics[b] online survey. In this preference elicitation questionnaire, participants were asked for basic demographic information and experience with recommenders and AR/VR technology. They were also asked to select preferences for each of the classification dimensions for our domain items. These consisted of size (large or small), mug shape (cylindrical or non-cylindrical) and color preference. For color, participants were shown images of 13 coffee mugs of different colors and were asked to select their favorite 3 and least favorite 3. This information was stored on a server which computed sets of high and low quality recommendations for each user, based on the algorithm previously described.

To compare the effect of recommendation quality among the three different modalities, two different virtual retail stores were created: *4Buy* and *iMart*. 4Buy always attempts to provide high quality recommendations and iMart always attempts to recommend items from the database that the user will dislike. Distinct logos for 4Buy and iMart were visible in each modality (see Fig. 1) to allow users to recognize which store they are in and form different perceptions of trust for each store.

### 4.1. *Item-space classification*

As it is difficult to find free high-quality 3D models, we chose to modify the models on the fly to provide variance in recommendations. We began with a total of 18 different models, and applied transforms over size and color parameters to provide different virtual mugs for participants. The patterns on the mugs varied. To ensure that the pattern variable would not impact user preference more than the controlled features (size, shape and color), an MTurk study of 110 users was performed where each participant provided ratings between 1 to 100 for each of the four features. The mean and standard deviation for these ratings are found in Table 2. We found no

[b] https://www.qualtrics.com/

Table 2. Validation of item classification features from 110 participants in an online survey.

| | Mean | Std. Dev. |
|---|---|---|
| Color | 57.84 | 25.49 |
| Pattern | 58.45 | 27.5 |
| Size | 71.27 | 23.15 |
| Shape | 60.75 | 26.17 |

significant difference between pattern and the other features and so assume that manipulation of the other three features will be sufficient to provide good or bad recommendations based on the user profile. This is further confirmed in our results which show user ratings for good recommendations are significantly higher than those for bad ones. Different patterns in the mugs can contribute to *novelty* and *diversity* in recommendations, but overall quality can still be controlled in a meaningful way through manipulations on the other features.

### 4.2. *Experimental procedure*

The experiment was conducted at an American university campus. Participants were assigned to particular orderings for each condition. Participants were given a brief introduction to the study by the experimenter. They were provided with a simple background story as follows: "You have just broken your coffee mug and are looking online to shop for a new one. You will shop at two different stores using a variety of their interfaces".

For the AR condition, participants were given a training task where they had to observe several virtual items and use interaction in AR to provide feedback ratings. Once comfortable with the AR environment and rating procedure, they began the main rating phase. Here, they were shown a sequence of three recommendations, either from iMart (low quality) or 4Buy (high quality). They were asked to walk around and inspect the items, and then provide a rating for how much they liked the recommended item on a scale of 1 to 5, and how much they trusted the system's current ability to provide good recommendations on a scale of 1 to 10. There was no time limit imposed during the rating phase. Participants typically took less than 30 s to provide a rating, irregardless of the modality. Similar training steps were performed for the browser-based conditions.

Participants complete all three conditions for a given store (and recommendation quality) first, before repeating the conditions in the same order for the other store. We alternated which store the participants start with. Once all conditions were complete, participants completed a post study questionnaire and were given a brief post study interview by the experimenter. In the post study questionnaire, participants were asked to rate their overall trust for each recommender, the helpfulness of the recommender, how much they liked interacting with each modality, and how much they liked each store overall. In the interview, participants were asked about

their thoughts on the AR device, and whether they would choose to use it over the other modalities in a real world shopping scenario.

### 4.3. *Novelty effect*

Since AR is a new and emerging technology, and there is a "wow factor" with cutting edge devices such as the Hololens, novelty effects will always be challenging to deal with. To mitigate novelty effects in the experiment, participants were allowed up to 10 min to familiarize themselves with each modality. In the AR condition, participants played with the built-in holograms application on the Hololens device. Note that this familiarization period takes place before the training task begins.

After the experiments, we compared performance between the participants who started with the AR condition versus those who started with the browser-based conditions. We ran paired *t*-tests on our key metrics but found no significant differences between the two groups, giving us confidence that our balancing and familiarization procedures were helpful in controlling novelty effects of the Hololens device in the AR condition. This was further supported through post study interviews, during which participants reported that the familiarization period helped them to "get comfortable" using the AR headset.

