I like it... I like it not: Evaluating User Ratings Noise in Recommender Systems

Xavier Amatriain, Josep M. Pujol, and Nuria Oliver

Telefonica Research

Recent growing interest in predicting and influencing consumer behavior has generated a parallel increase in research efforts on Recommender Systems. Many of the state-of-the-art Recommender Systems algorithms rely on obtaining user ratings in order to later predict unknown ratings. An underlying assumption in this approach is that the user ratings can be treated as ground truth of the user's taste. However, users are inconsistent in giving their feedback, thus introducing an unknown amount of noise that challenges the validity of this assumption.

In this paper, we tackle the problem of analyzing and characterizing the noise in user feedback through ratings of movies. We present a user study aimed at quantifying the noise in user ratings that is due to inconsistencies. We measure RMSE values that range from 0.557 to 0.8156. We also analyze how factors such as item sorting and time of rating affect this noise.

1 Introduction and Motivation

A common approach to handle digital information overload is to offer users a personalized access to information. Recommender Systems (RS), for instance, automatically suggest new content that should comply with the user's taste. In the RS literature, these predictions of user preferences are typically obtained by means of approaches such as collaborative filtering – *i.e.* taking into account other users rating history in order to model the taste of peers – or content-based – *i.e.* using existing content descriptions to uncover relations between items. Regardless of the approach, these personalized services share a common concern: modeling the user's taste. Therefore, such systems need to somehow capture likes and dislikes in order to model or infer the user's preferences.

User preferences can be captured via either *implicit* or *explicit* user feedback. In the implicit approach [12], user preferences are inferred by observing consumption patterns. However, modeling user preferences on the basis of implicit feedback has a major limitation: the underlying assumption is that the amount of time that users spend accessing a given content is directly proportional to how much they like it. Consequently, explicit feedback is the favored approach for gathering information on user preferences. Although this approach adds a burden on the users and different users might respond differently to incentives [6], it is generally accepted that explicit data is more reliable in most situations.

The preferred method for capturing explicit preference information from users consists of rating questionnaires [1], where users are asked to provide feedback – via a value point on a fixed scale – on how much they like some content. Typically, scales range from 0 or 1 to 5 or 10 and are quantized to integer values.

Approaches to inferring user preferences are evaluated on the basis of how well they can match a previously existing rating or anticipate future ones. However, little attention has been paid to how consistent users are in giving these ratings, how much input noise can be expected and how this noise can be characterized (see Section 2). The main contribution of this paper is a user study aimed at characterizing and quantifying the noise caused by user inconsistencies when providing ratings (see Section 4 for an overview of the experimental procedure and Section 5 for the results). This estimation is important because it represents a lower bound on the error of explicit feedback-based RS.

2 Related Work

The bias introduced in RS by noise in user ratings has been known for some time. Hill et al. [9] were aware of this issue and designed a small scale experiment to measure reliability in user ratings. They carried out a two trial user study with 22 participants and a time difference of 6 weeks between trials. Unfortunately, the noise in user ratings was a side issue in their overall study and they only reported pairwise correlations. Cosley et al. [4] carried out a similar experiment using a rate-rerate procedure with two trials on 212 participants. They selected 40 random movies in the center of the rating scale (*i.e.* 2,3 or 4 rating) that participants had already rated in the past – months or even years earlier, according to the authors. They reported participants being consistent only 60% of the time. In this study, the measured correlation between trials was 0.70. Herlocker et al. [8] discuss the noise in user ratings in their review of evaluating methods for RS. In particular, they introduce the concept of the “magic barrier” that is created by natural variability in ratings. The authors also highlight the importance of analyzing and discovering this inherent variability in recommender data sets and include it as a future line of work.

Mahony et al. [13] classify noise in RS into *natural* and *malicious*. The former refers to the definition of user generated noise provided in this paper, while the latter refers to noise that is deliberately introduced in a system in order to bias the results. Even though the focus of their work is on *malicious* noise, they do propose a de-noising algorithm that can be used to detect *natural* noise. Their baseline recommender algorithm reported a marginal improvement on a reduced data set once the ratings labeled as noise by the de-noising method are discarded.

