**How scales influence user rating behaviour in recommender systems**

# How scales influence user rating behaviour in recommender systems

Federica Cena[a], Cristina Gena[a], Pierluigi Grillo[a], Tsvi Kuflik[b], Fabiana Vernero[a]* and Alan J. Wecker[b]

[a] *Department of Computer Science, University of Torino, Torino, Italy;*

[b] *Department of Information Systems, The University of Haifa, Haifa, Israel.*

Word count: 13504

* *Corresponding Author:*

*Fabiana Vernero,*

*Corso Svizzera 185, 10140, Torino*

*vernerof@di.unito.it*

# How scales influence user rating behavior in recommender systems


## Abstract

Many websites allow users to rate items and share their ratings with others, for social or personalization purposes. In recommender systems in particular, personalized suggestions are generated by predicting ratings for items that users are unaware of, based on the ratings users provided for other items. Explicit user ratings are collected by means of graphical widgets referred to as "rating scales". Each system or website normally uses a specific rating scale, in many cases differing from scales used by other systems in their granularity, visual metaphor, numbering or availability of a neutral position. While many works in the field of survey design reported on the effects of rating scales on user ratings, these, however, are normally regarded as neutral tools when it comes to recommender systems. In this paper, we challenge this view and provide new empirical information about the impact of rating scales on user ratings, presenting the results of three new studies carried out in different domains. Based on these results, we demonstrate that a static mathematical mapping is not the best method to compare ratings coming from scales with different features, and suggest when it is possible to use linear functions instead.



**Keywords**: rating scales, recommender system, user studies, human-machine interface

## 1. INTRODUCTION

Following the advent of Web 2.0, many web-based applications (including social media and e-commerce websites) give users the opportunity to rate content, for social or for personalization purposes. For example, YouTube[1] allows users to rate videos and share their ratings with others. Users provide their ratings by means of "rating scales", i.e. graphical widgets that are characterized by specific features (e.g. granularity, numbering, presence of a neutral position, etc.) (Gena et al.2011). On the Web, we can find various examples of rating scales, differing, among other things, in their visual appearance, which is probably their most salient feature, for example: Amazon[2], Anobii[3] and Barnes & Noble[4] use stars, Facebook[5] and YouTube use thumbs, Tripadvisor[6] uses circles, LateRooms[7] uses squares and Criticker[8] uses bare numbers.

User ratings are especially valuable pieces of information for most recommender systems (Adomavicius and Tuzhilin 2005), where users express their preferences for items by rating them and, based on these ratings, they receive personalized suggestions, hopefully suited to their needs. It is, therefore, important to be able to correctly interpret and compare ratings even when they were expressed using different rating scales, as described in the following scenarios. For example, an application A (e.g., IMDb) may be able to import users' ratings from another system B (e.g., YouTube) in the same domain in order to improve its user models with more information about users. Since the two systems use two different rating scales (e.g., "stars" and "thumbs" with different granularities), system A needs a mechanism to correctly translate ratings from scale B to its scale, e.g. from a 2-point scale of thumbs to a 10-point scale of stars. In the same way, a new web-based application, that is not able to provide recommendations due to the cold start problem, can import user ratings from other systems which the user is served by (each with its own rating

[1] http://www.youtube.com
[2] http://www.amazon.com
[3] http://www.anobii.com
[4] http://www.barnesandnoble.com
[5] http://www.facebook.com
[6] http://www.tripadvisor.com
[7] http://www.laterooms.com
[8] http://www.criticker.com

scale), and use them to generate recommendations. Thus, it needs a method for converting these heterogeneous ratings data to the single, homogeneous scale it uses. Similarly, in a Web 2.0 scenario, social aggregators or mashup systems, like Trivago in the tourism domain, may aggregate different services with different features and provide an aggregation (or a summary) of the information from all the systems involved. Thus, they need to correctly aggregate users' ratings expressed with different rating scales. Moreover, application developers may want to change the rating scale in order to better satisfy users' needs, as YouTube did in 2009, moving from 5-point (stars) to 2-point (thumbs). In this case, the application needed to keep all previous ratings collected in the past, and to be able to properly convert them to the new scale. Finally, application developers may want to give users an opportunity to use different rating scales for different tasks (for example, stars to configure their interests in the user model and thumbs to rate items, as can be seen in certain websites, like booking.com[9] for instance) or in certain moments (for example, when users gain experience with the system, they may switch from one scale to another). Furthermore, application developers might give their users the possibility to rate items according to different parameters, each of which can be assessed on a specific rating scale. For example, an application in the tourism domain might allow users to rate hotels according to their cleanness and location, providing a 5-star rating scale for cleanness and a thumbs-up/thumbs-down one for location. Thus, in principle a system could allow users to choose which rating scale to use (Cena et al. 2010). Finally, in the context of in the so-called *cross-domain recommendation*, Cantador et al. (2015) propose to leverage all the available user data provided in various systems and domains in order to generate a more complete user model and better recommendations. As a practical application of knowledge transfer in the context of a single domain (e.g. movies), a target system may import user ratings of overlapping items from a source system - for example representing user rating with a given rating scale - as auxiliary data to a

[9] http://www.booking.com

target system - representing user rating with another rating scale with different granularity - in order to address the data-sparsity problem.

In all these cases, these systems need a way to compare ratings given in different rating scales. However, this is not a trivial task, since the rating scales themselves have an effect on the ratings, as many studies point out (Garland 1991, Friedman and Amoo 1999, Amoo and Friedman 2001). In fact, how people respond to different rating scales is primarily an issue of psychology rather than a mathematical question (Cummins and Gullone 2000). This paper particularly focuses on the influence the rating scales exert on user rating behavior, i.e., their capacity to induce users to assign higher or lower ratings than they would assign with a different rating scale. Some interesting insights about the effect of rating scales features on the user rating behavior in recommender systems can be found in the work of Vaz et al. (2013). Their results suggest that using a rating scale with a smaller granularity obtains better results, in terms of a lower MAE, in a rating prediction task. While the work of these authors deals with the impact of rating scales on user rating behavior in a broad sense, we are particularly focused on their influence on user ratings. While there does not seem to be much debate around this issue in the field of recommender systems and intelligent user interfaces, a notable exception is represented by (Cosley et al. 2003). The authors found out that ratings on different scales correlated well, and suggested that designers might allow users to choose their favourite rating scale and compute recommendations by means of mathematically normalized scores.

Conversely, in a similar experiment done by Cena et al. (2010), where users were asked to rate the same item on different rating scales, it was observed that 40% of the ratings departed considerably from mathematical proportion, suggesting that rating scales themselves might induce users to be more or less optimistic (or strict, or meticulous). This insight has been also confirmed in two further experiments done by Gena et al. (2011). In our vision, this effect is, at least in part, a consequence of the set of features (e.g., granularity, numbering, presence of a neutral position, etc) that characterize and differentiate a certain scale from the others. For example, a rating scale

exploiting a human metaphor and offering only one point (like the simple "thumbs-up") is perceived very differently in comparison with a scale consisting in bare numbers ranging from 0 to 100. Consequently, the ratings given by means of the two scales can be expected to be considerably different as well.

The main contribution of this paper is twofold: first, we provide new empirical evidence about the effects of rating scales on user ratings focusing on recommender systems with studies on large samples and in different domains and contexts, both on and off the web; second, we demonstrate that a static mapping is not the optimal solution to compare ratings originating from rating scales with different features.

By static mapping we refer to a standard mapping which does not take into account user behaviour, but only the number of points offered by a certain scale, and maps them to a chosen interval based on mathematical proportion. For example, in case the destination interval is 0-1 (i.e., the lowest value should be mapped to 0 and the highest to 1), the new rating can be obtained by applying the following formula:

$$\frac{1}{|points| - 1} * (point\_position - 1)$$

where |points| indicates the number of points of the original scale and point_position indicates the position of the point corresponding to the original rating. For example, for mapping 3 out of 5 in a 5-point rating scale to a destination interval of 0-1, |points| corresponds to 5, point_position corresponds to 3, and the obtained value is 0.5.

The paper is structured as follows. Section 2 provides the background of previous work on rating scales, while Section 3 provides an analysis of rating scales in order to point out their features and a general description of the three empirical evaluations; Section 4, Section 5, and Section 6 describe the different experiments in three different contexts: several movie websites, a museum guide, and a controlled web-based experiment. Afterwards, Sections 7 and 8 discuss and conclude the paper with considerations and future directions for work.

## 2. RELATED WORK

### 2.1 Specific features and their impact on user ratings

The effect of rating scales on user ratings is often reported in research works in different fields, such as economics, psychology and surveys design. These works generally focused on the study of how specific features, such as granularity, neutral point or labelling, can affect the final rating.

**Granularity**: The optimal number of points in rating scales (*granularity*) and its possible effect on user responses is a much-debated issue in the literature, even if the results obtained so far do not appear to be conclusive. Lim (2008) had 137 participants that repeatedly assessed their overall level of happiness with 4, 5, 7 and 11-point Likert-like rating scales, with the aim of investigating whether the number of points can affect the respondents' ratings. The author directly rescaled all ratings to the 11-point scale for comparison purposes and found that the mean happiness value was significantly higher for the 11-point scale with respect to the 4 and 7-point scales, while there were no significant differences between participants' mean ratings for the 11-point and the 5-point scales. This implicitly implies that higher granularity causes higher ratings. Dawes (2002), in a study that is related to our experiment 3 (Section 6.2), found that data collected with five-point scales could be easily translated to make them comparable with data collected with eleven-point scales. Similarly, in a subsequent study, they found that five and seven-point scales can easily be re-scaled; comparing five or seven-point data to 10-point data, a straightforward re-scaling and arithmetic adjustment easily facilitated the comparison (Dawes 2008).

Other works studied the effect of granularity on the *reliability* of the response. Preston and Colman (2000) examined Likert-like scales with 2 to 11 points, as well as a 101-point scale and found that ratings expressed by means of scales with 2, 3, or 4 points were the least reliable, valid, and discriminating; in addition, participants at their study preferred scales with a relatively high number of points (i.e., 7, 9 or 10), even if scales with 2, 3 or 4 points were judged quicker to use. Thus, they concluded that high granularity is more reliable and preferred by users.

Similar results were obtained by Weng (2004), who investigated how granularity can impact the test-retest reliability and found that scales with more response options (at least three) have a better chance of attaining higher reliability. More recently, Shaftel et al. (2012) compared different rating scales (a 5-point and a 7-point standard Likert rating scales, a 9-point numerical scale with verbal labels only at the endpoints, a 0-100 visual analogical scale marked in tens (i.e., 0, 10, 20, etc.) and a 6-point verbally labelled scale) and found that some items were ineffective at collecting distinguishing information with some rating scales but effective with others, concluding that the best granularity depends on the item content and purpose, and thus the decision is domain-dependent.

