A Study on User Preferential Choices about Rating Scales

Federica Cena, University of Torino, Torino, Italy

Fabiana Vernero, University of Torino, Torino, Italy

ABSTRACT

Websites usually offer the same rating scale for all users and all tasks, but users can have very different preferences. In this paper, the authors study rating scales from the point of view of preferential choices, investigating i) if user preferences for rating scales depend on the object to evaluate, and ii) if user preferences change after they have rated an object repeatedly, gaining a high level of experience with the evaluated object. The authors first defined a model of rating scales, identifying generic classes based on features like granularity and visual metaphor. Then, the authors had users choose between three scales, one for each class, for rating two objects with opposite features, first in a condition where users had a low level of experience, and then in a condition where their level of experience was high. Results showed that user choices depend on the evaluated objects, while their level of experience influences their overall preferences, but not their choices when they have to rate a specific object. The authors conclude with some insights and guidelines for designers of interactive systems.

Keywords: *Interactive Systems, Rating Scales, User Choices, User Interface, User Studies*

1. INTRODUCTION

Rating scales are visual widgets that are characterized by specific features (e.g. granularity, numbering, presence of a neutral position, etc.) which allow users to provide quantitative input to a system. Each system uses its own different rating scale, with different features such as granularity and visual presentation. Examples of rating scales are stars in Amazon (Amazon), Anobii (Anobi) and Barnes & Noble (Barnes & Noble), thumbs in Facebook (Facebook) and YouTube (Youtube), circles in Tripadvisor (Tripadvisor), squares in LateRooms (LateRooms), bare numbers in Criticker (Criticker). In recommender systems (Adomavicius and Tuzhilin, 2005), users rate items to receive personalized suggestions about other items (similar to the previous ones or liked by similar users).

Understanding how users perceive rating scales, and why they might prefer one to another, is very important for interface designers in order to create more effective and pleasant web sites. This problem can be framed in terms of preferential choices (Jameson, 2012), i.e., when two or more options are available, none of which can be defined as "incorrect", but one of which can be preferred for some reason (e.g., tasks, user skills, usage context, habits, etc. (Jameson et al., 2011)).

Rating scales are widely studied in the literature, especially in survey design (Garland, 1991; Colman et al., 1997; Amoo & Friedman, 2001; Dawes, 2008) and *Human-Computer*

DOI: 10.4018/ijthi.2015010103

Interaction (HCI) (Cosey et al., 2003; van Barneveld & van Setten, 2004; Nobarany et al., 2012; Herlocker et al., 2004), but not in terms of preferential choices, i.e. not focusing on users' decision making process. In the survey design field, scales are compared according to their psycometric properties, i.e., their ability to detect "real" user opinions. In the HCI field, scales are mainly studied from a usability point of view. In a sense, the question that all previous works aimed to answer was: "*What is the scale that measures the best?*". We instead aim to answer a different question: "*What is the scale that users would choose*?", assuming that there may be other criteria beside precision and usability. With this paper, we investigate *how users choose rating scales* when they are offered this opportunity. In particular, we study whether *users prefer different scales for evaluating different objects*, and whether *user's choices change* after they have rated a certain object repeatedly, gaining a higher level of experience with the evaluated object.

To answer these questions, we first analysed existing rating scales in order to define an abstract model, which allowed us to identify three generic "classes" of rating scales. Then, we carried out a user study to investigate user choices with respect to three scales chosen as representatives of each class. According to our findings, user choices are influenced by the evaluated objects and overall preferences for rating scales can change after their repeated use. Based on our results, we formulate some guidelines for systems designers.

The main contributions of this paper are:

- A general model of rating scales;
- The results of a user study on preferential choices about rating scales in a website;
- New insights which can help system designers to include the most appropriate rating scales.

The paper is structured as follows. Section 2 presents a preliminary study to devise a model for describing rating scales. Section 3 clarifies how we chose the scales, the objects to evaluate and the use case for our user study, while Sections 4 and 5 describe the study and its results. Section 6 provides the theoretical background for our research and analyzes related works, offering a systematic comparison with the results obtained by the most relevant ones. Section 7 concludes the paper with some guidelines derived from our results and with some possible directions for future work.

2. A MODEL OF RATING SCALES

Aiming at defining a general model of rating scales, we examined thirteen rating scales described by van Barneveld & van Setten (2004), Gena et al. (2011) and Nobarany et al. (2012): 3-, 5- and 10- point stars, bare numbers, smileys, sliders ranging -10/+10, -1/+1 and 0/10, likert-like scales ranging -10/+10 and 1/5 and 1-, 2-, and 3-point thumbs (see Figure 1). First, we identified a list of features which could be used to describe them, based on the literature and our insights, and we organized it in a model (Section 2.1). Then, observing how these features tend to combine in the examined scales, we identified three general classes of rating scales (Section 2.2).

2.1. Analysis of Rating Scales

van Barneveld & van Setten (2004), Gena et al. (2011) and Nobarany et al. (2012) described rating scales through different features, as reported in Table 1.

If we consider the thirteen rating scales we selected, however, some of these features do not seem useful to distinguish among them: continuity (all scales allow to input only discrete values), use of colour (colour is not used in a semantic way, i.e., to convey meaning), measurement scale (all scales allow only absolute input) and recall support (no information is provided about previously recorder opinions). Conversely, in our opinion, features such as "step", "icon", "point mutability", "positive/negative scale" and "diffusion" should be taken into account, and a feature such as "numbering"

Figure 1. Rating scales from [Van Barnevald and Van Setten 2004], [Nobarany et al. 2012] and [Gena et al. 2011]

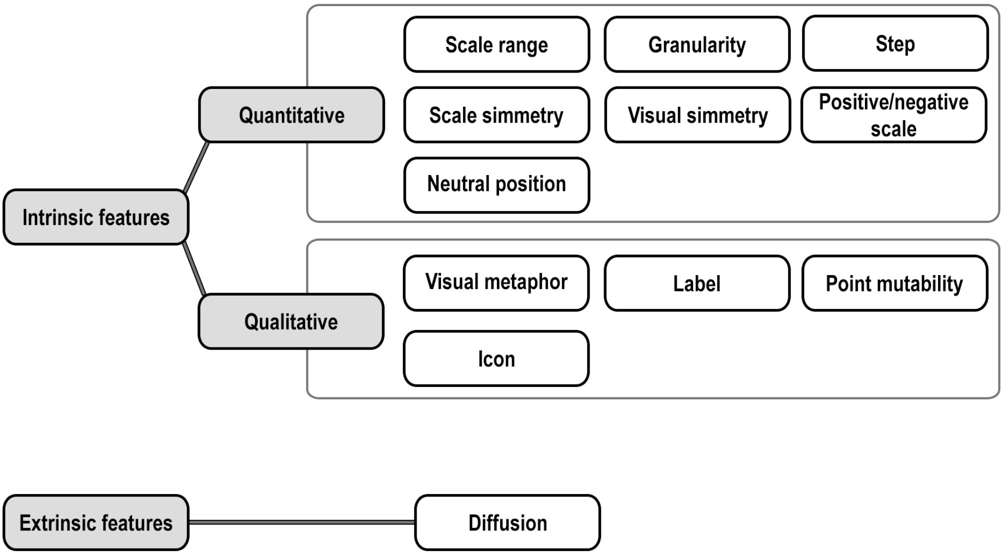


Table 1. Rating scales features identified in previous work

van Barneveld & van Setten (2004)	Gena et al. (2011)	Nobarany et al. (2012)
Presentation form Range Precision Symmetric vs asymmetric scale Continuous vs discrete scale Visual symmetry or asymmetry Use of colour	Visual metaphor Numbering Granularity Neutral position	Measurement scale Recall support

(which refers to the numbers, if any, associated to each position in a scale) should be substituted by a more general concept such as "label".