## 5. Results

To answer our research questions, we looked at user ratings for individual product recommendations and overall trust in the recommender system. We examined differences in ratings across each modality in order to assess relevant effects on user's perception of recommendations. Additionally, we examine self-reported UX metrics from a post study questionnaire and verbal interview.

### 5.1. *Participants*

In total, 31 participants completed the in-person study. Data from three participants were removed due to being provided incorrect instructions on the rating system. These participants misunderstood the task and rated other aspects such as the design of the logo. We also removed two additional participants due to system failure of the HoloLens during the experiment, leaving a total of 26 for analysis. Participants had a median age of 23, mean age of 27 with std. deviation of 9.58. 77% were male and 23%

Table 3. Accuracy: Pairwise comparison between modalities.

| Contrast | Estimate | SE | df | *t* ratio | *p* value |
|---|---|---|---|---|---|
| Browser2D – Browser3D | −0.2641 | 0.1095 | 440.51 | −2.413 | 0.0428 |
| Browser2D – AR | −0.2791 | 0.1091 | 438.92 | −2.557 | 0.0293 |
| Browser3D – AR | −0.0150 | 0.1091 | 438.42 | −0.138 | 0.9896 |

*Notes*: Results are averaged over the levels of: Recommendation quality.
*p* value adjustment: Tukey method for comparing a family of three estimates.

female. All had at least some college education. Participants were recruited through a user study pool at the university and were paid $10 for the study, which lasted about 40 min.

### 5.2. *Do users perceive product recommendations in AR differently than in a browser-based UI?*

To begin our analysis, we looked at the average accuracy ratings within each condition. The resulting data is graphed in Fig. 3. We tested for significance using paired $t$-tests.

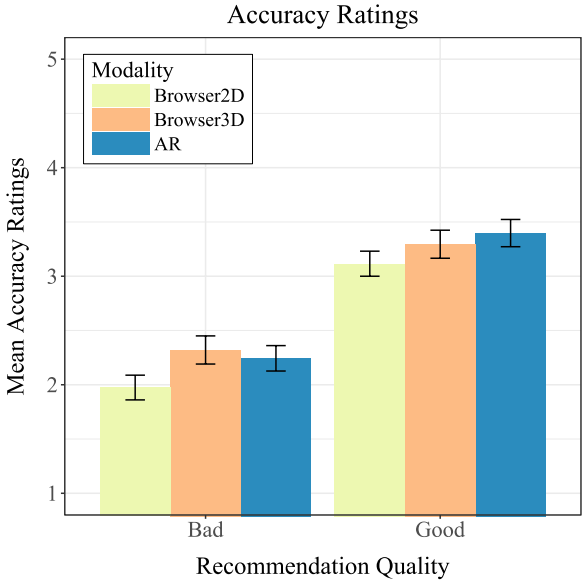


Fig. 3. Mean accuracy rating with standard error.

For these ratings, our initial hypothesis was that increased reality and immersion provided by the AR modality would amplify users' perception of recommendation accuracy. More realistic inspection methods might cause users to have a greater awareness of how well a product fits their preferences. Thus, we expected bad recommendations to be rated lower in AR compared to browser based methods, and likewise good recommendations would be rated higher in AR.

When looking at ratings in the bad recommender, we found a significant difference between the 2D modality ($\mu = 1.97$) and the 3D modality ($\mu = 2.32$) conditions; $p = 0.024$. There was almost significance between 2D modality and the AR modality ($\mu = 2.24$); $p = 0.064$. In the good recommender, we found significance between the 2D modality ($\mu = 3.12$) and the AR modality ($\mu = 3.4$); $p = 0.035$, but not between 2D and 3D modalities.

Additionally, we see a significant difference in ratings between the good and bad recommenders for all three modalities (all $p < 0.0005$). This gives us confidence that our recommendation algorithm is correctly providing high and low quality recommendations based on the user's preferences.

These results appear to reject our hypothesis. Irregardless of recommender quality, AR and 3D modalities seem to improve perception of recommendations. However, the signal does not appear consistently between bad and good recommenders. Thus, to look at the effects of each modality across both good and bad recommender conditions, we opted to perform further analysis using linear mixed effects models. Specifically, the modality type and recommendation quality are modeled as fixed effects, while participants and item design are modeled as random effects.