**Neutral point:** In Garland (1991), the author produces some evidence that the presence or absence of a neutral point on a scale produces some distortion in the results. In particular, they found that some respondents may choose the mid-point in order to provide a less negative answer, because of a social desirability bias. On the other hand, rating scales with no mid-point force the real indifferent to make a choice, causing a distortion towards higher or lower answers, depending on the content which is being assessed. Weijters et al. (2010) also investigated whether the presence or absence of a neutral midpoint can affect user responses. In particular, they found that it causes a higher NARS ("net acquiescence response style", i.e., the tendency to show more agreement than disagreement), a lower ERS ("extreme response style") and a lower MR ("misresponse to reversed items"). In relation to our goals, we can say that with neutral points in the rating scale, we will have less extreme responses and higher ratings.

**Labelling:** Various researchers observed that the *labelling* of a scale influences user responses. Weijters et al. (2010) took into account the format of labels, comparing a case where all points were labelled and one where only the endpoints were labelled. They observed that labelling all the points (w.r.t. to only the extremes) causes a higher NARS, a lower ERS and a lower MR. They concluded that the format of a rating scale can bias the mean, variance and internal consistency of the collected data, so that pieces of information obtained with different rating scales are not

directly comparable. Other authors studied the effect of the polarity of the labels, providing evidence of a bias towards the left side of a scale, possibly due to factors such as reading habits, a primacy effect or pseudoneglect, i.e., an asymmetry in spatial attention which favours the left side of space (Holzinger et al., 2011). Friedman et al. (1993), for instance, collected students' attitudes towards their college with two different scales, one where the items were labelled "strongly agree", "agree", "undecided", "disagree" and "strongly disagree" and one where the labels appeared in the opposite order, and found that the first scale resulted in a significantly greater degree of agreement. Similarly, Yan and Keusch (2015) varied the direction of an 11-point rating scale with numeric labels (from 0 to 10 and vice-versa) and found that the mean ratings were shifted toward the left end of the scale. Moreover, in (Amoo and Friedman, 2001), the authors show that the negative evaluation side of a scale is perceived as more negative when it is labelled with negative rather than positive numbers (e.g., -4 rather than 1), and this causes higher average ratings when scales with negative numerical labels are used. The same result was later achieved by Tourangeau et al. (2007). Interestingly, the authors also observed an analogous (even if less extreme) effect with colours: in fact, they found that user ratings tend to be higher when the end points of a scale are shaded in different hues, as compared to scales where both ends are shaded in the same hue[10].

### 2.2 Using different rating scales in recommender systems

Differently from the above-mentioned works we focus on rating scales as a whole, trying to highlight the influence of scales on user ratings in recommender systems. As seen in Section 1, this topic is particularly relevant, since the recommender's performance depends on ratings. However, apart from Vaz et al. (2013) and (Cosley et al. 2003), which studied the effect of different scales on user ratings also in relation to MAE, most of the studies in this area focused on

[10] A detailed analysis of the impact of color on user ratings is out of the scope of the current paper. The interested reader can refer to (Tourangeau et al. 2007) for more information on their study or to (Stickel et al., 2009) for a more general discussion of how colours can enhance usability and user experience.

other aspects, such as the design of rating scales and the effects of user's personality on rating behavior.

Vaz et al. (2013) carried out an experiment where they compared the performance of a collaborative filtering algorithm using ratings expressed on two scales with different granularity. More specifically, the authors obtained two different data sets by mapping their original set of ratings, which were expressed on a 5-point scale, to a 3-point scale (dislike/neutral/like), converting "1" and "2" ratings in "dislike", "3" ratings in "neutral", and "4" and "5" ratings in "like". Their results suggest that using a rating scale with a smaller granularity obtains better results, in terms of a lower MAE, in a rating prediction task. The intuition behind such a result is that general preferences, expressed at the level of like and dislike, are more stable than preferences expressed with a high degree of detail. While this work deals with the impact of rating scales on user rating behaviour in a broad sense, we are particularly focused on their influence on user ratings. Cosley et al. (2003) asked their subjects to re-rate 3 sets of movies on MovieLens (already evaluated by means of the original 5-point rating scale) with: a binary scale (thumbs up-down), a no-zero scale (range: -3, +3), and a half-star scale (range: 0.5, 5). The authors found out that ratings on all 3 scales correlated well with original user ratings, and suggested that designers might allow users to choose their favourite rating scale and compute recommendations by means of mathematically normalized scores. However, they also observed that users tended to give higher mean ratings on the binary and on the no-zero scales, and that new ratings on the binary scale correlated less strongly with original ratings than new ratings on the no-zero and half-star scales.

Differently from these studies, we aimed at conducting a more comprehensive study of rating scales, considering granularity, neutral point and visual metaphor and doing this in a variety of different usage scenarios.

### 2.3 The design of rating scales

Referring to the design of rating scales, Swearingen and Sinha (2002) suggested to adopt a mix of different types of questions (e.g., expressing binary liking versus rating items on a Liker-like

scale) and to provide constant feedback on user contributions in order to keep users from getting bored or frustrated during the rating. Herlocker et al. (2004) pointed out that the granularity of user preferences with respect to recommended contents may be different from the granularity managed by the specific recommender system. Thus, an appropriate rating scale should allow users to distinguish among exactly as many levels of liking as it makes sense to them.

In (van Barneveld and van Setten 2004), the authors defined the main elements of interface aspects for presenting system predictions and collecting explicit user feedback in the context of a TV recommender system: 1) presentation form; 2) scale of the prediction or rating (including range, precision, symmetric versus asymmetric and continuous versus discrete); 3) visual symmetry or asymmetry; and 4) use of colour. They also found that most users prefer to have predictions presented by means of 5-star interfaces, while they are less in agreement regarding interfaces to provide input to the system, consistently with the findings of Cena et al. (2010).

Also Nobarany et al. (2012) concentrated their work on the design of "opinion measurement interfaces[11]" (as they called the rating scales). They identified two axes: Measurement Scale (absolute rating vs. relative ranking) and Recall Support (previously recorded opinions). They experimented the use of two prototypes with different rating and ranking scales (Stars, Stars + Recall, Binary, List). The measures used to compare the scales were: speed, accuracy, mental demand, and suitability of organization, fun to use and overall preference. Quantitative and qualitative final results showed that 1) a rating interface that provides recall support with examples of users' previous choices is preferred, 2) rating accuracy is perceived as more important than rating speed.

For Usman et al. (2010), the star rating system is the de-facto standard for rating a product, being regarded as one of the most appealing rating systems for the direct interaction with the

[11] Notice that, by the term "opinion measurement interfaces", the authors specifically refer to rating scales. In the context of Web 2.0, and with the proliferation of user-generated contents, however, the analysis of people's opinions, attitudes and emotions expressed through natural language texts is gaining relevance. The interested reader can refer to Petz et. al. (2015) for a discussion of the challenges arising from user-generated content (with respect to more formal texts) and an evaluation of the effectiveness of several text preprocessing algorithms.

users. However, due to its limitation when comparing items with different numbers of raters (i.e., stars do not convey any information regarding the number of people who reviewed a certain item), the authors argue that the visual strength of the five stars is not enough to declare that they are the best option also for recommender systems. For this reason they proposed a Relative Ranking, where a benchmark item is used to compare other items until they reach its number of points. According to the authors, this method is more realistic and useful at a glimpse than the five stars system. Sparling and Sen (2011) investigated the costs, in terms of mental effort and time, which are associated to rating scales with different granularities and carried out an online survey where they compared the following four scales: unary ("like it"), binary ("thumbs up"/"thumbs down"), five-star, and 100-point slider. They found that users' average rating time increases with the granularity of rating scales, while there are no significant differences as far as cognitive load is concerned, if the unary scale (which requires a significantly less hard work on the part of users) is excluded. Moreover, the participants in their survey preferred the stars and the thumbs, disliking the scales at the ends of the granularity spectrum (unary scale, slider).

**2.4 User personality and its impact on user ratings**

Many studies in recommender systems focused on differences in rating behavior that can be related to users' personality. This is a complementary perspective with respect to the one proposed in our paper, but we put it here for completeness in the discussion. As explained in Schafer et al. (2007), "one optimistic happy user may consistently rate things 4 out of 5 stars that a pessimistic sad user rates 3 out of 5 stars", even if they actually mean the same. For example, Hu and Pu (2013) showed the influence of user personality characteristics on user rating behaviors. They conducted an online survey with 122 participants: they used the Five Factor Model (openness to experience, conscientiousness, extraversion, agreeableness, neuroticism) and the Big Five Inventory (John and Srivastava1999) in order to describe the personality of each user in the first part of the survey. An adjustment using averages could be adopted in order to compensate for

idiosyncratic behavior. For example, Schafer et al. (2007) predicted user ratings for an item *i* as a positive or negative variation with respect to their average rating, and the amount of such variation is determined as a function of the difference between the rating other users assigned to item i and their average rating. Similar approaches are described in Herlocker et al. (2004), Adomavicius and Tuzhilin (2005) and in Goldberg et al. (2001).

## 3. EMPIRICAL EVALUATIONS

We carried out three user studies in different domains and contexts, all of them with the aim of validating our main hypothesis regarding if and how rating scales have a personality and thus are able to influence users' ratings. The first evaluation (Section 4) focused on existing movie websites and aimed to evaluate how the same movies are rated using different scales. The second one (Section 5) exploited a system designed by some of the authors, which acts as a museum visitor's guide for the Hecht[12] museum in Haifa, Israel, and offers different rating scales (Kuflik et al. 2015). This second evaluation took place in a different context from the web, in order to assess external validity of the other results. The third one (Section 6) is a controlled experiment where participants have to explicitly translate ratings from one scale to another.

The three evaluations differ in various aspects, thus allowing us to explore the issue of rating scale personality from different points of view:

- **Domain**: the first and third evaluations were carried out in the movies domain, while the second one in the museum/cultural heritage domain;
- **Procedure**: the first and second evaluations used real-world data, generated by people using a particular system in situ, while the third evaluation was a controlled experiment;
- **Number of participants**: evaluation 1 is based on large datasets with thousands of ratings per system, evaluations 2 and 3 use data from about 300 subjects each;

[12] http://mushecht.haifa.ac.il/Default_eng.aspx

- **Type of comparison**: the first and second evaluations compare ratings given to a certain item by different people, and each person is expected to have used only one website/rating scale; in the third evaluation, each participant rated all the items with all the rating scales;
- **Context**: evaluation 1 used real-world data coming from ratings given in a context of interaction with real web sites, while evaluation 2 used real-world data coming from ratings given in the context of interaction with a museum guide and finally evaluation 3 was a controlled experiment performed using an ad-hoc web-based interface.