From these considerations, we therefore derived the following list of rating scale features[1]:

- **Visual Metaphor:** The visualization form which influences the comprehension and emotional connotation of each scale (e.g., a smiley face is a metaphor related to human emotions). Not all visualization forms are metaphors.
- **Granularity:** The number of positions allowed by the rating scale.
- **Scale Range:** The minumum and maximum values of a rating scale (e.g. from 0 to 10).
- **Scale Simmetry:** The presence of specular positive and negative points, explicitly identified with numerical or textual labels.
- **Visual Simmetry:** Symmetry in the visual representation of the positive and negative points.
- **Neutral Position:** The presence of an intermediate point.
- **Step:** The distance among the points in a rating scale.

- **Icon:** The specific image used in a rating scale.
- **Label:** The verbal cue added to a point in a rating scale. It can be a text or a number.
- **Point Mutability:** Indicates whether all points in a rating scale are represented in the same way or not.
- **Positive/Negative Scale:** the presence of only positive ratings or also negative ones.
- **Diffusion:** It indicates how often a certain rating scale is used in the Web and thus is familiar to users.

Then, we grouped rating scale features by means of *card sorting*. Card sorting is a quick and popular method for structuring information, often used in information architecture and user-centred design. We asked three colleagues, experts in HCI and recommender systems, to participate in a joint card sorting session aimed at grouping and structuring the features of rating scales. Each feature was written on a paper card. Participants were asked to sort cards into groups until they were all satisfied about the way cards were organized. Moreover, they were asked to name the different groups they created.

As shown in Figure 2, two main groups emerged from card sorting: intrinsic features, i.e., characteristics relating to the essential nature of a scale, and extrinsic features, i.e., external characteristics not forming an essential part of a thing, originating from the outside. Diffusion was the only extrinsic feature identified by evaluators. Intrinsic features, on the contrary, were further classified into quantitative and qualitative features.

As part of their card-sorting task, evaluators also singled out the most relevant features for the groups they identified: granularity was indicated as the most relevant quantitative feature, while the visual metaphor was deemed especially important among the qualitative features.

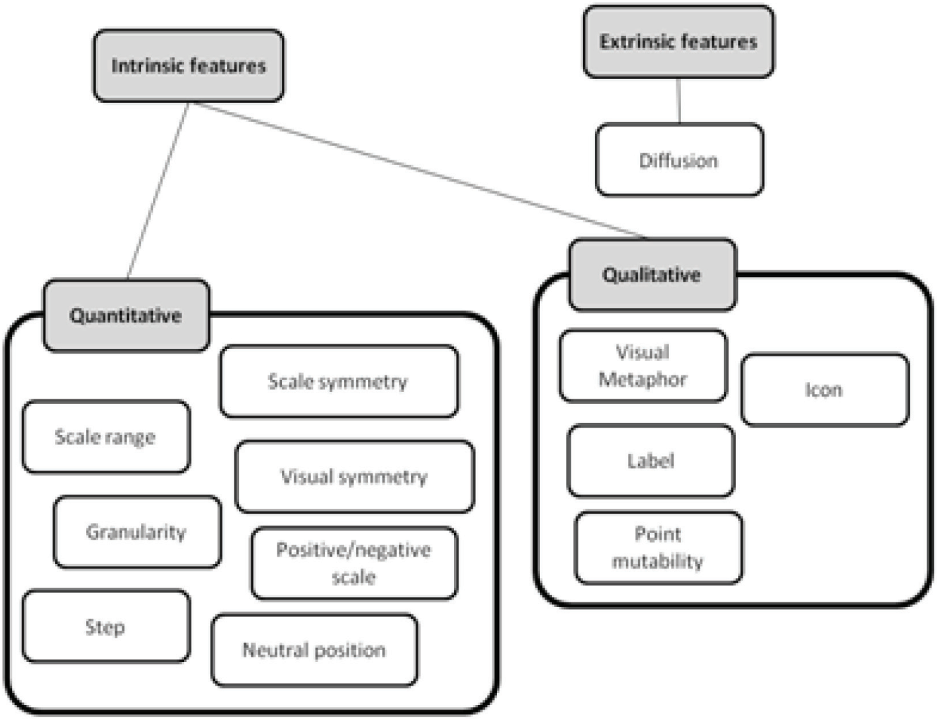
2.2. Identifying Classes of Rating Scales

Aiming at defining general “classes” of rating scales, we organized the thirteen rating scales in a grid-like visualization with features in the rows, and possible values in the cells (Figure 3 b). For example, a row with three cells, corresponding to the values “coarse”, “medium”, and “fine”, was used to represent the “granularity” feature. We then classified each scale (see Figure 3 a) by assigning it to all the relevant cells.

Through visual inspection, we identified three clusters of rating scales:

- **Human Scales** (Smileys, Thumbs): scales with a strong visual characterization, exploiting human visual metaphors to express judgements based on emotions rather than on precise quantifications. They have low granularity, no labels, and medium diffusion; they usually have negative points but no neutral position. They are usually visually symmetric and each point is represented by a different icon. Step, range and symmetry features are not applicable.
- **Neutral Scales** (Stars): Scales with no strong connotation, often considered a standard. Quantitative input corresponds to the number of points users select. They can have any granularity, but no labels, no visual metaphors, and no negative points. They are not symmetric. All their points are represented with the same icon and their step corresponds to 1. They usually have a neutral point and are very widespread.
- **Technical Scales** (Likert Scales, Sliders): scales recalling measurement tools, revealing a strong focus on quantitative evaluations. Differently from neutral scales, they can use abstraction to represent numbers. Their granularity is usually high and their range is quite variable. They have numerical labels, neutral and negative points. They either do not exploit metaphors or use technological ones, with no differences in the visual representation of scale points. Their step can be higher than 1, they are symmentric (both visually and as far as labels are concerned) and have a low spread.

Figure 2. The taxonomy of control features which resulted from card sorting



3. THE USER STUDY: PREPARATION

In order to carry out our study, we needed to select a use-case system and define which rating scales and objects to take into account.

3.1. Use Case Application

As a use case for our experiment, we selected iCITY (Carmagnola et al., 2008), a social adaptive website recommending cultural events, since it allows user to rate various items. In iCITY, users are offered personalized event lists, ordered depending on their preferences and on the context. Moreover, they are suggested to follow other users with similar interests. Users explicitly inform the system of their preferences with respect to event categories (e.g., music, art, etc.) and subcategories (e.g., classical and rock for the music category) using a star-based rating scale.

3.2. Rating Scales

We selected a representative for each of the "classes" of rating scales identified in Section 2.2: *3-point thumbs* for the "human", *5-point stars* for the "neutral" and *11-point sliders* for the "technical" class (see Figure 4). These rating scales were chosen for two reasons.

First, in addition to the metaphor, which depends on their class, all these scales have a different granularity. For stars and sliders, we chose the most common granularities. As for thumbs, 3 points are less common than 2 or 1 points, but we wanted to allow a more fair comparison with the other scales, considered that we aimed at studying rating scales "as a whole" and did not orthogonally vary their main features, so that their effect could be isolated.