To the best of our knowledge, the former are the only pieces of work in the literature on RS that explicitly address the problem of inconsistencies in user ratings. The work presented in this paper provides a more detailed study and in-depth analysis with the aim of characterizing the noise due to inconsistencies in user ratings.

3 Measures of Reliability in User Tests

Our effort to analyze and characterize noise and inconsistencies in user ratings is related to the concept of *reliability* of user tests from classical test theory. Reliability in this context is defined as the ratio of true score variance over the observed score variance. This ratio is used as a signal-to-noise measure of a given user test. Since true scores are unknown, it is not possible to compute reliability directly. However, there are methods to estimate it [10].

Of particular interest to us is the so-called *test-retest reliability*. This measure is often used in psychometry to quantify how reliable a particular "instrument" (*e.g.* survey or test) is [15]. The test-retest reliability is a function of the Pearson correlations between the different trials of the same test. However, it is not sufficient to compute the correlation between two different trials of the same test. As Heise explains [7], the correlation is aggregating two effects: the instrument's reliability and the stability of the user's judgements. That is, if we measure how much a user likes an item at two different times (separated by a month, for instance) and find a different rating, this could be due to either the reliability of the measure and the user's response or to the fact that the user's opinion has changed during that period. Therefore, three points in time are needed in order to distinguish between both effects. Once these are available, pairwise correlations r_{12}, r_{23}, and r_{13} can be computed to obtain (a) the overall reliability (Eq.1), and (b) the stability in users' opinions from time x to time y, (s_{xy}) (Eq. 2).

$$\mathbf{r_{xx}} = r_{12}r_{23}/r_{13} \tag{1}$$

$$\mathbf{s_{12}} = r_{13}/r_{23}; \quad \mathbf{s_{23}} = r_{13}/r_{12}; \quad \mathbf{s_{13}} = {r_{13}}^2/r_{12}r_{23} \tag{2}$$

Note that neither of the related surveys reviewed in the previous section [9] [4] take into account the reliability and stability of their studies. This is especially problematic in the case of Cosle's *et al.* experiment where ratings might be separated by months.

4 Experimental Setup

The research questions that we wanted to address with our experiment are: *Q1:* Are users inconsistent when providing ratings? *Q2:* If so, how large is the error due to such inconsistencies? *Q3:* What are the factors that have an impact on user inconsistencies?

Apparatus and Procedure We selected 100 movie titles from the Netflix Prize database [2]. The selection was done by using a stratified random sample on the movie popularity curve. We divided the 500000 movies in the database into 10 equal-density bins and random sampled 10 movies out of each bin – only 100 movies were selected in order to avoid user churn. By using this procedure, we obtained a sample that included a significant portion of unpopular movies that ensured an appropriate spread of the results.

Our experiment consisted of 3 trials (R_1, R_2, and R_3) of the same task: rating 100 movies via a Web interface. The three trials took place at different points in time, in order to assess the reliability of the user rating paradigm and to measure the variability of users. The minimum time difference between trials was set to 24 hours for the first and second and 15 days for the second and third. Users could stop and resume the trial at the same spot at any time.

User ratings were provided on a 1 to 5 star scale with a special crossed-out eye icon located on the left to indicate unseen movies. Information about the movie included title, year, director, cast and DVD cover. Users could follow a link to *IMDB* [1] if they needed further information.

[1] http://www.imdb.com.

We designed a two part test-retest experiment in order to discern the test reliability from the user's stability. In addition, we wanted to analyze whether the elapsed time between ratings and the order in which items were presented had any influence in the consistency of the participants' answers.

Participants were presented with movie titles in a predetermined sequential order so that the effect of the order of the responses could also be analyzed. Previous research has shown that sequential user tests generate what is known as the *assimilation/contrast effect* [5, 14]: a user is likely to give a lower rating to an item if the preceding one deserved a very high evaluation. However, if successive items are comparable in their ratings, the user is likely to assimilate the second item to the preceding one and give the same rating to both. In addition, and especially in the case of the first and second trials, we wanted to rule out the effect of any possible sequential memory effect (*i.e.* remembering the ratings from the previous trial and therefore not paying enough attention the next time). For these reasons, two different permutations of the movies were created: permutation 1 (used in trials 1 and 3) was a random order; and permutation 2 (used trial 2) ordered movies according to their popularity in Netflix.