To validate this approach, we assessed the fit of our models using pseudo-$R^2$ values [31]. Marginal pseudo-$R^2$ was computed for fixed effects, and conditional pseudo-$R^2$ for random effects. For the accuracy model, the marginal pseudo-$R^2$ was 0.216 and the conditional pseudo-$R^2$ was 0.369. Additionally, mixed effects models assume that the residuals of the model are normally distributed. We plotted the residuals of each model as Q–Q plots to check this assumption and found that the residuals fall about a fairly straight line, suggesting normality. These plots can be found in Fig. 4. Finally, we created separate models where Modality and Recommendation Quality were modeled as having an interaction effect. We performed a likelihood ratio test against these to determine any significant interaction effects, but did not find any significant inter-dependence between them thus we did not include interaction effects in our models.

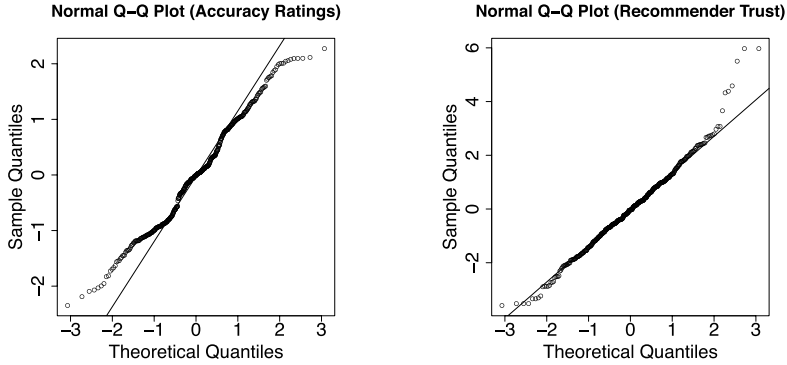


Fig. 4. Q–Q plots of residuals of LME models for accuracy and trust.

The full pairwise comparisons between each modality are shown in Table 3. These tables describe the difference in ratings after averaging over the levels of recommendation quality and performing $p$-value adjustment using the Tukey method. Here, we can see a significant difference between Browser2D and the AR modalities

($p = 0.0293$), as well as between Browser2D and Browser3D ($p = 0.0428$). This provides further evidence that AR may improve user perception of product recommendations.

When comparing AR against the 3D interface, pairwise comparisons within our model did not show a significant difference in product rating. We believe that this result was due to a hidden variable created through differing levels of control in the interaction. In the 3D browser, users could rotate the items and view them from all angles. However, in the AR condition, the item was in a fixed position, and therefore could not be viewed from the bottom angle, since it was positioned on a table. During the verbal interview, three participants mentioned they prefer the 3D view because it "allows you to see the mug in every possible orientation".

### 5.3. *Are there differences in recommender system trust when presented in AR versus a browser-based UI?*

This question focuses on the perception of algorithm quality within the different modalities. Similar to the product ratings, we hypothesized that the AR condition could help improve user awareness of a recommendation algorithm's performance, leading to lower ratings for the low quality recommender and higher ratings for the high quality recommender compared to the other modalities.

Our analysis on trust ratings mirrored the methods used for product ratings in the previous section. In Fig. 5, you can see the graphed trust ratings. In the bad recommender, we found significant differences between 2D ($\mu = 3.51$) and 3D


Fig. 5. Mean trust ratings with standard error.

($\mu = 4.27$); $p < 0.001$, as well as 2D and AR ($\mu = 3.95$); $p = 0.034$. In the good recommender, we also found significance between 2D ($\mu = 5.41$) and 3D ($\mu = 5.95$); $p = 0.005$, and also between 2D and AR ($\mu = 5.91$); $p = 0.008$. Again, we see a significant difference between the good and bad recommenders for all 3 modalities (all $p < 0.0005$). Figure 5 clearly show that users perceived a difference between good and bad algorithms in all conditions. For example, participants in the 2D browser condition rated trust in the iMart (low quality recommender algorithm) at 3.51 and 4Buy at 5.41, which is a relative improvement of 54% over the iMart algorithm.

We again used linear mixed models to analyze trust ratings across recommendation quality. We performed the same steps to validate the model as in the previous section. For the trust model, marginal pseudo-$R^2$ was 0.184 and conditional pseudo-$R^2$ was 0.555. Table 4 is the resulting pairwise comparisons. In particular, we highlight the differences between Browser2D and the AR modalities which are significant for trust ratings ($p = 0.0167$). Ultimately, the results we found did not support our hypothesis. Instead, our results suggest that Trust is improved for the AR and 3D modalities, despite the differences in recommender quality.