Second, users have different degrees of familiarity with them: while stars are very common, thumbs are less popular and sliders are

Figure 3. a) The representation of the chosen rating scales b) The grid-like representation used to define general classes of rating scales

(a)

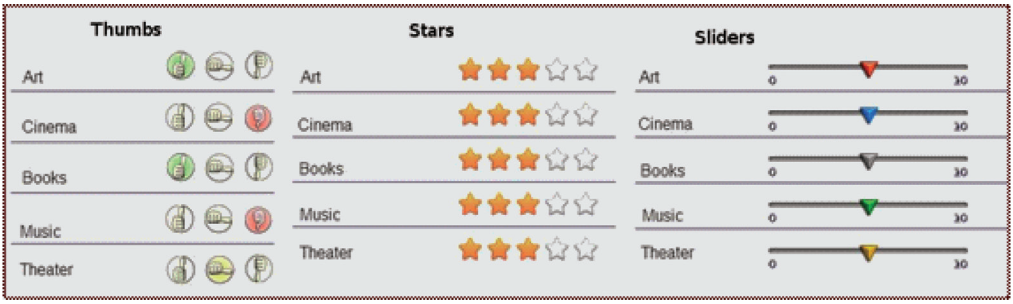
☺	❶	●	❍	■	□	▢
Smileys	Bare numbers	-10/+10 Likert-scale	1/5 Likert-scale	1-point thumbs	2-point thumbs	3-point thumbs

✦	★	✸	⬩	◆	❖
3-point stars	5-point stars	10-point stars	-1/+1 slider	0/10 slider	-10/+10 slider

(b)

Granularity	☺■□▢✦⬩ Coarse	❍★ Medium	❶●✸◆❖ High				
Scale range	☺■□▢ Not applicable	●❖ -10/+10	⬩ -1/+1	❶◆ 0-10	✦ +1/+3	❍★ +1/+5	✸ +1/+10
Step	☺■□▢ Not applicable	❶❍✦★✸⬩◆ 1	●❖ >1				
Label	☺■□▢✦★✸ No	❶●❍⬩◆❖ Numbers	❍ Text				
Icon	❶●❍ No	☺ Smiley	■□▢ Thumbs	✦★✸ Stars	⬩◆❖ Volume controller		
Visual metaphor	❶●❍✦★✸ No	☺■□▢ Human	⬩◆❖ Technological				
Point mutability	❶●❍✦★✸⬩◆❖ No	☺■□▢ Yes					
Neutral position	☺■□✸ No	❶●❍▢✦★⬩◆❖ Yes					
Positive/negative scale	❶❍■✦★✸◆ No	☺●□▢⬩❖ Yes					
Scale symmetry	☺■□▢ Not applicable	❶❍✦★✸◆ No	●⬩❖ Yes				
Visual symmetry	❶ Not applicable	❍■✦★✸◆ No	☺●□▢⬩❖ Yes				
Diffusion	☺❶●❍⬩◆❖ Low	■□▢ Medium	✦★✸ High				

Figure 4. The rating scales used in the experiment



rarely used as rating scales, although they are commonly adopted with other purposes (e.g., varying the zoom level of an image).

3.3. Objects to Evaluate

The *objects* to evaluate in websites (items, categories, users) can be characterized as follows:

- **General vs Specific:** Objects with high generality refer to concepts or categories and are usually less concrete than items which refer to single objects.
- **Simple vs Complex:** Objects can be perceived as unidimensional (e.g., a movie) or multi-faceted (e.g., an experience).

- **Not-Animated vs Animated:** Objects can be things, i.e., concepts or material items, or people.
- **Highly Mutable vs Scarcely Mutable:** Objects can change over time and/or can be evaluated differently according to the context, the affective state of the evaluator and other external factors.
- **Highly Social vs Scarcely Social:** Social objects can be people or contents created by somebody and the expected reaction of the people they are related to can influence their evaluation.

Among the objects users can evaluate in iC-ITY, we selected *event categories* and *suggested users*, since they have opposite features. Event categories are general, simple and scarcely mutable objects, with no social valence, while users are specific, complex and highly mutable objects, with high social valence.

4. THE USER STUDY: PROCEDURE

Our study consisted in two phases: an *exploration phase* and *an experimental evaluation.*

In the *exploration phase,* users familiarized with the three rating scales out of context (i.e., without using them for a specific task). The exploration phase was introduced since we assumed that not all participants were accustomed to using sliders and we wanted to reduce users' initial bias in favour of stars.

In the *experimental evaluation*, participants had to perform two specific tasks:

1. Expressing preferences for event categories;
2. Rating other users suggested by the system.

While the first task is quite common in personalized systems (e.g., in the option setting phase or when users edit their profile), assigning ratings to people is less frequent[2]. In addition, expressing one's own interests with respect to general categories is usually a one-shot task, while rating users is a task which can be carried out frequently, since new recommendations are provided on a constant basis.

4.1. Research Questions

We aimed at answering the following research questions:

- [RQ1] Do user preferences for rating scales depend on the object to evaluate?
- [RQ2] Do user preferences about the most suitable scale for a certain object change after having repeatedly rated it?
- [RQ3] Do user overall preferences for scales change after having used different scales repeatedly?
- [RQ4] What are the motivations for user overall preferences?

4.2. Design

We chose a within-subjects, multiple factors design.

We have two *independent variables*:

- "object" (possible values: *event categories, users*),
- "user experience in evaluating the object" (possible values: *low, high*), i.e., how experienced participants are at evaluating a certain object. We manipulated this variable as follows: in the "low" condition, participants choose their preferred scales before performing any rating task. In the "high" condition, participants make their choices after having performed the rating tasks repeatedly with all the available scales.

We have three *dependent variables*:

- the rating scale chosen by the user,
- the best rating scale,
- the worst rating scale.

We counterbalanced to control for order effects by randomizing the order of presentation of rating scales in the exploration phase. In the experimental evaluation, the "low" condition of the "user experience in evaluating the object" variable inevitably preceded the "high" condition; however, the order of the objects to rate was randomized.

4.3. Participants

We selected 32 participants[3], 56,3% males, 43,8% females, 15-54 years old, among friends and among colleagues and students at the Computer Science Department, University of Turin, according to an availability sampling strategy[4]. All of them were frequent Internet users, very familiar with social media. In accordance with the so-called "90-9-1 principle[5]", however, most of them had a relatively low level of experience with the tasks we took into account: none of them, in fact, declared to frequently rate items or other users, nor to frequently express their interests with respect to general categories in social websites, but most of them (91%) had occasionally rated items, while a few users (15%) had already provided feedback for other users and expressed their preferences with respect to general categories (9%).

4.4. Procedure

Participants first filled in a short *online questionnaire* about their demographics and technology-related habits, where they indicated their gender, age, frequency of Internet and social media usage, number of habitually used social media and frequency of rating behaviours on the Internet (with respect to general categories, to specific, non animated items and to other users).

In the *exploration phase*, participants familiarized with the 3 rating scales out of context, free to either try and interact with every one, or just give them a look-over. Then, they filled in another online questionnaire (see Appendix A) to express their overall preferences for the best and the worst rating scales and their motivations, choosing among options inspired by (Gabrielli and Jameson, 2009).

In the *experimental evaluation*, participants could freely explore iCITY for a few minutes. Then, they were prompted to execute the experimental tasks and encouraged to think aloud.

In the *first condition* (user experience with objects: low), participants had to imagine they performed tasks and had to choose which scale they would use.

In the *second condition* (user experience with objects: high), participants had to actually perform each task with all 3 scales before they chose their favourite one. Each task required that participants rated a set of five categories/users.

Finally, participants expressed again their overall preferences for rating scales using the same questionnaire as in the exploration phase, as a follow-up (see Appendix A).

5. THE USER STUDY: RESULTS

We report the results of the analysis of user choices (RQ1, Section 5.2, and RQ2, Section 5.3), patterns of change (RQ3, Section 5.4) and user motivations (RQ4, Section 5.5). A summary of our research questions, and the answers we could provide to them through our analysis, is reported in Table 6.