One possible concern in our experiment design was the short elapsed time between our trials. Another concern was that the different order introduced in trial 2 could be introducing a hard-to-isolate confound. To address these issues, we ran a fourth trial, R_4, with a subset of our population (36 users) seven months after our original survey. The results are reported separately in section 5.4 as a final support to our hypothesis.

Participants Participants were recruited via email advertisement in a large telecommunications company. A total of 118 distinct users completed the three trials in the study. The participants' age ranged from 22 to 47 years, with an average age of 31.2 years. Almost 90% of our participants were in the 22 to 37 age group and most of them were male (79.12%). This demographic group corresponds to the most active group in online applications such as RS [3].

Additionally, we collected data about their familiarity with the movie domain. Participants reported watching an average of 1.55 movies in the cinema, 3.8 TV movies, and 5.13 DVD movies per month. When asked about their familiarity with online rating systems, participants were somewhat unfamiliar with them (mean: 2.60 on a 5 point Likert scale). Finally, when asked about Web usage familiarity, our participants considered themselves to be proficient users, with an average of 4.74 on a 5 point Likert scale.

5 Results

In this section, we first compare the ratings obtained in our survey with the Netflix ratings for the same movies. We then present our results by evaluating the test-retest reliability of the experiment as well as user stabilities. Finally, we analyze three variables that might play a role in determining user inconsistencies: (a) the rating scale, (b) the order in which the movies were presented; and (c) the moment of time when movies were rated[2].

Comparison to Netflix The Netflix dataset is one of the most popular benchmarks in the RS community. Therefore and before further analysis, we compare the behavior of the participants in our experiment to that of Netflix' users. First, we compare the ratings obtained in our survey with those in Neflix. Figure 1 depicts the rating distribution of the three trials of the experiment, when compared to the Netflix ratings on the *same* 100 movies. Note how similar both rating distributions are. The main difference is that the Netflix data set distribution has a *higher* mean (*i.e.* Netflix users tend to rate the 100 movies with higher scores than the participants in our study). This observation might be due to several factors: Our experiment, as opposed to Netflix, asked users to rate movies that they did not explicitly choose to rate. In addition, our movie sample is biased towards non-popular movies, which in a different setting most users would have not rated. Finally, there might also be an effect of our biased demographics.

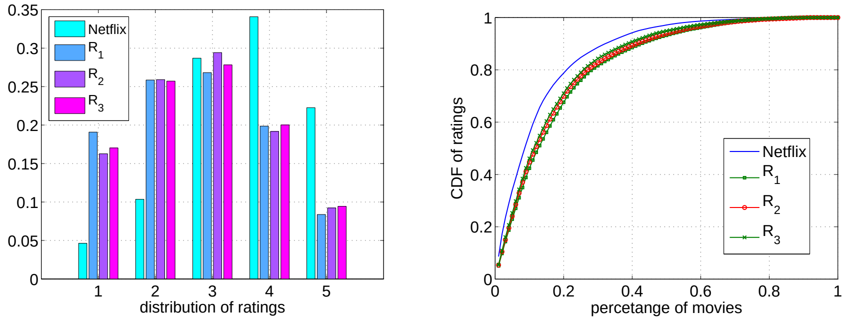


Fig. 1: User study data compared to Netflix. (a) Rating distribution in the 3 trials of our survey as compared to the Netflix data set. And, (b) Cumulative distribution of number of ratings by movie.

Next, we are interested in assessing whether our experiment design – *i.e.* having users rate movies in a batch – might be different enough from a real setting that would bias the results. In our experiment, we measure an average of 18.5 ratings per user in the worst case (first trial). If we analyze the Netflix dataset, we measure an average of 5.8 ratings per day (session). However, when we remove sessions with less than 4 ratings from the Netflix dataset, we measure an average of 20 movies per session, larger than in our study. Note that sessions not removed in this case (*i.e.* those with 4 or more consecutive ratings) account for 79.67% of the ratings in the Netflix dataset. Therefore, our experimental setting seems to be representative of high proportions of the Netflix dataset (and hence of similar real-life settings).