Table 4. Trust: Pairwise comparison between modalities.

| Contrast | Estimate | SE | df | $t$ ratio | $p$ value |
|---|---|---|---|---|---|
| Browser2D – Browser3D | −0.6474 | 0.1697 | 442 | −3.815 | 0.0005 |
| Browser2D – AR | −0.4679 | 0.1697 | 442 | −2.757 | 0.0167 |
| Browser3D – AR | 0.1795 | 0.1697 | 442 | 1.058 | 0.5410 |

*Notes*: Results are averaged over the levels of: Recommendation Quality.
$p$ value adjustment: Tukey method for comparing a family of three estimates.

#### 5.3.1. *Trust dynamics*

We build on recent work in recommender systems research by examining the perception of trust in the recommender system over time, for the high and low quality recommendation algorithms. We plot these trends for each modality in Fig. 6. The first clear effect from this is the separation between the high and low quality recommendation strategies (4Buy and iMart). This provides further support of the effectiveness of our recommender system, despite it's relative simplicity.

Looking at the slopes of these distributions, all but one of the data points for the low quality recommender (iMart) follow a downward sloping trend, while those for the high quality recommender (4Buy) have an initial upwards trend. This supports similar results found in [29], in which users trust in the system dropped swiftly following repeated interactions with poor recommendations. This is further supported by our post study questionnaire, where participants significant preferred 4Buy over iMart.

Additionally, we see a decrease in the rate of change of trust after repeated interaction, between the first and second recommendation to the second and third

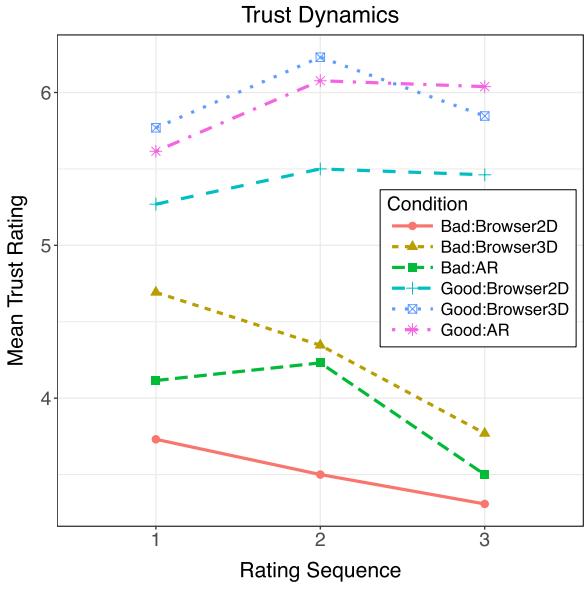


Fig. 6. Dynamics of trust for each condition.

recommendatio. However, this is only present for the good recommender system. This trend is similar to results found in [30].

For further analysis, we used Analysis of Covariance (ANCOVA) to compare trust ratings categorized by condition, controlling for rating sequence. Our test did not find any significant interaction between rating sequence and condition ($p = 0.515$), suggesting that there aren't any significant differences in the slopes of the regression lines between each condition. We believe this may be due to the limited amount of repeated interactions. Additionally, we looked at whether trust changed over time for high and low quality recommendations regardless of modality. We analyzed the average ratings for the first and last modality used for both the low and high quality recommender using a paired t-test, but did not find a significant difference in either case ($p = 0.783$).

### 5.4. *What is the general sentiment towards an AR recommender system for in-home shopping?*

Our primary source of analysis for this research question are through a post-study questionnaire and semi-structured verbal interview conducted immediately after the experiment.

#### 5.4.1. *Post-study questionnaire*

The results of the post questionnaire are shown in Fig. 7. The leftmost plot shows the perceived trust in the system's recommendations broken down for each of the six

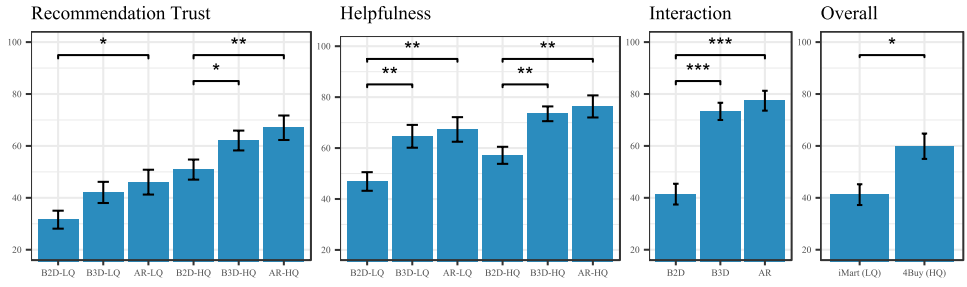


Fig. 7. Mean subjective ratings from the post study questionnaire with standard error. Participants were asked to rate how much they trust the recommendations, how helpful each store's interface was, the interaction quality of each modality, and overall preference for each store. Brackets show the level of significance between particular values ($^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$). Additionally, there was significance ($p < 0.01$) in recommendation trust between each HQ modality and their LQ counterparts.