Rating scales used in our evaluations (with their corresponding websites, as far as the first evaluation is concerned) were selected with an eye to the fact that they allowed us to cover different possible values for the most distinctive features we identified in a previous work (Cena and Vernero, 2015), i.e., visual metaphor, icon, granularity, range, positive/negative, neutral position, point mutability. A description of the meaning of these features is provided in Table 1.

**Table 1. The most distinctive features included in the model of rating scales devised by Cena and Vernero (2015)**

| Feature | Description |
|---|---|
| Visual metaphor | The metaphor, used in the visual appearance of a rating scale, which can impact on its interpretation and emotional connotation (e.g., a smiley face exploits a metaphor related to human emotions). Not all visual presentation forms make use of metaphors. |
| Icon | The specific image or presentation form used in a rating scale. |
| Granularity | The number of positions in the rating scale |
| Range | The minimum and maximum value of the scale (e.g. from 0 to 10) |
| Positive/negative | The presence of either only positive, or only negative, or both kinds of points |
| Neutral position | The presence of a neutral, intermediate position (middle point) |
| Point mutability | Whether the points in a rating scale are represented in the same way or not |

In fact, according to our vision, the personality of a rating scale is somehow related to its objective features, so that scales with different features can be expected to have different effects on users' ratings.

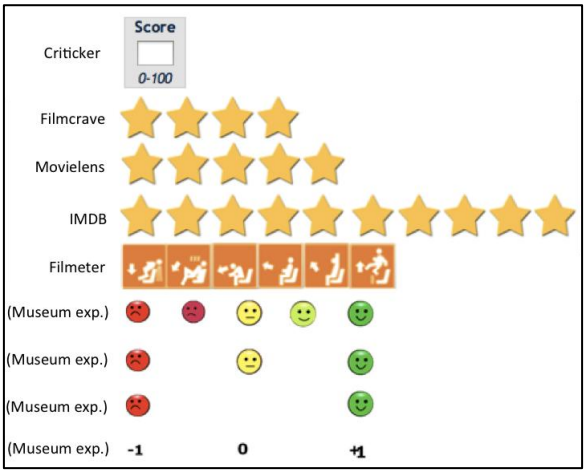
For the first evaluation, rating scale selection was also affected by the kind of data access facilities (e.g., public APIs) offered by websites. Thus, we selected:

- The free text numerical scale ranging from 1-100 which is used in Criticker[13];
- The 6-point rating scale with icons representing a hypothetical movie viewer (with mood from frustrated to enthusiast) exploited by Filmeter[14];
- The 4-star rating scale used in FilmCrave[15], which offers half-stars ratings ;
- The 5-star rating scale used in MovieLens[16], which offers half-stars ratings;
- The 10-star rating scale used in IMDb[17];

For the second evaluation, we used:

- A 5-star rating scale (it was included in all three evaluations because of its status of "standard");
- A 2-point, a 3-point, and a 5-point rating scale with icons representing smiling, sad or neutral faces;
- A 3-point numerical rating scale with the available positions explicitly labeled as "-1", "0" and "+1".

Finally, for the third evaluation, we re-used all the rating scales selected for the first two evaluations. A visual summary of the selected rating scales is presented in Figure 1, while a concise description of their distinctive features is given in Table 3.




[13] http://www.criticker.com
[14] http://www.filmeter.net. Note that the data were gathered from Filmeter in 2013. Currently the system has changed its rating scale.
[15] http://www.filmcrave.com
[16] http://www.movielens.umn.edu
[17] http://www.imdb.com

**Figure 1. The rating scales analyzed in our work, chosen from existent web sites and an ad-hoc experiment**

**Table 2. The analyzed rating scales described under different features**

| Source | Visual metaphor | Icon | Granularity | Range | Positive/negative | Neutral position | Point mutability |
|---|---|---|---|---|---|---|---|
| Criticker | None (but reference to school marks) | None (bare numbers) | 101 | 0-100 | only positive | YES | NO |
| Filmcrave | Neutral, standard rating system | Stars | 8 | 1 star - 4 stars | only positive | NO | NO |
| Movielens, | Neutral, standard rating system | Stars | 10 | 1 star - 5 stars with half-stars ratings | only positive | YES | NO |
| PIL | Neutral, standard rating system | Stars | 5 | 1 star - 5 stars | only positive | YES | NO |
| IMDb | Neutral, standard rating system | Stars | 10 | 1 star - 10 Stars | only positive | NO | NO |
| Filmeter | Human, domain- related (cinema) | Cinema-viewer | 6 | very negative mood - very positive mood | positive + negative | YES | YES (different icons) |
| PIL | Human | Smileys | 5 | very sad face - very happy face | positive + negative | YES | YES (different icons and colors) |
| PIL | Human | Smileys | 3 | very sad face - very happy face | positive + negative | YES | YES (different icons and colors) |
| PIL | Human | Smileys | 2 | very sad face - very happy face | positive + negative | NO | YES (different icons and colors) |
| PIL | Measurement tool | Numbers | 3 | -1 - +1 | positive + negative | YES | NO |

## 4. REAL MOVIE RATINGS ANALYSIS

**Goal.** The goal of this evaluation was to investigate the values of ratings given on heterogeneous rating scales (and using different systems) on the same items. Our hypothesis was that static normalization was not enough for mapping user ratings expressed with different rating scales. In studies done by Cena et al. (2010) and Gena et al. (2011), it was observed that ratings expressed on the same items by means of different rating scales depart considerably from mathematical proportion, and so we can assume that rating scales actually have an influence on user ratings. Thus, we decided to extend the experiments to a broader set of data in order to externally validate those findings.

**Hypothesis**. Our main hypothesis was that ratings given to the same movies using different rating scales could produce different results due to the features of the scales.

**Design**. Five factors (the five movie portals), between-subjects design.

**Subjects.** Anonymous subjects who left ratings on 1199 selected movies on the below described systems. Since we collected only anonymous ratings we do not know exactly the total number of subjects that left ratings on these systems, but just the total number of ratings (see Table 4), since one subject may have rated several times. Note that for selecting the movies we used an availability sampling approach (also known as sample of convenience[18]): the selected movies are those who have ratings on all the six movie portals, and similarly the same subjects and their ratings.

**Apparatus and Materials**. The raw data on ratings were collected both manually, and through public APIs, and through proprietary crawlers (see the detailed description below).

**Procedure**. In order to validate our hypothesis, we collected the ratings given to 1199 different movies by using five different rating scales, which differ in metaphor and granularity (see Section 2), belonging to five different popular movie portals:

- **Criticker**[19], which exploits as rating scale a free text numerical scale from 0 to 100, see Figure 5;
- **Filmeter**[20], which exploits a 1 to 6 rating scale with icons representing a hypothetical movie viewer (from frustrated to enthusiast mood), see Figures 2 and 3[21];
- **FilmCrave**[22], which exploits a 1 to 4 stars rating scale (with the possibility of half-star ratings), see Figure 4.
- **IMDb**[23] , which exploits a 1 to 10 stars rating scale, see Figure 6;
- **MovieLens**[24], which exploits a 1 to 5 stars rating scale (with the possibility of half-star ratings), see Figure 7.

[18] Even though random sampling is the best way of having a representative sample, these strategies require a great deal of time and money. Therefore much research in psychology 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, often used when the population source is not completely defined (Royce and Straits 1999).
[19] http://www.criticker.com
[20] http://www.filmeter.net
[21] Notice that the data were gathered from Filmeter on 2013. Currently the system has changed its rating scales.
[22] http://www.filmcrave.com
[23] http://www.imdb.com
[24] http://www.movilens.com

At the beginning of 2014 we started to collect data from Criticker, since, at that time, it provided an API for extracting movies' data. With the Criticker API, we selected the most popular films (the ones that had a higher number of ratings). We extracted movie information such as title, id, etc., about these films as well as the corresponding IMDb identifier and users' ratings. In particular, for every film we have extracted all the ratings (in the range 0-100) provided by the users. Then we have calculated the average rating values for every movie.

**Figure 2. Filmeter's rating scale**


**Figure 3. Filmeter's average rating's display**


**Figure 4. Filmcrave's rating scale and average rating's display**


**Figure 5. Criticker's rating scale and average rating's display**


**Figure 6. IMDb's rating scale and average rating's display**

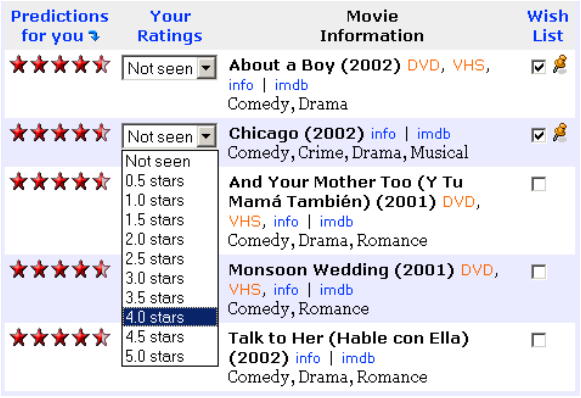



**Figure 7. MovieLens rating scales**

IMDb does not offer any kind of API to query its huge database of films and film-related information. To overcome this constraint and collect data about ratings in IMDb, we have used IMDb movie id extracted from Criticker film list to reach the movie's URL, and then a softbot automatically parsed the corresponding pages and extracted the average ratings and the corresponding number of ratings. Unlike Criticker, from the IMDb pages we did not collect all the single ratings, but only the average ratings, the total number of ratings (and the total number of users) for each movie. Similarly, we manually collected the average ratings, the total number of ratings on the same movies from Filmeter[25], and Filmcrave since there were no APIs available, while for MovieLens we utilized the data set consisting of 20 million ratings[26] and selected the corresponding movies through the IMDb id. Similarly to what was done for Criticker, we calculated the average rating values for each movie.

[25] Filmeter merely shows the integer value of the averages, while other systems, such as IMDb and Filmcrave, show the rounded averages.
[26] http://www.grouplens.org/taxonomy/term/14

We then compared the values coming from the five rating scales. In order to be able to compare the different scales, the users' ratings were normalized onto a 0-1 scale, according to the static mapping described in Section 1. With this conversion mechanism all the values are represented in a comparable 0-1 range.

We saved all the converted data in a database and calculated some descriptive statistics about ratings (mean, standard deviation, etc.), inferential statistic (Kruskal-Wallis test), correlations and regressions.

**Results**. The total number of ratings collected from each system is shown in Table 4, together with the mean and standard deviation. We are aware of the fact that the average number of ratings for movies greatly varies between one system and another. However these numbers represent the actual photograph of the ratings for movies at the time of the evaluation on these systems.