5.1. Method

To analyze the effects of objects and experience on user choices, we expoit data from the experimental evaluation. To study user motivations and the evolution of their overall preferences, we compare user answers to our questionnaire in the exploration phase and in the follow-up.

Our analysis mainly focuses on the study of frequency distributions. We use two statistical measures:

- When we consider a univariate frequency distribution, the Gini Heterogeneity index[6] is used to understand whether user choices, for example about the rating scale to use for a certain task, are homogeneous (i.e., users tend to choose the same option) or dishomogeneous (i.e., users tend to choose different options).
- When we consider a bivariate frequency distribution, the chi-squared test is computed in order to prove whether the observed frequencies are statistically different from a theoretical distribution where all cases

are evenly distributed among the possible options, thus allowing to assume a correlation between the two considered variables.

5.2. [RQ1] Do User Preferences for Rating Scales Depend on the Object to Evaluate?

Table 2 summarizes user choices about the rating scale to use for rating different objects with two different levels of experience. Simple visual inspection shows that user choices are very heterogeneous: different users tend to choose different rating scales for a certain object, and there is not a strong consensus about the most appropriate scale to use. This insight is confirmed by the high values of the Gini Heterogeneity index, the lowest value of which is 0,73, still indicating quite heterogeneous user choices. However, considering data at an aggregate level, we can notice that user choices are distributed differently for different objects: most users chose the stars for expressing their interests about event categories, while the thumbs were the most popular scale for rating suggestions about users to follow. Sliders represented the second choice for categories, while they were the least popular one for users. Such differences are apparent in Figure 5 and their significance is confirmed by chi-square analysis (χ(2): 7, 71 α = 0,05 in the "low" condition, χ(2): 18,30 α = 0,001 in the "high" condition), allowing to conclude that users tend to choose different rating scales for different objects. Interestingly, the strength and significance of this correlation get higher in the "high" condition, suggesting that acquiring a certain experience with objects sharpens user perception that rating scales can be used in a specialized way.

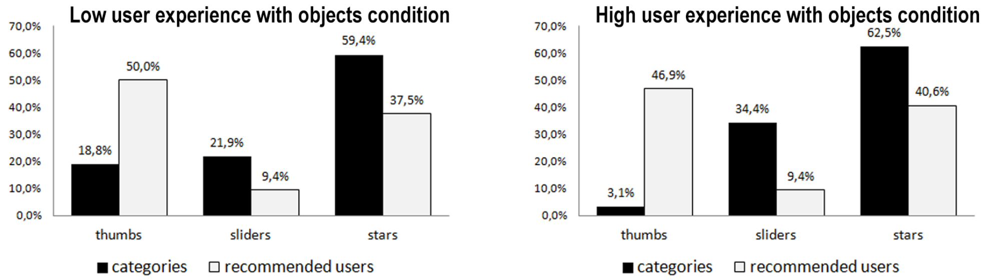
5.3. [RQ2]: Do User Preferences About the Most Suitable Scale for a Certain Object Change, After Having Repeatedly Rated It?

The inspection of Figure 5 also shows that, given a certain object, the distribution of user choices is very similar in the "low" and in the "high" user experience conditions, indicating no significant relationship between user choices and their level of experience, neither for event categories (χ(2):4,486) nor for users to follow (χ(2):0,072). Thus, rating a certain object repeatedly has no effect on user choices. Instead, at an aggregate level, users tend to associate a scale to an object steadily.

Table 2. User choices about the scale to use for rating different objects

	Frequencies			**Gini Heterogeneity Index**
	Thumbs	*Sliders*	*Stars*	
Low user experience				
Task 1 (object: categories)	18,8%	21,9%	59,4%	0,84
Task 2 (object: recommended users)	50%	9,4%	37,5%	0,86
High user experience				
Task 1 (object: categories)	3,1%	34,4%	62,5%	0,73
Task 2 (object: recommended users)	46,9%	9,4%	40,6%	0,87

Figure 5. A comparison of user choices for the "expressing preferences with respect to categories" and the "rating suggestions about users to follow" tasks in the "low" vs "high" "user experience with objects" conditions.



5.4. [RQ3]: Do User Overall Preferences for Scales Change After Having Used Different Scales Repeatedly?

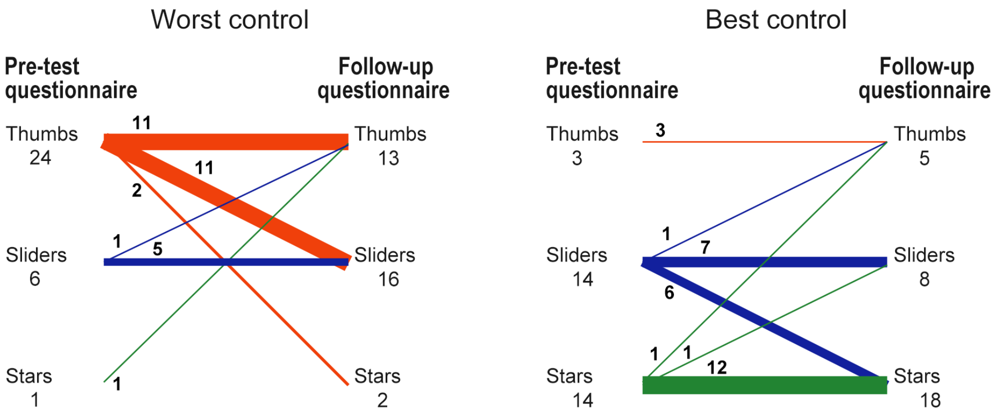
User overall preferences, i.e., general choices that are not associated to a specific object, are different before and after the experimental evaluation. Table 3 shows a clear change in user opinions about the worst scale: before the experiment, most users designated thumbs as the worst scale, while user choices were much more heterogeneous after the experiment and sliders resulted as the worst rating scale for most users. Chi-square analysis confirms that there is no connection in user choices for the worst scale before and after the evaluation. An analogous, although not statistically significant, change occurred for user choices about the best scales, with sliders having lost popularity after the experiment.

Figure 6 summarizes patterns of change in overall preferences. Most users (58,1%) changed their preferences at least once. Among them, 50% changed their opinion only about the worst rating scale, 16,7% changed their opinion only about the best one and 33,3% changed their opinion about both the worst and the best rating scale. Analyzing user preferences in detail, more than half the users who chose thumbs as the worst rating scale before the experimental evaluation then changed their opinion; of them, 45,8% indicated sliders and 8,4% stars as the worst rating scale after the experimental evaluation. On the contrary, most users who chose sliders as the worst rating scale at the beginning maintained their opinion also at the end. As far

Table 3. User choices about the best and worst rating scales in the pre-test and follow-up questionnaires

	Frequencies			Gini Heterogeneity Index
	Thumbs	Sliders	Stars	
Best rating scale				
Pre-test questionnaire	9,4%	43,8%	43,8%	0,87
Follow-up questionnaire	18,8%	25%	56,3%	0,88
Worst rating scale				
Pre-test questionnaire	75%	18,8%	3,1%	0,54
Follow-up questionnaire	40,6%	53,1%	6,3%	0,82

Figure 6. Patterns of change in user overall opinions



as the best rating scale is concerned, most users who chose thumbs and stars at the beginning did not change their preferences, while half the users who chose sliders opted either for stars or for thumbs after the experiment. The most relevant pattern we can derive shows that user opinions about sliders seem to worsen after that they have used this scale repeatedly during the experimental evaluation: in fact, the number of users who indicate them as the worst rating scale increases and, at the same time, fewer users choose them as the best rating scale.

5.5. [RQ4] What are the Motivations for User Overall Preferences?