conditions. Here the browser-based conditions are abbreviated to B2D and B3D, and algorithm quality is represented as HQ or LQ for high and low respectively.

The first point to note is that the questionnaire responses for trust in the recommendations align well with the observed ratings during the experiment, with both AR-HQ and B3D-HQ showing a significant rating improvement of about 20% over the traditional UI B2D-HQ. There was no significant difference between the B3D and AR conditions. However, our post study interviews revealed that people either had a strong preference for the 3D-browser condition or the AR condition. Those who strongly preferred AR, tended to mention the value of being able to see the item in real-world context (situated recommendations), while those who preferred the 3D browser version tended to like the familiarity of the interface for shopping.

Perceived helpfulness of the stores was also evaluated and showed a similar trend to trust, with AR and B3D having significant rating improvement over B2D for both recommender algorithms (LQ and HQ). However, the differences between recommender quality (LQ and HQ), was not as pronounced as it was on the trust metric. We believe this is an indication that users were considering other aspects than recommendation quality for their decisions on helpfulness, such as the quality of the UI design.

Figure 7 also shows results for perceived quality of interaction with the system. As expected, both AR and B3D received very positive ratings. This is consistent with our interview feedback where participants preferred B3D almost as much as the AR condition due to better inspection capabilities. Participants were also asked to rate each store overall. Here we see that participants did perceive the difference in algorithm quality across the two stores. The store with high quality recommendations (HQ) showed a 50% improvement over the LQ store. This was consistent with our observed ratings-based results.

5.4.2. *Verbal interview*

Participants were given a verbal interview immediately after the post-study questionnaire, which typically lasted about 5 min. For the verbal interview, we provided some structure by asking in order the following questions:

(1) What did you think of the HoloLens?
(2) Would you use this in a real world shopping scenario?
(3) Did you find the ability to walk around and view objects in a real world environment to be helpful or distracting in your shopping decisions?

Participants were asked all three questions regardless of if they had already mentioned any related comments in a prior question. This means that some participants touch on the same topic multiple times over the course of the interview.

When asked what they thought of the HoloLens, participant opinions were generally quite high. Most participants thought the AR device was cool, interesting, and enjoyable. Even before being prompted in question 3, participants often commented about the ability to walk up to the object, see it from different angles, and compare the object to it's surroundings. The most common negative opinion was the color fidelity of the display. Five participants had complaints about colors being washed out and difficult to perceive. Other complaints include the limited field of view, and discomfort due to weight of the device.

When asked about whether they would consider using the AR interface in a real world shopping scenario, participants responses were very positive. All but five of the participants reported that they would choose to use the AR system if it were available to them. Out of many different reasons cited, the most common was the "try before you buy" reason — to visualize and interact with the item in the context where it is to be used. An equally common opinion was the desire to use the interface for purchasing certain types of items. Typically, participants mentioned it would be very useful for purchasing large items such as furniture. A few participants commented that they would use AR shopping once the interface was improved. In this case, they felt the interface was very useful for shopping but wanted a more "polished" user interface design. The participants who did not want to use it argued that the interaction was not sufficient and that the 3D browser version allowed for a better inspection of the item. Additionally, participants reported feelings of frustration and discomfort that would dissuade them from using the device.

A summary of the most common responses to the first two questions can be found in Fig. 8. Note that participants may have commented on multiple topics during the course of answering each question.