[2] In addition, and in order to rule-out a possible effect of our movie selection procedure, we computed all values for the 20% most popular movies, observing no significant difference.

5.1 Test-retest Reliability and Stability

In order to compute the reliability of our test, we first compute the correlation coefficients between different trials, which result in $r_{12} = 0.8986$, $r_{23} = 0.9028$, and $r_{13} = 0.8783$. From these values and using Eq. 1, the overall reliability of our experiment is $r_{overall} = 0.924$. As a first conclusion, we observe that our test has high overall reliability – any value over 0.9 is usually considered "good" in classical test theory [11]. This result validates the procedure of asking users for their ratings – in the context of Web-based movie rating – as a good measure of whether they like/dislike **these** particular movies. A different question, that we will address later in our analysis, is whether this procedure is a good way to quantify user preferences. The overall reliability also sets an upper bound for a predictive algorithm based on this explicit user feedback.

Using Eq. 2, we compute the temporal pairwise stabilities to be: $s_{12} = 0.973$, $s_{23} = 0.977$, and $s_{13} = 0.951$. These stability factors are all high as well. This should be expected given the short times elapsed between trials: user preferences are not likely to change in two weeks. Also as expected, the lowest stability coefficient (s_{13}) corresponds to the longest time interval between trials (at least 15 days between trials 1 and 3). However, it comes as a surprise that the stability between trials 1 and 2 (at least 1 day apart) is slightly lower than the that between trials 2 and 3 (at least 15 days). Note that the stability coefficient might also be accounting for the user's "learning effect". Such intuition is supported by the fact that the stability effect between trials 1 and 2 is not closer to 1.0 – it is hard to imagine that the users opinions have changed in about 24 hours. The lower values in s_{13} could in fact be accounting for both change in opinion and a learning effect. We leave this issue to future work.

These inter-test correlations are the only measures that can be compared to the works of Hill et al. [9] and Cosley et al. [4], with reported correlations of 0.83 and 0.70 respectively (see Section 2). However, their measures include the effect of both reliability and stability.

Additionally, we are interested in measuring the impact that a given rating value has on the overall reliability. Therefore, we compute new reliability values by ignoring all triplets of ratings where at least one rating equals the value to remove. Removing ratings 2, 4, and especially 3, improves the reliability, yielding new values of 0.93, 0.925 and 0.95, respectively – as compared to the overall reliability of 0.924. On the other hand, removing extreme ratings (1 and 5) yields lower reliability – 0.88 and 0.89, respectively. This finding seems to indicate that recommender algorithms could benefit from giving lower weight or importance to ratings in the middle of the rating scale.

5.2 Analysis of Users Inconsistencies

Next, we shall study the inconsistencies of user ratings across different trials. Table 1 summarizes the results of the experiment when grouping the trials by pairs, where R_k corresponds to trial $k = 1, ..., 3$.

Let us define the aggregated rating of user u's ratings of movie m as a tuple $\langle r_k \rangle_{um}$, where r_k corresponds to the rating at trial R_k. Therefore, for a given user u and movie m we have vector of three ratings $\langle r_{um1} r_{um2} r_{um3} \rangle$, Note that there are user × movies tuples (*i.e.* $118 \times 100 = 11800$ in our case). A rating is considered to be *consistent* across trials, when all values of r_k are the same. Note that we are not interested in those tuples where all r_k are zeros, which is the value used to represent a *not-seen*.

Effect of "not seen" values In order to analyze the effect that the "not seen" value has in our study, we consider two different subsets: a) the *intersection* or only tuples where all ratings are *seen* (> 0) and b) the *union*, where not seen values are included. For instance, ratings $\langle 4, 4, 5\rangle_{um}$ would be inconsistent, because user u changed her evaluation of movie m from 4 to 5 in the last trial. This tuple, however, would be included both in the intersection and the union set. However, the tuple $\langle 4, 4, 0\rangle_{um}$ would not be included in the intersection set, because one of the ratings is a *not-seen*.