As we can notice in Table 5 the mean values show differences that need to be evaluated in order to discover if they are significant. In particular, IMDb and Criticker show higher average ratings, followed by MovieLens, FilmCrave, and Filmeter. Our hypothesis in interpreting these results, is that the higher granularity and the presence of a neutral point encourage users to rate higher. Criticker offers the finest explicit granularities – and probably pushes the users to be more precise and even stricter- and, in term of granularity, it is followed by IMDb, whose granularity is probably perceived very close to the one of Criticker (100 vs 10). MovieLens has the same granularity of IMDb through the use of half-star ratings but it offers the possibility to choose an explicit neutral point, which collects the 23,6% of ratings (see Figure 8). Figure 8 indeed shows the percentage of ratings per rating position. We can observe that the half-star positions are not frequently used, with respect to full-star positions. So we can conclude that, generally, users perceive as main granularity of the scales the one associated to the full star position. The same probably happens with FilmCrave, however we could not compute these calculation since we only have available the average ratings per movie. Filmcrave shows an average value closer the one of Filmeter, which shows the lowest average: they both do not have a neutral point, and their granularity is close and

based on even numbers. Indeed Filmeter shows the lowest average, also associated with the widest rating range (0-1) probably because the icons associated to rating positions push the users to use the lowest rates, and lacks of a neutral point.

In order to evaluate the significance of these average results, we performed Kruskal-Wallis one-way analysis of variance by ranks, a non-parametric method of testing equivalent to the parametric one-way analysis of variance (ANOVA) on the movie average per systems. When the Kruskal-Wallis test leads to significant results, at least one of the samples is different from the other samples. The Kruskal-Wallis test rejected the null hypothesis and confirmed that the values are different with significance at the 0.05 level. However, we also wanted to determine which of these groups significantly differ from each other. Thus we run a post-hoc pairwise comparison that showed that the differences between the means are all significant.

**Table 3. Main statistic for the movie rating data: number of ratings.**

| | **Criticker** | **Filmeter** | **FilmCrave** | **IMDb** | **MovieLens** |
|---|---|---|---|---|---|
| **Number of total ratings** | 5,311,371 | 167,798 | 346,521 | 116,601,797 | 11,872,347 |
| **Average number of ratings per movie** | 4,429.83 | 142.56 | 292.92 | 97,249.21 | 9,926.71 |
| **Standard deviation** | 2,668.28 | 126.66 | 249.19 | 80,499.32 | 9,892.37 |

**Table 4. Main descriptive statistics for the movie average rating values.**

| | **Criticker** | **Filmeter** | **FilmCrave** | **IMDb** | **MovieLens** |
|---|---|---|---|---|---|
| **Mean** | 0.6715 | 0.5655 | 0.5703 | 0.6935 | 0.6537 |
| **Std. Error of Mean** | 0.00295 | 0.0042 | 0.00363 | 0.00298 | 0.00333 |
| **Std. Deviation** | 0.10201 | 0.1441 | 0.1247 | 0.10305 | 0.11518 |
| **Minimum** | 0.24 | 0 | 0.06 | 0.19 | 0.19 |
| **Maximum** | 0.88 | 1 | 0.85 | 0.96 | 0.88 |

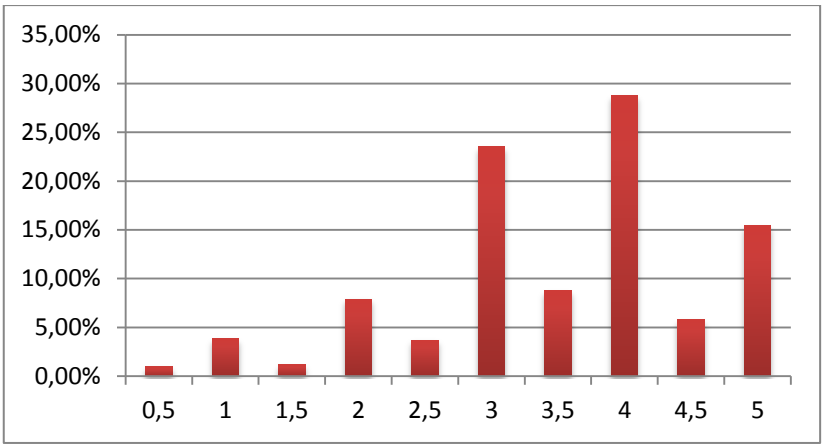



**Figure 8. Distribution of ratings per scale position in MovieLens.**

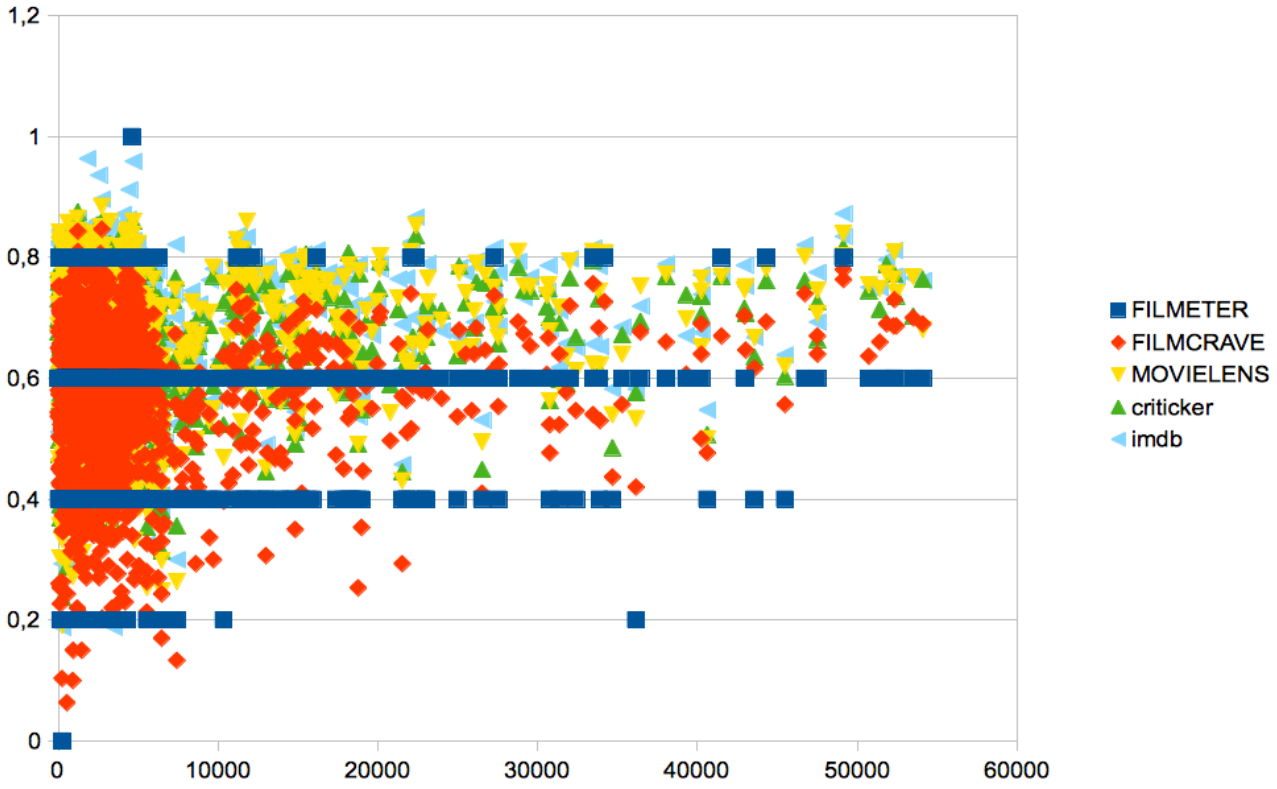



**Figure 9. Graphic of the correlations among all the scales.**

**Table 5. Main statistic for the movie average rating values: correlations.**

| | | CRITICKER | FILMETER | FILMCRAVE | IMDB | MOVIELENS |
|---|---|---|---|---|---|---|
| **CRITICKER** | Pearson Correlation | 1 | ,791** | ,943** | ,944** | ,946** |
| **FILMETER** | Pearson Correlation | ,791** | 1 | ,783** | ,767** | ,740** |
| **FILMCRAVE** | Pearson Correlation | ,943** | ,783** | 1 | ,934** | ,915** |
| **IMDB** | Pearson Correlation | ,944** | ,767** | ,934** | 1 | ,933** |
| **MOVIELENS** | Pearson Correlation | ,946** | ,740** | ,915** | ,933** | 1 |

****All the correlations are significant at the 0.01 level.**

**Table 6. Main statistic for the movie average rating values: regressions**

| Independent variable | Dependent variable | Pearson's *R* | Explained variance | p | Regression coefficient | A | Equation |
|---|---|---|---|---|---|---|---|
| **Criticker** | Filmeter | 0.791 | 62% | 0.000 | 1.114 | -0.182 | Y = 1.114X – 0.182 |
| | Filmcrave | 0.943 | 88% | 0.000 | 1.162 | -0.209 | Y = 1.162X – 0.209 |
| | IMDb | 0.944 | 89% | 0.000 | 0.953 | +0.053 | Y = 0.953X + 0.053 |
| | MovieLens | 0.946 | 89% | 0.000 | 1.067 | -0.063 | Y = 1.067X – 0.063 |
| **Filmeter** | Criticker | 0.791 | 62% | 0.000 | 0.562 | +0.354 | Y = 0.562X + 0.354 |
| | Filmcrave | 0.783 | 61% | 0.000 | 0.683 | +0.184 | Y = 0.683X + 0.184 |
| | IMDb | 0.767 | 59% | 0.000 | 0.549 | +0.383 | Y = 0.549X + 0.383 |
| | MovieLens | 0.740 | 55% | 0.000 | 0.592 | +0.319 | Y = 0.592X + 0.319 |
| **FilmCrave** | Criticker | 0.943 | 88% | 0.000 | 0.765 | +0.235 | Y = 0.765X + 0.235 |
| | Filmeter | 0.783 | 61% | 0.000 | 0.897 | +0.053 | Y = 0.897X + 0.053 |
| | IMDb | 0.934 | 87% | 0.000 | 0.763 | +0.258 | Y = 0.763X + 0.258 |
| | MovieLens | 0.915 | 84% | 0.000 | 0.838 | +0.157 | Y = 0.838X + 0.157 |
| **IMDB** | Criticker | 0.944 | 89% | 0.000 | 0.934 | +0.024 | Y = 0.934X + 0.024 |
| | Filmeter | 0.767 | 59% | 0.000 | 1.071 | -0.177 | Y = 1.071X - 0.177 |
| | Filmcrave | 0.934 | 87% | 0.000 | 1.143 | -0.222 | Y = 1.143X - 0.222 |
| | MovieLens | 0.933 | 87% | 0.000 | 1.042 | -0.069 | Y = 1.042X - 0.069 |
| **MovieLens** | Criticker | 0.946 | 89% | 0.000 | 0.838 | +0.124 | Y = 0.838X + 0.124 |
| | Filmeter | 0.740 | 55% | 0.000 | 0.925 | -0.039 | Y = 0.925X – 0.039 |
| | Filmcrave | 0.915 | 84% | 0.000 | 0.999 | -0.082 | Y = 0.999X – 0.082 |
| | IMDb | 0.933 | 87% | 0.000 | 0.836 | +0.147 | Y = 0.836X + 0.147 |

In order to understand the general trend of rating behavior, we have first plotted the rating trends on a graph that orders the Criticker rating values from the smallest to the largest and accordingly the other ratings, see Figure 9.