For the third question, we were able to bucket the responses into four categories. 13 participants said the ability to view products in-situ was very helpful, while seven participants said it was only slightly helpful. Five participants said it was neither helpful nor distracting to the shopping task. Only one person said it was somewhat distracting to see while shopping. This same participant ultimately

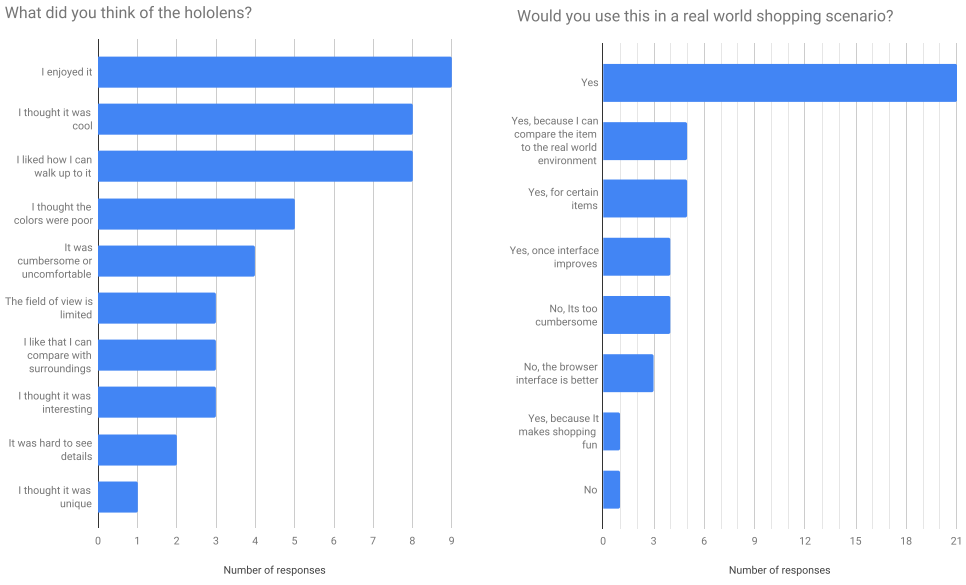


Fig. 8. Summary of common responses in the verbal interviews for questions 1 (left) and 2 (right).

commented that they preferred the browser version because it was much faster and more efficient to use.

Ultimately, these responses show a lot of positive sentiment for the future of AR recommender systems. Our study participants don't view *in situ* shopping as more distracting or more difficult. The biggest concerns came from hardware limitations or lack of polished designs, issues that would surely be addressed by future commercialization efforts and design improvements that were not the focus of our controlled test interfaces. Already, recently deployed or announced products such as the MagicLeap One[c] or HoloLens 2[d] are lighter, more comfortable, and have twice the field of view. These additional technological capabilities should lead to better audience acceptance regarding these issues.

### 5.5. *Demographic analysis*

We wanted to look at potential differences between demographics in their experience of the AR recommender system. We recorded participant demographics and looked at mean trust and accuracy ratings between each demographic group. We grouped participants by age, gender, and their familiarity or prior experience with AR devices.

When looking at gender, we had 20 males and six females take part in our study. Our analysis showed that females gave higher product ratings ($\mu = 2.96$) in AR

[c] https://www.magicleap.com/magic-leap-one
[d] https://www.microsoft.com/en-us/hololens

conditions when compared to males ($\mu = 2.66$); $p = 0.045$, using Welch's $t$-test. However, there were no significant differences for mean trust ratings. Additionally, females reported having less familiarity with the modality than males. When asked on a scale of 1 to 5 about their prior experience with AR, the mean for females was 3.12 compared to 3.4 for males.

To look at age, we grouped participants into two age bins around the median, and a similar analysis was performed. We performed similar $t$-tests for mean ratings and mean trust but did not find any significant differences between the two age groups. Additionally, older participants tended to have more prior AR experience, with an average rating of 3.21 vs. 3.0 for the younger participants.

Finally, we separated participants based on their familiarity with AR into two groups: those with low or no experience with AR, and those with some prior AR experience. The low experience group was composed of the 12 participants whose prior experience with AR was 1 or 2 out of 5, leaving 14 participants for the other group. Using Welch's $t$-test, we found that participants with little AR experience had significantly higher product ratings ($\mu = 2.91$) compared to those with some AR experience ($\mu = 2.57$); $p < 0.005$. They also had higher trust ratings ($\mu = 5.21$) compared to those with more AR experience ($\mu = 4.51$); $p < 0.005$.

To help explain these results, we look to participant comments in their post study questionnaire. Participants who have more AR experience tend to be more critical of the AR device's limitations, noting things like the weight of the device or the poor resolution of the display. Whereas those who are newer to the interface are more excited about it's potential, and are more willing to forgive these faults.

## 6. Discussion

The results from our study contribute to an emerging body of work focused on understanding user perception of AR with predictive AI systems such as recommender systems. Throughout the study, AR and Browser3D modalities performed on par with each other, whereas both tended to improve ratings and other metrics compared to Browser2D. Participants generally fell into two camps, those preferring Browser3D and those preferring AR.