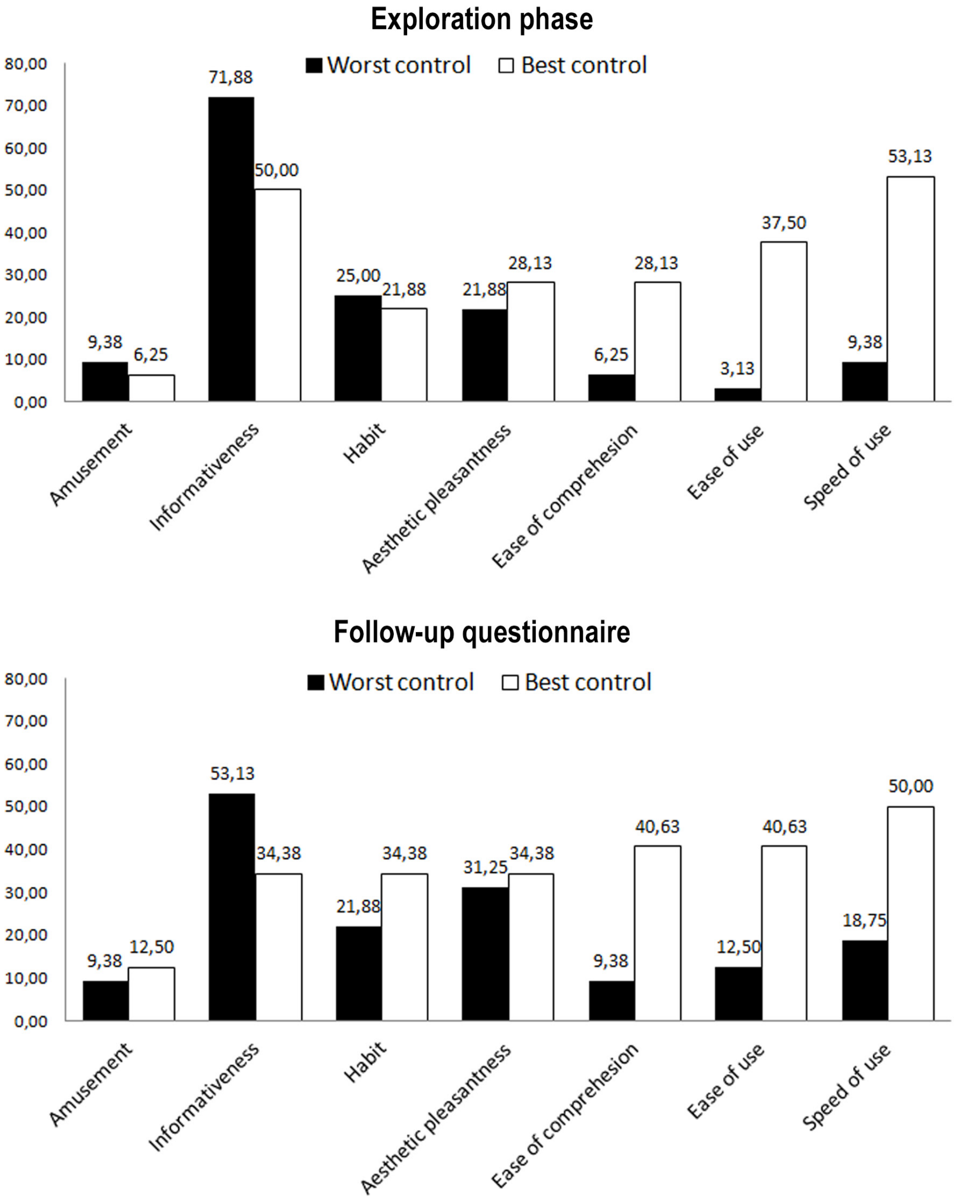
In order to understand why users indicated a certain scale as the worst or as the best one, we considered both their answers in the questionnaire[7] (Section 5.5.1), and their free comments provided through thinking aloud during the experimental evaluation (Section 5.5.2).

5.5.1. Answers in the Questionnaire

Figure 7 summarizes user answers. Motivations for the choice of the best rating scale are more heterogeneous than in the case of the worst one, both in the exploration phase and in the follow-up questionnaire. The main motivations for declaring a scale the worst one are informativeness, habit and aesthetic pleasantness, while the best scales are chosen because of their speed and ease of use, informativeness (in the exploration phase) and ease of comprehension (in the follow-up questionnaire). Thus, user motivations for the two choices only partially overlap, with "informativeness" being the only common main reason in the exploration phase. It seems that qualities the lack of which makes users indicate a certain scale as the worst one are different from qualities the presence of which can cause a scale to be especially appreciated. Finally, it is worth noticing that users provided more motivations after the experiment, possibly as a consequence of their increased experience.

Table 4 and Table 5 report the frequency distribution of user motivations with respect to the three different scales, in the exploration phase and in the follow-up questionnaire, respectively. The data suggest that certain motivations are associated more frequently to specific scales than to the others, in the context of a certain choice. As far as the worst scale is concerned, informativeness is significantly related to thumbs ($\chi(2)$: 12, 938 $\alpha = 0, 01$ in the exploration phase and $\chi(2)$: 20, 071 $\alpha = 0, 001$ in the follow-up questionnaire) and speed of

Figure 7. User motivations for choices



use to sliders (χ(2): 13, 839*[8] α = 0, 001, "exploration"; χ(2): 6,516* α = 0,05, "follow-up").

For the best scale, only the connection between informativeness and sliders is significant in both questionnaires (χ(2): 17, 748 α = 0, 001, "exploration"; χ(2): 13, 345* α = 0, 01, "follow-up"). The connections between stars and speed of use (χ(2): 7, 254 α = 0, 05), ease of comprehension (χ(2): 9, 852* α = 0, 01) and aesthetic pleasantness (χ(2): 6,27* α = 0,

Table 4. User motivations for their overall preferences in the exploration phase

	Worst scale			**Best scale**		
Motivations	**Thumbs**	**Sliders**	**Stars**	**Thumbs**	**Sliders**	**Stars**
Amusement	2	1	0	2	0	0
Informativeness	21	1	1	1	13	2
Habit	8	0	0	1	2	4
Aesthetic pleasantness	5	2	0	1	1	7
Ease of comprehension	1	1	0	0	1	8
Ease of use	1	0	0	1	3	8
Speed of use	0	3	0	2	4	11

Table 5. User motivations for their overall preferences in the follow-up questionnaire

	Worst scale			**Best scale**		
Motivations	**Thumbs**	**Sliders**	**Stars**	**Thumbs**	**Sliders**	**Stars**
Amusement	0	3	0	2	0	2
Informativeness	13	3	1	1	7	3
Habit	2	5	0	1	0	10
Aesthetic pleasantness	4	6	0	3	1	7
Ease of comprehension	1	2	0	1	2	10
Ease of use	0	4	0	3	1	9
Speed of use	0	6	0	4	2	10

05), as well as the connection between thumbs and amusement ($\chi(2)$: 19, 954* $\alpha = 0, 001$) are significant only in the follow-up questionnaire. Finally, stars appear to be significantly related to habit in the exploration phase ($\chi(2)$: 8, 604* $\alpha = 0, 025$). Such relationships highlight distinctive features of the three scales which are especially relevant to justify user choices. On the negative side, thumbs appear little informative, while sliders seem too slow. On the positive side, informativeness is relevant to choose sliders. Stars, which were the most popular rating scale both in the exploration phase and in the follow-up questionnaire, were judged faster to use, easier to understand, more aesthetically pleasant and more familiar with respect to the other scales. Amusement, on the contrary, seems a good reason for preferring thumbs.