	$\#R_i$	$\#R_j$	#		$RMSE$	
			$\cap$	$\cup$	$\cap$	$\cup$
R_1, R_2	2185	1961	1838	2308	0.573	0.707
R_1, R_3	2185	1909	1774	2320	0.637	0.765
R_2, R_3	1969	1909	1730	2140	0.557	0.694

Table 1: Summary of results on the pairwise comparison between trials. The first and second column contain the number of ratings in trials R_i and R_j. The third and forth column depict the number of elements in the intersection and the union for R_i and R_j. The intersection set contains ratings in which no element is *not-seen*, whereas the union set allows for *not-seen* elements. The last two columns report the root square mean error of the intersection and the union sets.

Table 1 summarizes the users' inconsistency results. For example, in R_1, users provide 2185 out of the potential 11800 ratings. Thus, 9615 positions in the rating matrix of R_1 are *not-seen* values. Without taking the actual value of the rating into consideration, the divergence in the number of ratings illustrates how users are not even able to consistently determine whether they have seen a movie or not. Only 1838 ratings in R_1 also appear in R_2 – the intersection. If we take the union, we obtain 2308 ratings. The results are similar on all pairs of trials. With these results, we are able to answer our first research question *Q1*.

RMSE due to inconsistencies We shall now look at the inconsistencies due to a *different rating value* in different trials. We use the *root mean squared error* (RMSE) for easy comparison with previous and related work in the RS literature and in particular with the Netflix Prize threshold (*i.e.* desired RMSE of 0.8563) [2]. The right side of Table 1 contains the RMSE for the intersection and union sets across all trials.

The RMSE for the intersection sets ranges between 0.55 and 0.63, depending on the trials. Note that the previously computed stability is inversely correlated with the RMSE. The most stable comparison is between R_2 and R_3, 0.977, which gives the smallest RMSE (0.5571).

In the case of the union sets, we replace the *not-seen* value with the average rating for that movie. The RMSE is now higher as it is accounting for two types of user inconsistencies: inconsistencies in labeling as *seen or not-seen* and inconsistencies in the actual values. The RMSE ranges from 0.694 to 0.765 in this case.

Note that these values of RMSE represent a lower bound of the RMSE that could be achieved by a RS built from the data in our study. Therefore, and in the context of our study, current RS algorithms would not be able to predict the movie ratings with lower RMSE that the ones described in Table 1 (unless they are overfitting the training data). Of course, the particular RMSE values are dataset dependent. With this analysis, we address our second research question *Q2*.

5.3 Variables that have an Impact on User Inconsistencies

In order to answer our third research question (*Q3*), we analyze the variables that might play a role in increasing the likelihood of user inconsistencies. In particular, we explore the impact that the rating scale, item order and user input speed might have on inconsistencies.

Rating Scale Effect In the initial reliability analysis presented in Section 5.1, we showed that removing 2 and 3 star ratings yields higher reliability. We shall now investigate this further by analyzing which are the most common inconsistencies. Figure 2a shows the probability of inconsistency by the value of the rating between pairwise trials (R_1,R_2), (R_2,R_3) and (R_1,R_3). In other words, the probability that if users gave a rating of X in trial R_i, they will give a different rating in trial R_j.