From this general view we can see that there are emerging trends between: IMDb and Criticker (in Figure 9a the Criticker ratings are partially covered by the ones of IMDb), followed by MovieLens, FilmCrave, and Filmeter. In order to investigate the nature of these relationships we calculated correlations. In particular, we calculated Pearson's *r* coefficient to highlight significant linear correlations among pairs of scales and then we performed a linear regression analysis to define linear functions which allow predicting user ratings on a particular scale, given their ratings on another scale.

As we can notice in Table 6 all the rating values are significantly correlated. Following the Cohen's classification system (Cohen 1988), we consider only large relationships, i.e., the ones with $r > 0.5$ and explained variability > 0.25. All the correlations among scales are large, and the ones between Criticker, MovieLens, IMDb and Filmcrave are extremely large. Notice that, the extremely

large correlations between IMDb and Criticker is also due to the fact that Criticker offers to IMDb users the possibility of importing their IMDb ratings into Criticker.

Looking at Table 7, we can observe that the relationship among all the variables is significant, and we can use linear regression to predict one variable from another. If the variables were bound by a precise mathematical mapping, the regression coefficient would be equal to 1. However Table 7 shows regression coefficients larger and smaller than 1. So we can assume that the different values of the regression coefficients are caused by the influence of the features of the rating scales. We can also notice that regression coefficients larger than 1 describe faster changes in the dependent variable, while smaller coefficients describe lower changes. Negative signs of the intercepts highlight the presence of lower values in the dependent variable.

## 5. RATING SCALES IN THE MUSEUM

**Goals**. In order to further explore the idea that rating scales may influence user behaviour and test its external validity in a context other than the web, we designed a study using a museum visitors' guide system. Initial results of that study, published in (Kuflik et al. 2012), showed that indeed, users gave different scores when they were using different rating scales. Here we report on the results of the complete study (overall 251 logs analyzed out of about 600 visit logs of real participants' visits; 3-4 times the number of participants reported in the initial study).

**Hypothesis**. Our main hypothesis was that ratings given to the same presentations using different rating scales would produce different results due to the influence of each rating scale.

**Design**. Five factors (the five rating scales), between-subjects design.

**Subjects**. The participants (about 600, out of them only 251 logs were used due to various problems) were normal museum visitors that used a museum visitors' guide during their visit. The visitors were not identified (only by a user ID that was defined when the guide was given to them). No personal information was collected.

**Apparatus and Materials**. In early 2011 a museum visitors' guide system was introduced at the Hecht museum, a small archaeological museum located at the University of Haifa, Israel. The system is described in (Kuflik et al. 2015). It is a web-based system that allows users to freely walk around in the museum, wearing a small proximity sensor and carrying an iPod touch. When they are detected in the vicinity of a point of interest, they are offered a selection of multimedia presentations about objects of interest. Once they have selected a presentation and viewed it, they are required to provide feedback about their satisfaction from the presentation before continuing the visit (i.e. providing feedback is mandatory before the user continues to use the system). As part of the design of the user interface 5 different feedback mechanisms designs (presented in Fig. 9) were implemented and integrated into the system in order to explore whether the interface design of the rating scales has an impact on the ratings, as suggested by Kaptein et al. (2010). The scales in this experiment differ in granularity (from 2 to 5 points), in metaphors (human emotions: the smiley faces; school marks/degrees: the numerical scale; scoring/ranking: the stars), in the presence of a neutral position (present in stars, in 3-points faces, in the numerical scale) and in numbering (there is a scale consisting of -1, 0, 1). They were selected due to their popularity (stars and faces) and due to the fact that stars are neutral while smiley faces have an emotional aspect. Numbers were chosen in order to see what may be the impact of a clear negative value on ratings.

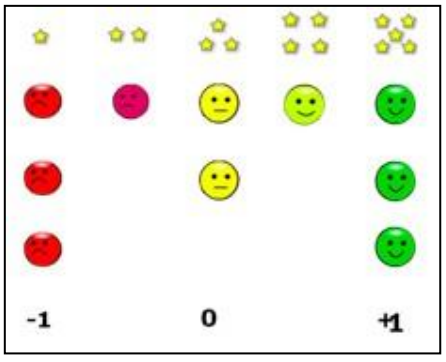



**Figure 10. The five rating scales that could be randomly selected for use in the museum visitors' guide system**

**Procedure**. We collected visit logs as part of our regular operation where the visitors guide was offered to normal museum visitors during regular museum opening hours. For experimentation

purposes, whenever a visitor logged in and started using the system, a randomly selected rating scale was activated for them and used throughout the entire visit. The interactions of the visitors with the system were logged, where the log contained (among other data) the location of the visitor, the presentation viewed, whether it was completed or stopped and the rating the visitor gave to the presentation. The experimentation stated in October 2011 and ended in January 2014 when we had about 600 logs of visitors out of them, 292 logs were good (there were quite a few logs that had problems: they were part of another specific experimentation, they were not recorded properly etc.) and 251 that had 3 ratings or more were analyzed. On average, a visitor rated 14.8 presentations (min=3, max=91, median=12, STD=12.2).

With respect to the rated presentations, out of 346 different presentations available for the users to rate, 288 were actually rated by users during the visits. We had 3747 rating records, hence a presentation got an average of 13 ratings in 5 rating scales, resulting an average of 2.6 ratings in every rating scale. Moreover, in many cases items were not rated in all five rating scales but only in some of them. This study differs considerably from the other studies reported here with respect to the numbers of rating available for analysis. Hence we provide only the overall results without any further statistical analysis that may not provide any additional insights given the relatively small number of ratings per rating scale and item.

**Results**. In general, all visitors scored the presentations high. Figure 13 presents the average ratings of the 5 different methods used in the experiment. In order to be able to compare the various scales, all were converted to 0-1 scale according to the above presented static mapping, i.e., when there were 2 values, then 0 and 1 were used, when there were 3 values, then 0,0.5 and 1 were used.

Looking at Figure 10 we can see that the average of the stars (1-5) and 5 faces are close as well as numerical values (-1, 0, +1) and 3 faces. It is interesting to see among the faces scales that 2 faces provided highest scores, then 3 faces and then 5 faces. Looking at the medians it seems that all the scales with the smallest range (2 stars and then -1, 0 ,+1 and 3 stars) had similar scores,

while the two scales with the highest but identical range (five faces and stars) had, again, a lower but similar score. Given the specific setting of the museum, it is no wonder that in general scores were high, as it is also common in other areas. The museum is a pleasant environment where visitors come to enjoy, the presentations are informative, interesting and in high quality. Still, when visitors had the choice, that is when the scale had more granularity, they were a bit more critical. Statistical analysis (Wilcoxon test and Mann-Whitney with Bonferroni correction) revealed that the differences between all scales are significant with one exception - the difference between "-1, 0, +1" and "3 faces" was not statistically significant.

Given all the above, we can say that whenever the range of the rating scale is similar, than the conversion between them should be relatively easy as it looks like here they have similar averages and medians. We can also note that when the range of scores gets smaller, the average score gets higher – users tend to rate positively, but when there is more flexibility, they use it. So if you'd like to get a better feedback, than the range of 5 enables a better feedback.

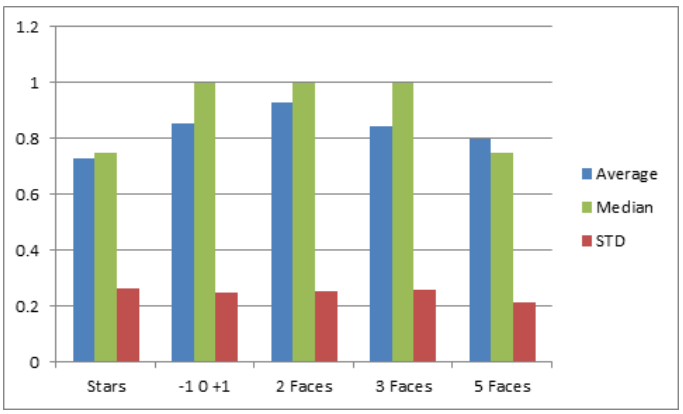



**Figure 11. Average, median and standard deviation of the ratings for the 5 rating scales used in the museum experiment**

## 6. USERS' RATING BEHAVIOUR IN A WEB-BASED USER STUDY

**Goals**. We set up a web-based user study in order to further analyse the differences within the set of rating scales used in our previous studies (see Figure 2). The main goal of this study was *understanding how users map their ratings on different rating scales.* More specifically, we

focused on the conversion of three input ratings (2 stars, 3 stars, and 4 stars) from the 5-star rating scale to all other scales. The 3-star rating was chosen since the use of a midpoint in rating scales is much debated in the literature and we aimed at observing the way people translate it when they are forced to make a choice. The other two ratings were chosen in order to understand how people convert ratings that, although being quite clearly "low" or "high", are not extreme values (1 and 5 ratings were ignored as these are extreme values that are easy to convert). The 5-star scale was chosen as the input scale due to its familiarity to most users. In fact, according to a pilot survey we carried out on 45 web-based rating systems in book, movie and travel domains, stars were the most popular icon (20/45), while 5 points was the most exploited granularity (16/45).

**Hypothesis**. We hypothesized that there are statistically significant differences between the ratings of users when they convert a given rating in a given scale to different rating scales.

**Design**. 3 x 8 factorial design, i.e. input rating x output rating scale, within subjects.

**Subjects**. We recruited 330 participants among the Facebook contacts of the authors, according to an availability sampling strategy[27]. Of them, 206 were Italians, 105 were Israelis and 19 were from other countries. 266 participants (174 Italians, 74 Israelis and 18 from other countries) completed the whole test.

**Apparatus and Materials.** Participants took part in our study through web pages (in English) that they could access on their own, at their convenience. All questions were accompanied by figures or interactive representations of the input ratings and/or rating scales they mentioned.