5.5.2. Free comments

We distinguish comments according to the evaluated object and users' level of experience in rating.

Object: categories.

Low experience. Most participants' comments made direct or indirect reference to the granularity of the rating scales. For example, some users chose thumbs for the task of evaluating event categories because they are "appropriate for expressing sharp, clear judgments, without graduations". Instead, most users preferred stars since they are very familiar and have "the right granularity." Interestingly, users who chose stars because of their granularity expressed opposite motivations with respect to users who appreciated the raw granularity of thumbs: for example, stars "allow to express intermediate positions of judgment".

High experience. Passing to a condition where users actually have to repeatedly perform tasks, the mere granularity of rating scales loses its importance in favour of other, more specific features. For example, participants say that thumbs "can be easily associated to simple ideas such as 'I like it/I don't like it'", stars "allow to express an order", and sliders "are suitable for comparing items and expressing 'relative' assessments". Participants' comments also become more precise, e.g., "sliders are more convenient if there are lots of categories to evaluate; if there are just a few, stars are preferable". Finally, participants stopped mentioning habit and familiarity with scales, perhaps because more relevant features emerged as a consequence of their direct experience.

Object: users

Low experience. Participant's comments are influenced by the social nature of the object to evaluate - other users. For most participants, thumbs are very suitable for this task since they are ironic ("thumbs are friendly, quizzical, jokey") and not too precise, thus allowing them to put many users at the same level ("All the people I like can be "thumbs up", without unfriendly distinctions"). Also in this case, most participants mentioned granularity when they commented on the scales they had chosen.

High experience. Similarly to what we observed for the task of rating categories, participants tend to cite granularity less often after that they have actually carried out the experimental tasks. Instead, they comment on more specific features, such as the capability of a certain scale to express intermediate judgments, the fact that another scale is more appropriate for a public or a private context of use or the fact that the sliders "allow to express only positive evaluations".

Free comments analysis shows that users find it very important that the scales have the right *granularity* to allow them to express a suitable amount of information. Most participants preferred raw scales for simple objects and finer scales for complex ones, confirming what emerged from questionnaire analysis. "Informativeness"[9], in fact, was among the top motivations for the choice of both the worst and the best rating scale. However, as we noticed, granularity/informativeness partially loses its importance after that users have repeatedly performed rating tasks with different scales. As a further point, the stars were generally very appreciated, probably due to the fact that participants were already accustomed to them. Many participants actually explained their choices in terms of *habit* or *familiarity*. However, this motivations were more important in the case of the worst scale and in the exploration phase, indicating that unfamiliar rating scales can be very penalized if users are not "stimulated" to try them out. Finally, the *social connotations* of the object, in the case of other users, were very often mentioned in comments. Some users cited ethical issues, while others made reference to emotional aspects such as irony and joking. Moreover, some participants commented on the way other users might perceive certain scales (e.g., as "offensive" or "playful"), an aspect which is absent when the evaluated object has no social connotations . Many users also distinguished between private and public contexts of use.

In Table 6, we provide the answers for each research question.

6. THEORETICAL BACKGROUND AND RELATED WORK

In this section, we will provide the theoretical background of our research, which relies on studies about rating scales (Section 6.1) in the context of HCI and survey design and, more in general, about decision making (Section 6.2). In Table 7, we sum up the main comparison points between our work and the most relevant related ones in such fields.

Table 6. Research questions summary

Research Questions	Our Answers
[RQ1] Do user preferences for rating scales depend on the object to evaluate?	Yes Users preferred stars for rating categories and thumbs for rating recommended users.
[RQ2] Do user preferences about the most suitable scale for a certain object change, after having repeatedly rated it?	No Users tend to associate a certain rating scale to a certain object steadily.
[RQ3] Do user overall preferences for scales change after having used different scales repeatedly?	Yes We observed a statistically significant change in users' opinions about the worst rating scale.
[RQ4] What are the motivations for user overall preferences?	Different motivations, in particular: - granularity (it is especially important when users have a low experience with the task of evaluating a certain object) - habit/familiarity (its lack can impress users negatively) - socially related motivations, e.g., privacy, social perception (they are relevant only for objects having social connotations, such as recommended users)

6.1. Rating Scales

In the *survey design* field, many works try to define the features of "good" (i.e., unbiased) rating scales, examining aspects such as the presence of a neutral point (Garland, 1991; Friedman and Amoo, 1999), the use of numeric labels to mark the different points (Amoo & Friedman, 2001), the possible imbalance in the number of positive and negative points (Friedman and Amoo, 1999) or the granularity (Colman et al., 1997; Dawes, 2008). Differently from these studies, we do not focus on finding the best rating scales but on discovering which ones users would choose and why.

In the HCI field, rating scales are studied to find which features can promote usability. According to Herlocker et al. (2004), appropriate rating scales should allow users to express exactly as many levels of liking as they wish to distinguish. The authors thus implicitly suggest that rating scales with different granularities might be appropriate for different objects, as we demonstrate in our work. Cosley et al. (2003) studied granularity, numeric labels and visual metaphors in a series of user studies, concluding that designers can allow users to choose the scales they prefer, since they found no significant difficulties in managing multiple scales. Nobarany et al. (2012) focused on the measurement scale, distinguishing between "absolute rating" and "relative ranking", and on the recall support, i.e., the ability to offer information about previously provided input. They concluded that users take advantage of recall support, while ranking interfaces may be quite challenging since they force users to specify a total ordering for items.

Differently from these studies, all focusing on specific features, we aim at investigating the process of selecting a rating scale as a human decision making problem. We think that this choice can be influenced not only by scale features but also by other factors, such as user experience in the task, the evaluated object or the social context. To this respect, our work is similar to the one by van Barneveld and van Setten (2004), who investigated user preferences for rating scales comparing two different situations: *receiving output from* vs *giving input to a system*. The authors found that most users prefer to have system output presented by means of five-star rating scales, but they are less in agreement regarding input, consistently with our findings. In our work, we extend their results studying user preferences for rating scales with

Table 7. Comparison with the most relevant related work

Work	Common topic	Results	Differences with our study
Herlocker et al., 2004	Rating scale features: granularity	Rating scales with different granularities are appropriate for different objects	Rating scales with different granularities and visual metaphors are appropriate for different objects. We also studied whether user choices change according to the level of experience.
Nobarany et al., 2012	Rating scale features: measurement scale, recall support	Users take adavantage of recall support, while ranking interfaces are challenging	We only consider rating scales in common use, which do not normally offer recall support and are all characterized by the same measurement scale (absolute). The scales we consider have different granularities and visual metaphors.
Gena at al., 2011	Rating scale personality	Each scale has its own "personality" which influences the ratings of the users	We did not consider the influence of the scale on user ratings, but what influences users' preferences for a particular scales
Van Barneveld and van Setten 2007	User preferences for scales in input and output situation	Users prefer five stars for output, while there is no consensus for input.	User preferences in input situations are very heterogeneous. While we did not consider output situations, we took into account different objects to evaluate and the level of experience. We found that five stars are preferred for event categories, but not for recommended users.
Gabrielli and Jameson, 2009	Factors which influence user choices: level of experience	Users' level of experience can determine changes in users' choices	Users' level of experience influences overall preferences for rating scales, but has no effect on choices related to a specific task.
Jameson 2014	Choices in HCI	Users are considered as "choosers". There are six main patterns of choice, based on social influence, experience, consequences, attributes, policies, trial and error	We focused on user preferences for rating scales as the result of a choice process. Beside the level of experience, we studied in particular the features of the objects to evaluate, corresponding to attributes in Jameson's model.

respect to different objects to evaluate and different levels of experience with them.