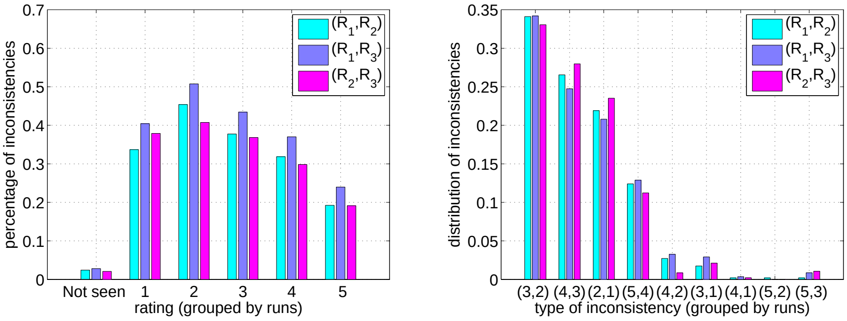


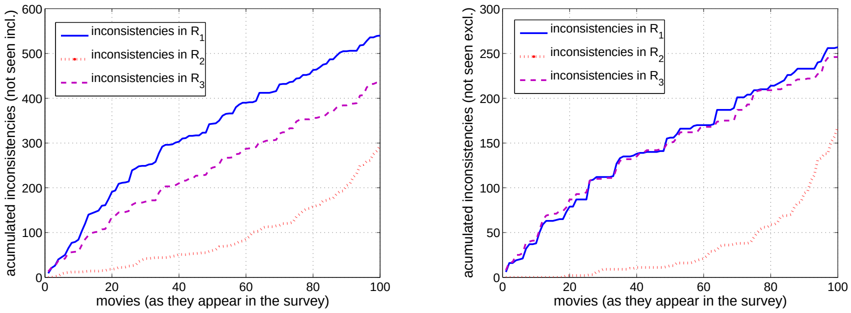
Fig. 2: Users Inconsistencies. (a) Percentage of inconsistencies by rating value and (b) Distribution of types of inconsistencies

Note how ratings with extreme opinions (*i.e.* the lowest and highest ratings in the scale) are more consistent across different trials: the probability of inconsistencies is highest for 2 and 3 stars ratings. The average ratings in our study are 2.73, 2.79 and 2.79 for R_1, R_2 and R_3 respectively. Also note that the probability of inconsistency with *not-seen* is lower.

We shall investigate next what are the most common inconsistencies. Figure 2b depicts the distribution of inconsistencies by switching the score – note that the Figure does not include inconsistencies due to *not-seen* items. The two most common inconsistencies are due to a rating drifting between 2 and 3 (about 34%) and between 3 and 4 (25%). Ratings with a ±1 drift account for more than 90% of the inconsistencies.

Thus, ratings in the middle of the rating scale seem to be more prone to inconsistencies than extreme ratings. This observation makes intuitive sense for several reasons: First, extreme ratings have a lower or higher bound (*e.g.* you cannot get higher than 5). Also, users are probably more consistent about remembering very good and very bad movies, which somehow impacted them. Finally, extreme ratings seem to be less prone to assimilation and contrast effects. These intuitions, however should be further investigated in future work.

Item Order Effect Next, we shall analyze the effect of time on user inconsistencies. Figure 3 depicts the inconsistencies as they appeared over time while participants filled out each of the surveys. Note that now inconsistencies are not computed by pairwise comparisons across trials, but reckoned across the three trials. In our analysis, we compute the *ground truth or valid* rating for each movie and participant as the rating that appears *at least twice* across the three trials. Thus, we assume that the trial with the different value is the one causing the inconsistency. Note that movies where the three ratings for the three trials are different from each other are discarded (they represent a 10.69% of the total).



(a) Taking into account "not seen" values.

(b) "Not seen" values not taken into account.

Fig. 3: Accumulated error across movies. An error is assigned to R_i if its rating is different than the other R. The movies are set as they appear in R_0 and R_3.

Figure 3a shows the accumulated inconsistencies over time as movies were presented to the user, including inconsistencies due to *not-seen*. Figure 3b excludes the *not-seen* inconsistencies.

As Figure 3a illustrates, the first trial R_1 is responsible for most of the inconsistencies, followed by the third trial R_3. The decrease of inconsistencies in the last trial R_3 might be caused by the learning effect, as users would have undergone the survey twice before. However, when discarding the effect of the *not-seen* value (Fig. 3b), R_1 and R_3 exhibit a very similar behavior. This result suggests that a learning effect might only affect the consistency on discriminating between *seen* and *not-seen* movies.

Interestingly, the second trial R_2, which took place at least one day after R_1 and where the movies were sorted by increasing popularity, displays the lowest level of inconsistencies. The improvement in consistency in R_2 might be explained by several factors: First, the short time between trials – only 24 hours. However, neither the pairwise stability nor the RSME support this hypothesis. Therefore, it seems that the *order* in which the movies are presented (*i.e.* showing popular movies first) could be the factor for the consistency gain. Additionally, this result might be related to the minimization of the *contrast effect*, as similar movies are shown together.