**Procedure.** This study was carried out in 2013. Participants were asked to think of three movies that they would rate with 2, 3, and 4 stars out of 5, respectively, and to rate them using each one of the eight output scales[28].

[27] The availability sampling is a sampling of convenience, based on subjects available to the researcher, often used when the population source is not completely defined. Even though random sampling is the best way of having a representative sample, these strategies require a great deal of time and money. Therefore much research in psychology is based on samples obtained through non-random selection (Royce and Straits 1999).

[28] Notice that both rating scales (in the first and second part) and adjectives (in the second part) were always presented to participants in a fixed order. This might be considered as a limitation for this study, since it might determine some order-related bias.

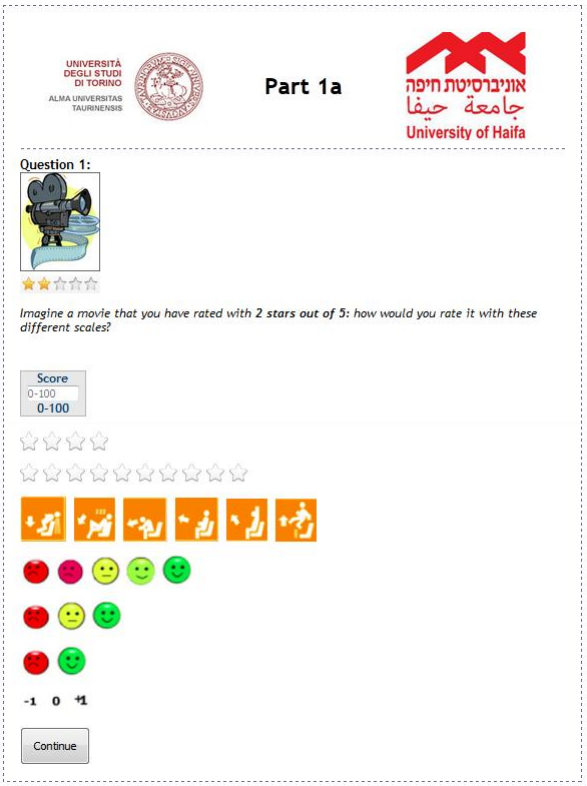



**Figure 12. An example web page used for our user study**

**Results.** To allow comparison, users' ratings were converted to a 0-1 range following the same static mapping we used for the previous experiments Results are presented in Figure 13, where many small differences can be noticed.

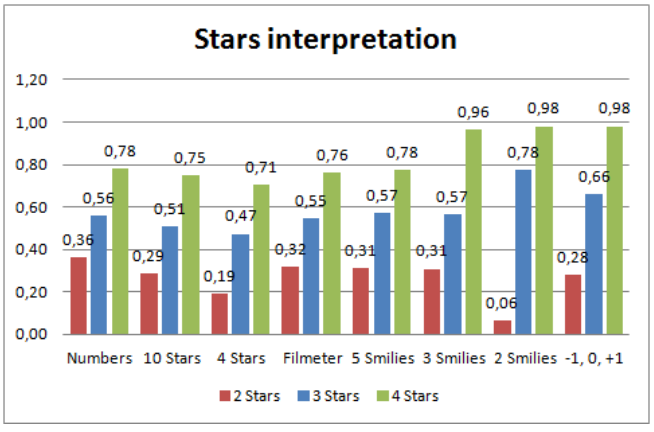



**Figure 13. 2,3 and 4 stars rating interpretation in 8 different rating scales**

Considering user conversions of the 2-, 3- and 4-stars ratings separately, the Friedman test[29] confirmed our intuition that there are significant differences among user ratings on the eight output scales in all three cases, Then, we compared all possible pairs of scales using Wilcoxon signed ranks test (note that, with eight rating scales, there are 28 different pairs). Results showed statistically significant differences for most of them, thus supporting our idea that scales can actually exert an influence on user ratings. However, an agreement (i.e., no significant differences, indicating that two scales have a similar influence on users) could still be observed in at least 5 cases out of 28 for each one of the three rating conversions. The most significant cases of agreement are the following:

- 3 and 5 smileys were found to be in agreement with 10 stars in the translation of 2 and 3, and of 3 and 4 stars respectively. Similarly, they were in agreement with Filmeter in the translation of 2 and 3 star ratings (3 smileys) and 3 and 4 star ratings (5 smileys). Three smileys and five smileys themselves are found to be in agreement for the translation of 2 and 3 star ratings, which is not surprising since these two scales are quite similar (same metaphor, use of color, use of explicitly negative points, close granularity, and a neutral point in both cases).
- 3 smileys were found to be in agreement with "-1, 0, +1" for the translation of 3 and 4 star ratings. Also for these two scales, which were already found to be in agreement in the museum experiment (Section “Rating scales in the museum”), similarity in user ratings can be explained by the fact that they have the same granularity, a neutral point and explicitly represented negative positions, even if they have very different visual metaphors.

Since we noticed some differences between Israelis and Italians, we analysed their answers separately. Considering the 2-star rating (Figure 14), we can observe that, in cases where the means do not agree, the Israelis were usually more critical than the Italians (the exception being

[29] The Friedman test is a non-parametric statistical test similar to ANOVA. It is used to detect whether at least one of the examined samples/items is significantly different from the others.

the “numbers” scale). Statistical analysis (MANOVA[30] with Bonferroni correction) revealed that there are significant differences between the two national groups in the “numbers” scale, in the "-1, 0, +1" scale, and in the 5 smileys scale ($p<0.05$). Notice that the observed average difference is smaller for numbers than for the other two scales, while relatively large differences between Italians and Israelis (e.g., in the case of the 3 smiley scale) were not found to be significant.

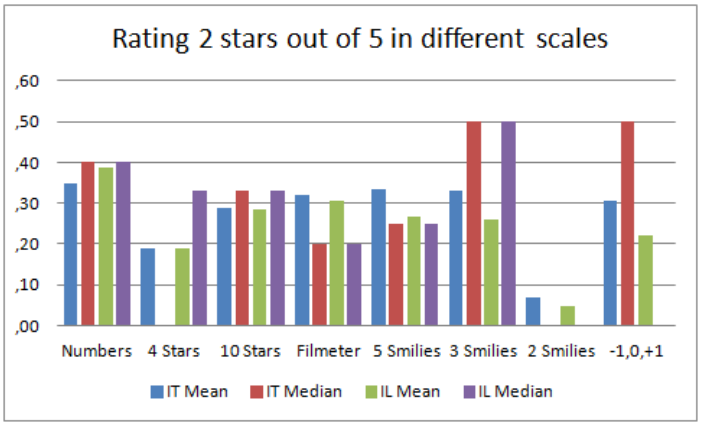



**Figure 14. Comparison between Italians and Israelis: 2 stars rating conversion**

As far as the 3-star rating is concerned, visual inspection of Figure 15 suggests that differences between the two groups are systematically larger, with the exceptions of the “numbers” scale (smaller difference) and of the 4 stars, 10 stars and Filmeter scales that again elicited very close rating interpretations in the two groups. Again, the Israelis were a bit more critical in their ratings. Statistical analysis (MANOVA) revealed that there are significant differences regarding the 5-, 3-, and 2-smileys scales, and the "-1, 0, +1" scale ($p<0.05$).

On the contrary, differences are generally very small for the 4-star rating (Figure 16), and they are statistically significant only for what concerns the interpretation of the numbers and 5-smileys scales ($p<0.05$).

[30] MANOVA was used instead of ANOVA since we had to consider two dependent variables, i.e. ratings by Israeli and ratings by Italians.

Following the same approach we described in Section 4, we joined the data deriving from users' translations of the three input ratings and performed further analyses to study whether and how ratings can be translated from a scale to another, concentrating on the existence of simple linear correlations. More specifically, we considered pairs consisting of five stars and each one of the other scales. Five stars played the role of independent variables in the linear regression study and were chosen because they were used as the standard input scale in our web-based study.

Results are presented in Table 8, including the linear-regression-based equation that allows to translate ratings from a scale to another. Notice that scales are ordered according to the value of *r* (or, equivalently, to the percentage of explained variability), so that those which show a higher linear correlation with the independent variable are listed first. All results are significant, with p = 0,01, and, following Cohen's classification system (Cohen 1988), all correlations can be considered very large.

**Table 7. Linear regression analysis: results for five stars as the independent factor.**

| | **Dependent Variable** | **Pearson's *R*** | **Explained variance** | **p** | **Regression Coefficient** | **A** | **Equation** |
|---|---|---|---|---|---|---|---|
| **All** | ten stars | 0.874 | 76% | 0.000 | 0.927 | 0.049 | Y = 0.93X + 0.05 |
| | numbers | 0.83 | 69% | 0.000 | 0.838 | 0.146 | Y = 0.84X + 0.15 |
| | five smileys | 0.823 | 68% | 0.000 | 0.934 | 0.084 | Y = 0.93X + 0.08 |
| | Filmeter | 0.804 | 65% | 0.000 | 0.896 | 0.09 | Y = 0.90X + 0.09 |
| | three smileys | 0.791 | 63% | 0.000 | 1.306 | -0.043 | Y = 1.31X − 0.04 |
| | −1, 0, +1 | 0.781 | 61% | 0.000 | 1.429 | -0.081 | Y = 1.43X − 0.08 |
| | two smileys | 0.772 | 59% | 0.000 | 1.851 | -0.322 | Y = 1.85X − 0.32 |
| | four stars | 0.761 | 58% | 0.000 | 1.036 | -0.064 | Y = 1.04X − 0.06 |
| **Italians** | ten stars | 0.86 | 74% | 0.000 | 0.919 | 0.055 | Y = 0.92X + 0.06 |
| | numbers | 0.81 | 66% | 0.000 | 0.838 | 0.137 | Y = 0.84X + 0.14 |
| | Filmeter | 0.8 | 63% | 0.000 | 0.884 | 0.101 | Y = 0.88X + 0.10 |
| | five smileys | 0.79 | 62% | 0.000 | 0.906 | 0.118 | Y = 0.91X + 0.12 |
| | three smileys | 0.77 | 60% | 0.000 | 1.266 | -0.005 | Y = 1.27X − 0.01 |
| | −1, 0, +1 | 0.76 | 58% | 0.000 | 1.351 | -0.016 | Y = 1.35X − 0.02 |
| | two smileys | 0.76 | 58% | 0.000 | 1.816 | -0.285 | Y = 1.82X − 0.29 |
| | four stars | 0.75 | 56% | 0.000 | 1.045 | -0.064 | Y = 1.05X − 0.06 |
| **Israelis** | ten stars | 0.9 | 82% | 0.000 | 0.936 | 0.047 | Y = 0.94X + 0.05 |
| | five smileys | 0.9 | 80% | 0.000 | 0.972 | 0.026 | Y = 0.97X + 0.03 |
| | numbers | 0.88 | 77% | 0.000 | 0.838 | 0.168 | Y = 0.84X + 0.17 |
| | three smileys | 0.83 | 68% | 0.000 | 1.391 | -0.118 | Y = 1.39X − 0.12 |
| | Filmeter | 0.82 | 67% | 0.000 | 0.923 | 0.075 | Y = 0.92X + 0.08 |
| | -1,0,+1 | 0.81 | 66% | 0.000 | 1.518 | -0.162 | Y = 1.52X − 0.16 |
| | two smileys | 0.79 | 62% | 0.000 | 1.916 | -0.393 | Y = 1.92X − 0.39 |
| | four stars | 0.79 | 62% | 0.000 | 1.016 | -0.061 | Y = 1.02X − 0.06 |

The strongest correlation exists between five stars and ten stars ($r$ = 0,874). This result is not surprising, since we previously observed statistical agreement between ratings on these two scales

and we could already find an extremely large correlation in the real movie ratings analysis, where MovieLens was treated as a 5-point scale by excluding half-star ratings.