Finally, Gena et al. (Cena et al., 2010; Gena et al., 2011) adopted a different perspective, studying the influence of rating scales on user ratings. They introduced the concept of "rating scale personality" (resulting from user perception of a scale), and showed that it can influence ratings. The authors concentrated on the effects of rating scales on user rating behavior, while in this work we want to understand what influences user preferences for rating scales themselves.

6.2. Decision Making

This research field has attracted much attention in psychology, from the classical work of Tversky & Kahneman (2008) to more recent approaches such as those of Ariely (2008) and Cialdini (2007). According to Jameson (2012), partially related literature in the field of HCI can be found in the areas of *usability guidelines,* which are meant to help users make the right choices when using an interactive system, *recommender systems*, which support users' preferential choices about items to consume, e.g., products to buy (Adomavicius and Tuzhilin, 2005), and persuasive technologies, i.e., interactive technologies which aim at influencing users' decisions, in terms of preferences and behaviours, in a veriety of domains (Fogg, 2003; Oinas-Kukkonen & Harjumaa, 2008). Jameson introduced the idea of "user as chooser" (Jameson et al., 2011) with the aim of defining choices and decisions with respect to computer systems in a practical way. In this line of research, Gabrielli and Jameson (2009) outlined several factors which can have an effect on user choices: tasks to perform, skills, usage context, usability priorities, aesthetic preferences, personality traits, habits formed with other systems, desire for novelty, system properties. In Jameson et al. (2014), the author adopted a more general point of view and highligthted the following broad aspects as a basis for choice patterns: social influence, experience, consequences, attributes, policies, trial and error. In this line, with our user study, we focus on user preferences for rating scales as the result of a decision making process, and try to understand which factors can have an influence on it.

An interesting aspect in the study of choices regards their evolution, e.g. how they change over time or as a consequence of other factors. There are only few longitudinal studies which allow to observe such changes (Gabrielli and Jameson, 2009; Jameson et al., 2009). In this paper, we investigated how choices can change depending on user experience with the evaluated objects, a factor connected to time, without being completely equivalent to it. Gabrielli and Jameson (2009) identified an increase in the level of experience as the main reason for changes in factors (such as user skills or habits) which can in turn influence choices.

In Table 7, we provide a conclusive comparison with the most important related work described above, in order to highlight the differences in results and methods with our study.

7. CONCLUSION

Differently from most previous related research, we did not study how specific features can bias or improve the ability of a rating scale to collect actual user preferences, but focused on the way users choose rating scales when they are confronted with a certain number of options and have to perform a realistic task. Among the most important findings of our experiment is the fact that user choices depend on the object they have to rate. The level of user experience in rating objects influences user overall preferences (as well as the motivations users provide for them), but it is not so relevant in determining user choices, given a specific object. Our focus on the point of view of users represents a new perspective to the study of rating scales, and one of our main strengths. In this way, we were able to bring in the discussion contextual factors which were normally overlooked in previous work, such as the evaluated object and user experience with it.

Implications of the results. According to our findings, we can provide the following insights related to user choices and motivations:

- Users tend to disagree about the rating scale to use for a certain object. However, stars are always quite popular, possibly due to their familiarity.

- Users choose different rating scales for objects with different features and maintain their choices, even when their level of experience with objects and knowledge of possible alternatives increases.
- Users tend to change their overall preferences when their level of experience in rating objects increases.
- Informativeness (granularity) is very important for determining both negative and positive choices; however, it is more important when user experience with the evaluated objects is low.
- Habit and familiarity are especially relevant in determining negative preferences: unfamiliar scales might be very penalized if users are not stimulated to try them out.
- The most relevant motivations for negative preferences are (the lack of) informativeness, habit, and speed of use.
- The most relevant motivations for positive preferences are informativeness, speed and ease use, and ease of comprehension.

Regarding the rating scales we can conclude that:

- Sliders are appreciated for their fine granularity and informativeness; on the negative side, they are the most time consuming rating scale of the three. Negative features prevail over positive features after that users have tried them out repeatedly.
- Stars are very familiar to users and therefore can be used as a default solution for most tasks. They are especially useful when users want to express precise evaluations and if the evaluated objects have low social valence.
- Thumbs are suitable for expressing simple evaluations, when users want to be ironic and when the evaluated objects have high social valence.

Guidelines for designers of interactive systems. Our results would be of particular interest for designers of interactive systems, for choosing the best rating scale to use in their systems

- Designers should offer stars (or, possibly, another neutral scale) as a default option for expressing interests with respect to general categories.
- Designers should offer thumbs (or, possibly, another human scale) as a default option for rating users.
- In case of a new task (i.e., other than expressing interests for categories or rating users), stars (or, possibly, another neutral rating scale) are usually a good default option.
- Designers might allow users to choose the rating scales they prefer according to the objects they have to evaluate and allow them to change their choices later.
- When they are assessing a novel interface/system, designers should repeat their evaluations after that participants have acquired a certain level of experience, in order to verify whether their initial findings still hold.
- When they are assessing an existing interface/system, designers should pay attention to involve both novice and experienced users, since their opinions might be different.

Limitation of our study. The main limitation of our study regards the number of selected rating scales. In fact, if we can certainly conjecture that all scales in a certain class will be perceived and used in a similar way, we need to carry out a study where we compare more representatives from the same class to be able to actually draw such generalizations. Similarly, we did not try to isolate the effect of specific rating scale and object features on user choices, an aspect which, if properly addressed, might allow us to make predictions about user choices with respect to new combinations of scales and objects to evaluate, without having to empirically test them. Finally, the usage of scales in time is limited to a simulation through repetitive usage; however, this stratagem might not allow us to capture all relevant changes in user choices which are connected to time passing.

Possible future directions of our work.

Free comments analysis showed that there are some aspects we have not yet considered which might influence user choices and which should therefore be taken into account in future work, such as the *level of privacy* assigned to ratings: many comments pointed out that different scales would be appropriate in a private and others in a public context, especially in case of items with a high social valence. Moreover, we want to overcome some of the limitations of our current study. In particular, we want to investigate the role of time (rather than the level of user experience with the evaluated object) in preferential choices, considering a wider time interval. We also intend to consider other objects, in order to study user choices with respect to different combinations of object features, and other scales.

Finally, we are planning to interpret our results in the light of general frameworks on user behaviour and choices, such as Jameson's (2014).

ACKNOWLEDGMENT

We would like to thank Anthony Jameson, Cristina Gena and Silvia Gabrielli for their help in the study.

REFERENCES

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transaction on Knowedge Data Engeneering*, *17*(6), 734–749. doi:10.1109/TKDE.2005.99

Amoo, T., & Friedman, H. H. (2001). Do Numeric Values Influence Subjects Responses to Rating Scales*? Journal of International Marketing and Marketing Research*, *26*, 41–46.

Ariely, D. (2008). *Predictably Irrational: The Hidden Forces That Shape Our Decisions*. New York: Harper Perennial.

Carmagnola, F., Cena, F., Console, L., Cortassa, O., Gena, C., & Goy, A. et al. (2008). Tag-Based User Modeling for Social Multi-Device Adaptive Guides. *User Modeling and User-Adapted Interaction*, *18*(5), 497–538. doi:10.1007/s11257-008-9052-2

Cena, F., Vernero, F., & Gena, C. (2010). Towards a customization of rating scales in adaptive systems. In P. De Bra, A. Kobsa, D. N. Chin (Eds.), User Modeling, Adaptation, and Personalization, 18th International Conference, UMAP 2010, Big Island, HI, USA, June 20-24, 2010 (pp. 369–374). Springer. doi:10.1007/978-3-642-13470-8_34

Cialdini, R. B. (2007). *Influence: The Psychology of Persuasion* (Revised edition). New York: Collins.