To sum up and according to our experiment, a rating interface that groups movies that are likely to receive similar ratings should help minimize user inconsistencies.

User Rating Speed Effect The data logs collected in the user study include the exact time at which each user rating was generated. This allows us to analyze how the speed with which users rate movies might affect their consistency.

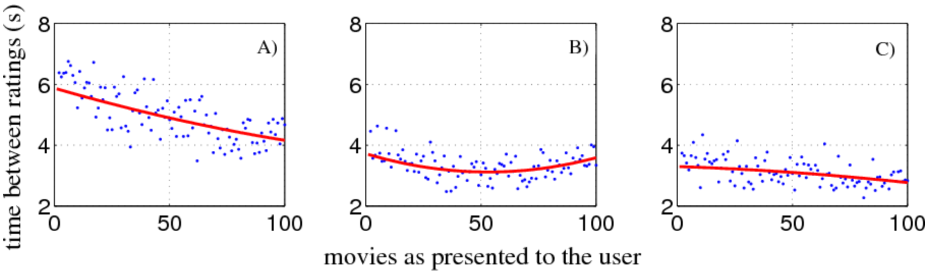


Fig. 4: Graphs depicting time between ratings for (a) R_1, (b) R_2, and (c) R_3. Note that all plots have the same temporal scale. The clicking time is always between 2 and 8 seconds. The average clicking time is 4.93, 3.30, 3.08 seconds for R_1, R_2 and R_3 respectively. For reference, a quadratic fit is also plotted as a line.

Figure 4 depicts the average evaluation time by movies where movies are sorted as they were presented to the user. Note how in the case of R_1 and R_3 (sorted at random), the evaluation time decreases as the survey progresses. This result makes intuitive sense, as users were probably getting tired or used to the setting. However, in the case of R_2 (Fig. 4.b), the evaluation time decreases at first, but then increases again during the last half of the survey. This behavior might be caused by the way the movies in R_2 were presented: users were fast in assessing unpopular movies, many of which they might not have seen, at the beginning of the survey. Then, when popular movies appear (and therefore probably seen by participants), users seem to spend more time thinking about the rating.

We measure an average rating time of 4.93, 3.30, and 3.08 seconds respectively for each of our trials. One might expect that faster clicking could introduce more inconsistencies due to input error. However, the percentage of inconsistencies per trial are 42.5%, 23.2%, and 32.3%. So, a shorter time between ratings does not imply more inconsistencies on the ratings.

5.4 Long-term Errors and Reliability

In this section, we measure the reliability and RMSE of our experiment when removing the original R_2 trial and adding a new one (R_4). This new trial was conducted 7 months after $R3$, and using the same random movie permutation as R_1 and R_3. Therefore, we now have three trials with the same movie order, separated 15 days and 7 months respectively. Our goal is to evaluate if there are significant differences in the values because of the longer elapsed time and the removal of the different sorting in the intermediate trial.

First, and in order to rule out the effects of this smaller – and maybe biased – population, we recomputed the correlations, stability factors, reliability, and RMSE in the three original trials for this subset of 36 users, observing no significant differences with the original values reported for the entire population.

Using this new setting, we obtain an overall reliability of 0.8763 – compared to the original 0.924. Although this is only a 5% difference, we are now below the 0.9 threshold. This is an indication that this kind of rating surveys might not be an appropriate way to measure user preferences over a long period of time. Our new stability factors are measured as $s_{13} = 1.0025$, $s_{34} = 0.9706$, and $s_{14} = 0.9730$. Now, and as it would be expected, we see a much clearer trend: very high stability between the trials separated 15 days and significantly lower for any two trials separated by 7 months.

Finally, we measure our new RMSE values as $R_{13} = 0.6143$, $R_{14} = 0.6822$, and $R_{34} = 0.6835$ for the intersection, and $R_{13} = 0.7445$, $R_{14} = 0.8156$, $R_{34} = 0.8014$ for the union. First, we observe that the RMSE for trials separated by 7 months, is significantly larger than in the original setting (see Table 1, columns 6 and 7). In the original setting, we also measured lower values between consecutive trials, arguably due to the memory effect. However, when the ellapsed time between consecutive trials is long enough (*e.g.* 7 months), this effect is no longer noticeable and the RMSE is larger for sessions separated a long time, regardless of whether they are consecutive or not. Note that if we want to measure the effect of both the long time interval plus a change in movie ordering, we can compute R_{24} – error between trial 2, sorted by popularity, and trial 4 with random order and conducted 7 monhts after. The measured RMSE is now 0.832.