Many of the verbal interview responses seem to indicate that participants appreciate qualities from both mediums. In the case of AR, participants enjoy being able to visualize a product in a real world context and grasp the actual scale of the object. However, AR is marred by issues with a low quality display and headset discomfort. 3D on the other hand is quick and easy to use, and still allows users to view recommended products from a variety of viewing angles.

While some of these problems will be solved in future iterations of AR devices, it's important to understand what the role of interaction should be moving forward. It's clear that users are accustomed to browser-based interaction methods. For many shopping experiences, they may prefer it over an AR experience. However, AR has potential to excel when delivering recommendations that have great impact on daily

life, or where scale and contextual information is important, such as home appliances and interior design. These qualities should be emphasized and communicated when designing for the future of AR driven recommender systems.

## 7. Conclusion

This paper presented a study that to our knowledge is the first empirical analysis of the effects of Augmented Reality interfaces on the perception of recommender systems. A $3 \times 2$ within subjects experiment assessed user perception of high and low quality personalized recommendations in three modalities: Augmented Reality with recommended items placed in a real world scene, web browser with 2D images, and web browser with 3D interaction. Quantitative metrics for product ratings and recommender trust were assessed, along with perception of the system through a post study questionnaire and verbal interview.

Results of our main research questions show that overall product ratings for recommended objects, and trust in the recommender, are significantly higher in AR and interactive 3D than in a traditional browser UI. However, there is no significant difference in either metric between interactive 3D and AR modalities. Furthermore, people perceive differences between high and low quality algorithms in all three modalities, but there is no significant trend that suggests better awareness of quality differences in AR. Finally, a majority of participants preferred to use AR over browser based interfaces for product recommendations, finding it helpful for visualizing in the context where it will be used.

## Acknowledgments

This work was funded in part by the United States Department of the Navy, Office of Naval Research grant #N00014-16-1-3002 and the National Science Foundation grant IIS-1845587.

## References

[1] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom and J. Riedl, Grouplens: An open architecture for collaborative filtering of netnews, in *Proc. ACM Conf. Computer Supported Cooperative Work*, 1994, pp. 175–186.

[2] J. L. Herlocker, J. A. Konstan and J. Riedl, Explaining collaborative filtering recommendations, in *ACM Conf. Computer Supported Cooperative Work*, 2000, pp. 241–250.

[3] B. Sarwar, G. Karypis, J. Konstan and J. Riedl, Item-based collaborative filtering recommendation algorithms, in *Proc. 10th Int. Conf. World Wide Web*, 2001, pp. 285–295.

[4] V. Tamturk, *The-ROI-of-Recommendation-Engines*, 2017, http://bit.ly/2nW2aUz, accessed on 1 April 2017.

[5] Zacks Equity Research, *Is Apple Looking to Expand in the Augmented Reality World?* 2017, http://bit.ly/2oTwjBv, accessed on 1 April 2017.

[6] S. Erickson, *Microsoft HoloLens and Lowe's, Working to Redefine Your Next Home Renovation*, 2017, http://bit.ly/2owhIju, accessed on 2 April 2017.