Colman, A., Norris, C., & Preston, C. (1997). Comparing rating scales of different lengths: Equivalence of scores from 5-point and 7-point scales. *Psychological Reports*, *80*(2), 355–362. doi:10.2466/pr0.1997.80.2.355

Cosley, D., Lam, S. K., Albert, I., Konstan, J. A., & Riedl, J. (2003). Is seeing believing?: how recommender system interfaces affect users' opinions. In *SIGCHI Conference on Human Factors in Computing Systems* (585-592). ACM, New York, NY, USA.

Dawes, J. (2008). Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *Internation Journal Market Research* (50), 61–77.

Fogg, B. J. (2003). *Persuasive technology: Using computers to change what we think and do*. San Francisco: Morgan Kaufmann.

Friedman, H. H., & Amoo, T. (1999). Rating the rating scales. *Journal of Marketing Management*, *9*, 114–123.

Gabrielli, S., & Jameson, A. (2009). Difference and changes in preferences regarding personalisation systems: a user-center design perspective. Paper presented at the Sixth Workshop on User-Centred Design and Evaluation of Adaptive Systems in conjunction with UMAP2009, Trento, Italy.

Garland, R. (1991). The Mid-Point on a Rating Scale: Is it Desirable? *Marketing Bulletin* (2), 66–70

Gena, C., Brogi, R., Cena, F., & Vernero, F. (2011). The impact of rating scales on user's rating behavior. In J. A.

Herlocker, J., Konstan, J., Terveen, L., & Riedl, J. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, *22*(1), 5–53. doi:10.1145/963770.963772

Hornbæk, K. (2011). Some Whys and Hows of Experiments in Human–Computer Interaction. *Foundations and Trends in Human–Computer Interaction*, (5: 4), 299–373.

Jameson, A. (2012). Choices and decisions of computer users. Jacko, J.A. (Ed.), The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications. CRC Press, Boca Raton, FL, 3rd edition, 77–91.

Jameson, A., Berendt, B., Gabrielli, S., Gena, C., Cena, F., Vernero, F., & Reinecke, K. (2014). Choice Architecture for Human-Computer Interaction. *Foundations and Trends in Human-Computer Interaction*, *7*(1–2), 1–235.

Jameson, A., Gabrielli, S., Kristensson, P. O., Reinecke, K., Cena, F., Gena, C., & Vernero, F. (2011). How can we support users' preferential choice? In *International Conference on Human Factors in Computing Systems, Extended Abstracts Volume* (409–418), ACM, New York, NY, USA.

Jameson, A., Gabrielli, S., & Oulasvirta, A. (2009). Users' preferences regarding intelligent user interfaces: differences among users and changes over time. In *International Conference on Intelligent User Interfaces* (497–498). ACM, New York, NY, USA.

Konstan, R. Conejo, J. Marzo, N. Oliver (Eds.), User Modeling, Adaption and Personalization - 19th International Conference, UMAP 2011, Girona, Spain, July 11-15, 2011. Proceedings (pp. 123–134). Springer.

Nobarany, S., Oram, L., Rajendran, V. K., Chen, C. H., McGrenere, J., & Munzner, T. (2012). The design space of opinion measurement interfaces: exploring recall support for rating and ranking. In *SIGCHI Conference on Human Factors in Computing Systems* (2035-2044). ACM, New York, NY, USA. doi:10.1145/2207676.2208351

Oinas-Kukkonen, & H., Harjumaa, M. (2008). A systematic framework for designing and evaluating persuasive systems. In international conference on Persuasive Technology (164–176), Berlin, Heidelberg.

Tversky, A., & Kahneman, D. (2008). The framing of decisions and the psychology of choice. *Science* (211), 453–458. In SIGCHI Conference on Human Factors in Computing Systems (2035-2044). ACM, New York, NY, USA.

van Barneveld, J., & van Setten, M. (2004). Designing Usable Interfaces for TV Recommender Systems. In L. Ardissono, A. Kobsa, & M. Maybury (Eds.), *Personalized Digital Television. Targeting programs to individual users*. Kluwer Academic Publishers. doi:10.1007/1-4020-2164-X_10

ENDNOTES

[1] When there is some overlapping between the features (visual metaphor and presentation form, granularity and precision) described in van Barneveld and van Setten (2004) and Gena et al. (2011), we follow Gena's terminology.

[2] In online auction and shopping websites such as Ebay, rating other users based on their behaviour as buyers or sellers allows to fuel trust and reputation dynamics. Similarly, in the context of forums and online communities, users are often evaluated according to their expertise, to the quality of their contributions, to their conduct as merchants (see Maxima.org) or as video-game players (see http://it.wikipedia.org/wiki/Animal_Crossing).

[3] This number of participants is in line with customs and pragmatic recommendations in the area of HCI studies, which typically employ 20 participants per condition (Hornbæk, (2011)).

[4] Much research in social science is based on samples obtained through non-random selection, such as the availability sampling, i.e. a sampling of convenience, based on subjects available to the researcher.

[5] This principle expresses participation inequality on the Web, stating that contents are created by about 1% of people, are edited by about 9% of them and viewed (without contributing) by the remaining 90%. See for example: http://en.wikipedia.org/wiki/1%25_rule_%28Internet_culture%29#cite_note-participation_inequality.

[6] Gini Heterogeneity Index is a measure of dispersion for categorical variables. The normalized version of this index, to which we refer here, ranges from 0 (maximum homogeneity) to 1 (maximum heterogeneity).

[7] See the Appendix for the questionnaire

[8] Chi-square values are marked with a wildcart in case more than 20% of the expected frequency values are lower than 5.

[9] Although "granularity" *per se* did not appear in the list of possible motivations in the questionnaires, we included "informativeness", which is a highly related, although slightly more general, concept.

Federica Cena is an Assistant Professor at the Department of Computer Science of the University of Turin (Italy). She obtained her Ph.D at the University of Turin with a thesis on user model interoperability process, in March 2007. Her main research interests are related to semantic knowledge representation and semantic reasoning for user modeling and adaptation, user model interoperability. In the last years, she is mainly devoted in studying the implications of semantic knowledge representation for Social Web and for Internet of Things.

Fabiana Vernero is a short-time researcher at the Department of Computer Science of the University of Turin (Italy), and a co-founder at Sinbit s.r.l., a university spin-off which mainly deals with mobile applications. She obtained her Ph.D at the University of Turin with a thesis on double-sided recommendations, in January 2011. Her main research interests lie in the areas of Human-Computer Interaction, Interaction Design and Recomemnder Systems.

APPENDIX A

Questionnaire Used in the Exploration Phase and in the Experiment Follow Up

Please, answer the following questions.

1. Which control do you like the most?
    - ◦ Thumbs
    - ◦ Sliders
    - ◦ Stars

    1.2. Which are the factors (max 3) which have influenced your choice the most?
    - ◦ It's amusing
    - ◦ I'm used to use it
    - ◦ It's cute
    - ◦ It's very informative
    - ◦ It's easy to understand
    - ◦ It's easy to use
    - ◦ It's fast to use
    - ◦ Other (specify)

    ……………………………………………………………………………………………

2. Which control do you like the least?
    - ◦ Thumbs
    - ◦ Sliders
    - ◦ Stars

    2.1. Which are the factors (max 3) which have influence your choice the most?
    - ◦ It's not amusing
    - ◦ I'm not used to use it
    - ◦ It's not cute
    - ◦ It's not very informative
    - ◦ It's not easy to understand
    - ◦ It's not easy to use
    - ◦ It's not fast to use
    - ◦ Other (specify)

    …………………………………………………………………………………….