Moreover, ratings on the 5-star scale are good predictors for ratings on numbers, five smileys and Filmeter scales. While five stars, five smileys, ten stars and numbers all have "related" granularity, Filmeter is similar to the other scales in that it has a fine granularity.

Considering the two national populations separately, we found no big differences between Italians and Israelis, but we can observe that correlations are systematically higher for Israelis. In fact, the average value for Pearson correlation is 0.83 for Israelis and 0.78 for Italians.

In order to allow comparison with the real movie ratings analysis, we also studied the existence of linear relationships in pairs consisting of the following scales: numbers (Criticker), Filmeter, four stars (Filmcrave) and ten stars (, IMDB); see Tables 9 and 10. For simplicity, the scale with the finest granularity played the role of independent variable. Our data show that there are large relationships for all the pairs, thus confirming our findings for the real movie ratings analysis.

**Table 8. Main statistics for the web-based user study: correlations.**

| | | Numbers | Filmeter | Four stars | Ten stars |
|---|---|---|---|---|---|
| **Numbers** | Pearson Correlation | 1 | ,768** | ,731** | ,868** |
| **Filmeter** | Pearson Correlation | ,768** | 1 | ,706** | 0,817** |
| **Four stars** | Pearson Correlation | ,731** | ,706** | 1 | ,780** |
| **Ten_stars** | Pearson Correlation | ,868** | 0,817** | ,780** | 1 |

******All the correlations are significant at the 0.01 level.

**Table 9. Linear regression analysis: results for numbers, Filmeter, 10 stars and 4 stars pairs**

| **Independent variable** | **Dependent variable** | **Pearson's *R*** | **Explained variance** | **p** | **Regression coefficient** | **A** | **Equation** |
|---|---|---|---|---|---|---|---|
| **Numbers** | ten stars*<br>Filmeter*<br>Four stars* | 0.868<br>0.768<br>0.731 | 75%<br>59%<br>53% | 0.000<br>0.000<br>0.000 | 0.912<br>0.849<br>0.986 | -0.003<br>0.058<br>-0.105 | Y = 0,912X – 0.003<br>Y = 0.849X + 0.058<br>Y = 0.986X - 0.105 |
| **10 stars** | Filmeter*<br>Four stars* | 0.817<br>0.780 | 67%<br>61% | 0.000<br>0.000 | 0.859<br>1.000 | +0.098<br>-0.060 | Y = 0.859X + 0.098<br>Y = 1.000X + 0.060 |
| **Filmeter** | Four stars* | 0.706 | 50% | 0.000 | 0.862 | -0.012 | Y = 0.862X + 0.012 |

* Large relationship according to Cohen's classification (Cohen 1988),

Summing up our main findings, we can observe that:

- In most cases, there are statistically significant differences between user ratings between any two rating scales we tested. This result allows us to accept our hypothesis and claim that rating scales have an influence on users' rating behavior.
- Cases of agreement between two scales can be explained by the fact that they have various objective features in common. In most cases, the common feature is granularity, which can be the equivalent (as for three smileys and "-1,0,+1"), sometimes quite close (5 smileys and Filmeter), and sometimes directly related (5 smileys and 10 stars). From the cases of agreement we observed, only 3 smileys did not have a similar granularity to 5 smileys and (particularly) 10 stars.
- All rating scales are significantly correlated, so that ratings can be translated from a scale to another using linear-regression-based equations.
- The way different rating scales are perceived is likely to be influenced by cultural background aspects. Israelis usually seem more critical than the Italians; at the same time, they seem to be more consistent when rating items on different scales.

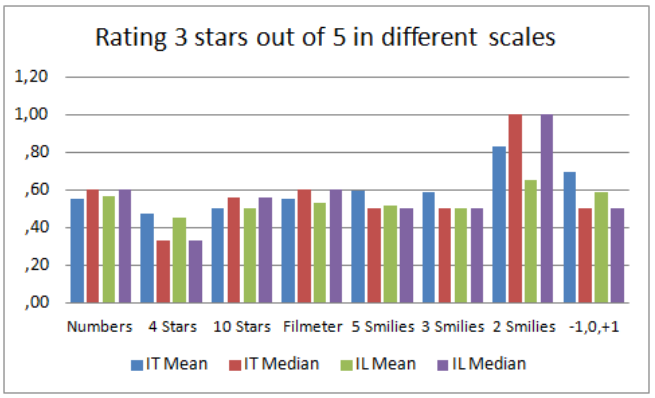



**Figure 15. Comparison between Italians and Israelis: 3 stars rating conversion**

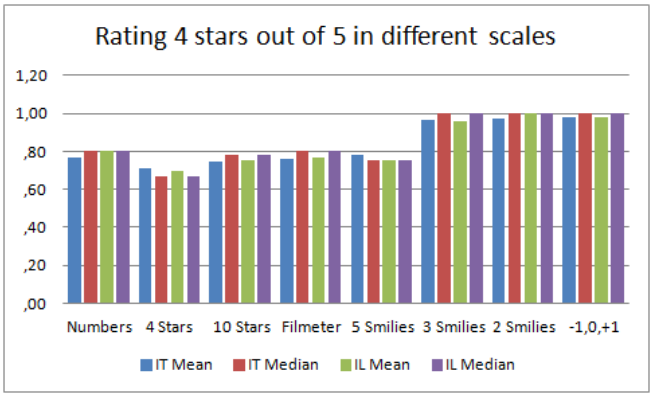



**Figure 16. Comparison between Italians and Israelis: 4 stars rating conversion**

## 7. DISCUSSION

Our work complements and adds to the large body of knowledge about rating scales that already exists. Our study in particular focused on the impact of the appearance of rating scales on user ratings in recommender systems.

In this paper we have provided new evidence that rating scales do have some distinctive features that somehow influence the users' rating behavior. This is confirmed by our statistical analysis that showed that the average differences between rating scales (when transformed to a common scale) are significant in most cases, for all the three empirical evaluations we carried out (Sections 4, 5 and 6). In particular, in experiment 1 (Section 4) and 3 (Section 6) most of the rating data are significantly correlated, thus showing a similar behavior of the users while they rate items, but with up/down differences, in our opinion due to the specific characteristics of the rating scales. Notice that, for experiment 2 (Section 5), the smaller number of ratings (2.6 per rating scale per item) prevents us from calculating correlations.

### 7.1 The impact of granularity on rating

In particular, for what concerns experiment 1, IMDb, Criticker and MovieLens show higher average ratings, followed by FilmCrave, and Filmeter. Our hypothesis is that the 10-based granularity encourages users to rate higher. Criticker offers the finest granularity, and probably pushes users to be more precise and even stricter. IMDb offers an explicit 10-point granularity and, in comparison with MovieLens (which offers an implicit 10-point granularity through half-stars

ratings and gives the possibility to choose a neutral point) its average results are indeed higher. MovieLens offers half-star ratings, which are not so frequently used (see Figure 8) as full-star ratings. Filmeter shows the lowest average, also associated with the widest rating range (0-1). We can hypothesize that the absence of a neutral point, and the icons associated to rating positions push the users to express lower ratings (see Figure 2). Neutral icons like the stars do not seem to affect the users as much. When they interact with star-based scales, users are probably more influenced by the granularity and the presence or absence of a neutral point. On the contrary, very descriptive and representational icons, such as those of Filmeter, exert a greater influence on users' ratings.

At the same time, performing correlational studies and linear regression analyses, we have highlighted strong relationships among user ratings on different scales: such relationships tell us that, when using these scales, users express their ratings in a comparable and consistent way, so that ratings on a scale could be easily predicted, provided that user's ratings on a related scale are known. In particular, regarding experiment 1, we observed very large correlations between Criticker, MovieLens, IMDb and Filmcrave, which are all above $r>0.9$, and large correlations ($0.740 > r < 0.791$) between Filmeter (which is characterized by more influencing icons) and the remaining ones.

All these correlations were confirmed in experiment 3[31], where we also identified a large correlation between 5 and 10 stars. Interestingly, we observed that all correlated scales show similar or closely related granularity, as well as sharing other objective features. Likewise, we have observed some cases of statistical agreement between pairs of rating scales (e.g., three smileys and "-1,0,+1" scale): again in this case, the involved scales usually have similar objective features, and they always have the same or a closely related granularity. Moreover, we observed that scales with a larger granularity seem to stimulate reasoning and, therefore, the expression of

[31] Notice, however, that here we used 5-star and 4-star rating scales with no half-star ratings, differently from MovieLens and Filmcrave.

pondered ratings, while scales with lower granularity lead to extreme ratings, in accordance with some state-of-art work (Preston and Colman, 2000).

Thanks to experiment 2, we could extend our results to the museum context, and confirm that ratings given to the same item with different rating scales are significantly different. Also in this case, granularity was useful to explain the gathered data: on the one hand, ratings were closer for scales with a similar granularity, on the other hand, user ratings got higher when granularity was lower. Interestingly, this last insight differs from our results from experiment 1, where users tended to give higher ratings when granularity was higher (as in Lim, 2008). A simple way to explain these apparently contrasting behaviours is to ascribe them to domain-specific effects (museum vs. cinema). In general users liked and appreciated the high quality multimedia presentations they watched and found them informative, hence the high evaluations they got. This may explain why with the rating scales with two or three options, the ratings were very high in general, while in the rating scales with five options (stars and smileys ratings were similar), the evaluations were a bit more moderate. However, if we consider that rating scale granularity ranges 4-101 in experiment 1 and 2-5 in the experiment 2, our data may actually highlight a tendency to give higher ratings with scales that have an "extreme" granularity, namely, either very low or very high, and to be more critical when using scales with an intermediate granularity.