6 Conclusions

In this paper, we have presented a user study aimed at quantitatively analyzing user inconsistencies in a movie rating domain. Since recommender systems commonly rely on user ratings to compute their predictions, inconsistencies in these ratings will have an impact on the quality of the recommendations. We believe that the characterization of these inconsistencies is of key importance in the RS field.

Our study shows that, although the reliability of the survey as an instrument and the stability of user opinions are high, inconsistencies negatively impact the quality of the predictions that would be given by a RS. The calculated RMSE between different trials ranged between 0.557 and 0.8156, depending on the ellapsed time and whether the “not seen” ratings effect is ruled out. These RMSE values represent a lower bound (*magic barrier*) for any explicit feedback-based RS built from the data of our study unless overfitting to this data. We plan on carrying out additional studies in order to understand how well our results generalize to other domains and settings. It is interesting to note how close these values are to current state-of-the-art recommendation algorithms.

We have also presented a detailed analysis on the nature of user inconsistencies. Our main findings can be summarized as follows: (1) Extreme ratings are more consistent than mild opinions; (2) users are more consistent when movies with similar ratings are grouped together; (3) the learning effect on the setting improves the user's assessment on whether she has seen the movie, but not the stability of the rating itself; and (4) faster user clicking does not yield more inconsistencies.

We believe that these insights will benefit the design of RS, which could take this characteristic distribution of inconsistencies into consideration. Future work should validate how much our findings can be generalized across settings, datasets and domains. In addition, we plan on using the information gathered in this study to analyze how different recommendation algorithms behave to this type of noise and design strategies to overcome it.

References

1. G. Adomavicius and A. Tuzhilin. 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(6):734–749, 2005.
2. J. Bennet and S. Lanning. The netflix prize. In *Proc. of KDD Work. on Large-scale Rec.. Sys.*, 2007.
3. Choicestream. Personalization Survey. Technical report, Choicestream Inc., 2007.
4. D. Cosley, S. K. Lam, I. Albert, J. A. Konstan, and J. Riedl. Is seeing believing?: how recommender system interfaces affect users' opinions. In *Proc. of CHI '03*, 2003.
5. A. Dijksterhuis, R. Spears, and V. Lepinasse. Reflecting and deflecting stereotypes: Assimilation and contrast in impression formation and automatic behavior. *J. of Exp. Social Psych.*, 37:286–299, 2001.
6. M. Harper, X. Li, Y. Chen, and J. Konstan. An economic model of user rating in an online recommender system. In *Proc. of UM 05*, 2005.
7. D. Heise. Separating reliability and stability in test-retest correlation. *Amer. Sociol. Rev.*, 34(1):93–101, 1969.
8. J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. on Inf. Syst.*, 22(1):5–53, 2004.
9. W. Hill, L. Stead, M. Rosenstein, and G. Furnas. Recommending and evaluating choices in a virtual community of use. In *Proc. of CHI '95*, 1995.
10. F. M. Lord and M. R. Novick. *Statistical theories of mental test scores*. Addison-Welsley, 1968.
11. K. Murphy and C. Davidshofer. *Psychological testing: Principles and applications (4th edition)*. Addison-Welsley, 1996.
12. D. W. Oard and J. Kim. Implicit feedback for recommender systems. In *AAAI Works. on Rec. Sys.*, 1998.
13. M. P. O'Mahony. Detecting noise in recommender system databases. In *Proc. of IUI'06*, 2006.
14. M. Sherif and C. I. Hovland. *Social judgment: Assimilation and contrast effects in communication and attitude change*. Yale University Press, 1961.
15. G. Torkzadeh and W. J. Doll. The test-retest reliability of user involvement instruments. *Inf. Manag.*, 26(1):21–31, 1994.