[7] D. Williams, *3 Retail Giants Who Used Augmented Reality to Sell*, 2017, http://bit.ly/1Ufnaeq, accessed on 2 April 2017.
[8] K. N. Jr., *Nike Store in Paris Lets Customers Test Sneaker Colors Using Augmented Reality*, 2017, http://bit.ly/2jvcuSe, accessed on 2 April 2017.
[9] J. Stoyanova, R. Goncalves, A. Coelhc and P. Brito, Real-time augmented reality shopping platform for studying consumer cognitive experiences, in *2013 2nd Experiment@ Int. Conf.*, 2013, pp. 194–195.
[10] Y. Lu and S. Smith, Augmented reality e-commerce system: A case study, *J. Comput. Inform. Sci. Engi.* **10**(2) (2010) 021005.
[11] T. Olsson, E. Lagerstam, T. Kärkkäinen and K. Väänänen-Vainio-Mattila, Expected user experience of mobile augmented reality services: A user study in the context of shopping centres, *Personal Ubiquitous Comput.* **17** (2013) 287–304.
[12] B. Wang, M. Ester, J. Bu and D. Cai, Who also likes it? Generating the most persuasive social explanations in recommender systems, in *AAAI Conf. Artificial Intelligence*, 2014.
[13] M. Balduini, I. Celino, D. Dell'Aglio, E. D. Valle, Y. Huang, T. Lee, S.-H. Kim and V. Tresp, BOTTARI: An augmented reality mobile application to deliver personalized and location-based recommendations by continuous analysis of social media streams, *Web Semantics: Sci. Serv. Agents World Wide Web* **16** (2012) 33–41.
[14] K. McCarthy, J. Reilly, L. McGinty and B. Smyth, Experiments in dynamic critiquing, in *Proc. 10th International Conf. Intelligent User Interfaces*, 2005, pp. 175–182.
[15] N. Tintarev and J. Masthoff, Effective explanations of recommendations: User-centered design, in *Proc. ACM Conf. Recommender Systems*, 2007, pp. 153–156.
[16] N. Tintarev and R. Kutlak, Explanations — making plans scrutable with argumentation and natural language generation, in *Intelligent User Interfaces (demo track)*, 2014.
[17] J. O'Donovan, B. Smyth, B. Gretarsson, S. Bostandjiev and T. Höllerer, Peerchooser: Visual interactive recommendation, in *Proc. SIGCHI Conf. Human Factors in Computing Systems*, 2008, pp. 1085–1088.
[18] K. Verbert, D. Parra, P. Brusilovsky and E. Duval, Visualizing recommendations to support exploration, transparency and controllability, in *Int. Conf. Intelligent User Interfaces*, 2013, pp. 351–362.
[19] D. Parra, P. Brusilovsky and C. Trattner, See what you want to see: Visual user-driven approach for hybrid recommendation, in *Proc. 19th Int. Conf. Intelligent User Interfaces*, 2014, pp. 235–240.
[20] S. Bostandjiev, J. O'Donovan and T. Höllerer, Linkedvis: Exploring social and semantic career recommendations, in *18th Int. Conf. Intelligent User Interfaces*, 2013, pp. 107–116.
[21] J. L. Herlocker, J. A. Konstan, L. Terveen and J. T. Riedl, Evaluating collaborative filtering recommender systems, *ACM Trans. Inf. Syst.* **22**(1) (2004) 5–53.
[22] J. Schaffer, P. Giridhar, D. Jones, T. Höllerer, T. Abdelzaher and J. O'Donovan, Getting the message? A study of explanation interfaces for microblog data analysis, in *Intelligent User Interfaces*, 2015, pp. 345–356.
[23] B. P. Knijnenburg, S. Bostandjiev, J. O'Donovan and A. Kobsa, Inspectability and control in social recommenders, in *Conf. Recommender Systems*, 2012, pp. 43–50.
[24] M. Nilashi, D. Jannach, O. B. Ibrahim, M. D. Esfahani and H. Ahmadi, Recommendation quality, transparency, and website quality for trust-building in recommendation agents, *Electron. Commer. Rec. Appl.* **19** (2016) 70–84.
[25] J. O'Donovan, B. Smyth, V. Evrim and D. McLeod, Extracting and visualizing trust relationships from online auction feedback comments, in *IJCAI*, 2007, pp. 2826–2831.

[26] S. Fazeli, B. Loni, A. Bellogin, H. Drachsler and P. Sloep, Implicit vs. explicit trust in social matrix factorization, in *Proc. 8th ACM Conf. Recommender Systems*, 2014, pp. 317–320.
[27] P. Massa and P. Avesani, Trust-aware recommender systems, in *Proc. ACM Conf. Recommender Systems*, 2007, pp. 17–24.
[28] R. Guha, R. Kumar, P. Raghavan and A. Tomkins, Propagation of trust and distrust, in *Proc. 13th Int. Conf. World Wide Web*, 2004, pp. 403–412.
[29] J. L. Harman, J. O'Donovan, T. F. Abdelzaher and C. Gonzalez, Dynamics of human trust in recommender systems, in *Proc. 8th ACM Conf. Recommender Systems*, 2014, pp. 305–308.
[30] K. Yu, S. Berkovsky, R. Taib, D. Conway, J. Zhou and F. Chen, User trust dynamics: An investigation driven by differences in system performance, in *Proc. 22nd Int. Conf. Intelligent User Interfaces*, 2017, pp. 307–317.
[31] S. Nakagawa and H. Schielzeth, A general and simple method for obtaining r2 from generalized linear mixed-effects models, *Metho. Ecolo. Evolution* **4**(2) (2013) 133–142.