Considering all three empirical studies, our intuition is that granularity is the main and most clearly identifiable feature responsible for the effects that ratings scales can exert on user ratings, an insight that also confirms the assumption that granularity is one of the most important features of rating scales, in accordance with the model we adopted (Section 2) and related work (such as Lim (2008) and Daves (2002). According to our experiments, other scale features, such as the presence of a neutral point or particularly expressive icons, influence unpredictably users ratings.

### 7.1 Comparing ratings scales and their impact on recommender systems performance

Another intuition garnered from our results is that the rating scales also impact on the performance of the recommender, as already introduced by (Vaz et al., 2013). This implies that comparing, using

MAE for instance, the results of recommender systems that collected user preferences by means of different rating scales may lead to incomparable results. As highlighted by all three experiments (see Sections 4, 5, 6) some scales seem to push users to a similar, and thus comparable rating behavior, while many other scales do not. Thus, our suggestion is that MAE and RMSE can be compared only when recommender systems use the same rating scale to collect user preferences.

As far as rating conversion is concerned, our results show that a uniform, predefined mapping is not enough if we want to compensate for the effects of rating scales. Luckily, we have found that strong linear relationships exist between most pairs of rating scales that can be used for rating translation. However, the specific coefficients to use for rating conversions differed in our evaluations, thus preventing us from providing a "one-size-fits-all" recipe for rating translation. Our intuition is that, given the existence of linear relationships between two types of scales, a specific conversion formula should be derived from an analysis of sample data which can be considered representative, given the domain and audience that a certain service/website aims at targeting.

## 8. CONCLUSIONS AND FUTURE WORK

The studies reported here confirmed our past results that there are significant differences in user's average ratings on different scales. However, we also confirmed (Cosley et al. 2003)'s result that ratings consisting of different scales correlate well in most cases. This means that ratings on a scale can be predicted based on ratings on other scales, which is good, but a static, pre-defined mapping is not enough, in that it does not fully take into account the effects of rating scales on users' rating behaviour. According to our understanding, such effects are responsible for the differences we could observe in users' average ratings. When there is a strong positive correlation (according to Pearson's r coefficient), we suggest to translate ratings using linear equations derived from regression analysis. For example, this can be done when ratings are translated from a 10-point star scale to a 101-point scale where users input bare numbers. Based on our studies, however, there are

no fixed parameters for such equations, and we suggest to derive them from an analysis of real ratings expressed by the target group of users in the target domain.

As future work, we aim to conduct further studies on rating scales, in different contexts and different domains, examining the relationship to the rating scale features. This will be conducted with the goal of trying to determine a general model capable of understandding users' rating behaviour given a particular scale.

ACKNOWLEDGEMENTS

**9. We are very grateful to Fabrizio Garis and Sathya Del Piano, the two students that helped us in collecting the large amount of data we studied in the real movie ratings analysis (Section 5). We are also very grateful to Martina Deplano, who helped us in carrying out the web-based user study (Section 6).REFERENCES**

Adomavicius, G. and Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. on Knowl. and Data Eng*., 17:734–749.

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

Benedek, J. and Miner, T. (2002). Measuring desirability: New methods for evaluating desirability in a usability lab setting. In *Proceedings of Usability Professionals Association 2003*, pages 8–12.

Berkovsky S., Herder E., Lops P., Santos O. C (2013). *Late-Breaking Results, Project Papers and Workshop Proceedings of the 21st Conference on User Modeling, Adaptation, and Personalization.*, Rome, Italy, June 10-14, 2013. CEUR Workshop Proceedings 997, CEUR-WS.org

Cantador, I. Fernandez-Tobıas, S. Berkovsky, and P. Cremonesi. *Recommender Systems Handbook*, chapter Cross-Domain Recommender Systems, pages 919–959. Springer US, Boston, MA, 2015.

Cena, F. and Vernero, F. (2015). A study on preferential choices about rating scales. *International Journal of Technology and Human Interaction*, 11(1):33–54.

Cena, F., Vernero, F., and Gena, C. (2010). Towards a customization of rating scales in adaptive systems. In UMAP, pages 369–374.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). New Jersey: Lawrence Erlbaum Associates.

Cosley, D., Lam, S. K., Albert, I., Konstan, J. A., and Riedl, J. (2003). Is seeing believing?: how recommender system interfaces affect users' opinions. In *Proceedings of the SIGCHI conference on Human factors in computing systems, CHI '03*, pages 585–592, New York, NY, USA. ACM.

Cummins, R. and Gullone, E., 2000. Why we should not use 5-point Likert scales: The case for subjective quality of life

measurement. In: Second International Conference on Quality of Life in Cities, 74–93.

Dawes, J. (2002). FivePoint vs. Eleven Point Scales: Does It Make A Difference To Data Characteristics? *Australasian Journal of Market Research*, 10, 39–47.

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. *International Journal of Market Research*, 50:61–77.

Friedman, H. H. and Amoo, T. (1999). Rating the rating scales. *Journal of Marketing Management*, 9(3):114–123.

Friedman, H., Herskovitz, P., and Pollack, S. (1993). The biasing effect of scale-checking styles on response to a Likert scale. In: *Proceedings of the Joint Statistical Meeting*, 792–795.

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

Gena, C., Brogi, R., Cena, F., and Vernero, F. (2011). The impact of rating scales on user's rating behavior. In *User Modeling, Adaption and Personalization - 19th International Conference, UMAP 2011, Girona, Spain, July 11-15, 2011. Proceedings*, Lecture Notes in Computer Science 6787, pages 123–134.

Goldberg, K. Y., Roeder, T., Gupta, D., and Perkins, C. (2001). Eigentaste: A constant time collaborative filtering algorithm. *Inf. Retr.*, 4(2):133–151.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. (2004). Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53.

Holzinger, A., Scherer, R., Ziefle. M. (2011). Navigational User Interface Elements on the Left Side: Intuition of Designers or Experimental Evidence? INTERACT (2) : 162-177.

Hu, R. and Pu, P. (2013). Exploring Relations between Personality and User Rating Behaviors. *UMAP Workshops 2013)*, CEUR-WS.org.

Kaptein, M.C., Nass, C., and Markopoulos, P. (2010). Powerful and Consistent Analysis of Likert-type Rating scales. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, Atlanta, Georgia, USA New York, NY, USA: ACM, 2391–2394.

John, O. P. and Srivastava, S. (1999). The big five trait taxonomy: History, measure- ment, and theoretical perspectives. Handbook of personality: Theory and research, *Behaviour & Information Technology output 2*(1999):102–138.

Kuflik, T., Wecker, A. J., Cena, F., and Gena, C. (2012). Evaluating rating scales personality. In Masthoff, J., Mobasher, B., Desmarais, M. C., and Nkambou, R., editors, *User Modeling, Adaptation, and Personalization - 20th International Con ference, UMAP 2012*, Montreal, Canada, July 16-20, 2012. Proceedings, volume 7379 of Lecture Notes in Computer Science, pages 310–315.

Tsvi Kuflik, Alan J. Wecker, Joel Lanir, Oliviero Stock (2015). An integrative framework for extending the boundaries of the museum visit experience: linking the pre, during and post visit phases. J. of IT & Tourism 15(1): 17-47. Lim, H.e. (2008). The Use of Different Happiness Rating Scales: Bias and Comparison Problem?. *Soc Indic Res*, *87*, 259–267.

Maydeu-Olivares, A., et al. (2009). The effect of varying the number of response alternatives in rating scales: Experimental evidence from intra-individual effects. *Behavior Research Methods*, 41 (2), 295–308.

Nobarany, S., Oram, L., Rajendran, V. K., Chen, C.-H., McGrenere, J., and Munzner, T. (2012). The design space of opinion measurement interfaces: exploring recall support for rating and ranking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 2035–2044, New York, NY, USA. ACM.

Petz, G., Karpowicz, M., Fürschuß, H., Auinger, A., Stríteský, V., Holzinger, A. (2015). Reprint of: Computational approaches for mining user's opinions on the Web 2.0. Inf. Process. Manage. 51(4): 510-519.

Preston, C. and Colman, A. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta psychologica*, 104(1):1–15.

Royce, S.A. and Straits, B.C. (1999). Approaches to social research, 3rd edn. Oxford University Press, New York.

Schafer, J. B., Frankowski, D., Herlocker, J., and Sen, S. (2007). *The adaptive web. chapter Collaborative filtering recommender systems*, pages 291–324. Springer- Verlag, Berlin, Heidelberg.

Shaftel, J., Nash, B.L., and Gillmor, S.C. (2012). Effects of the Number of Response Categories on Rating Scales, *Roundtable presented at the annual conference of the American Educational Research*, https://cete.ku.edu/sites/cete.drupal.ku.edu/files/docs/Presentations/2012_04_Shaftel%20et%20al.,%20Number%20of%20Response%20Categories,%204-9-12.pdf

Sparling, E.I. and Sen, S. (2011). Rating: how difficult is it? In: B. Mobasher, R.D. Burke, D. Jannach and G. Adomavicius, eds. *RecSys ACM*, 149–156.

Stickel, C., Maier,K., Ebner,M. and Holzinger, A. (2009). The modeling of harmonious color combinations for improved usability and UX. ITI 2009: 323-328

Strauss, A. and Corbin, J. (1998). *Basics of Qualitative Research : Techniques and Procedures for Developing Grounded Theory*. SAGE Publications.

Swearingen, K. and Sinha, R. (2002). Interaction design for recommender systems. In *Proceedings of Designing Interactive Systems 2002*. ACM. Press.

Tourangeau, R., Couper, M.P. and Conrad F. (2007) Color, Labels, and Interpretive Heuristics for Response Scales. *Public Opin Q* 71 (1): 91-112

van Barneveld, J. and van Setten, M. (2004). Designing Usable Interfaces for TV Recommender Systems. In: *Personalized Digital Television. Targeting programs to individual users*. L. Ardissono, A. Kobsa and M. Maybury editors, Kluwer Academic Publishers.

Vaz, P. C., Ribeiro, R., and de Matos, D. M. (2013). Understanding the temporal dynamics of recommendations across different rating scales. In Berkovsky, S., Herder, E., Lops, P., and Santos, O. C., editors, UMAP Workshops, volume 997 of CEUR Workshop Proceedings. CEUR-WS.org.

Weijters, B., Cabooter, E., and Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, 27 (3), 236 – 247.

Weng, L.J. (2004). Impact of the Number of Response Categories and Anchor Labels on Coefficient Alpha and Test-Retest Reliability. *Educational and Psychological Measurement*, 64 (6), 956–972.

Yan, T. and Keusch, F., (2015). The Effects of the Direction of Rating Scales on Survey Responses in a Telephone Survey. *Public Opin Q* 79 (1): 145-165

**List of Figures**