# Rating: How Difficult is It?

E. Isaac Sparling
Socialcast
San Francisco, California
isaac.sparling@gmail.com

Shilad Sen
Math, Stats, and Computer Science Dept.
Macalester College
St. Paul, Minnesota
ssen@macalester.edu

## ABSTRACT

Netflix.com uses star ratings, Digg.com uses up/down votes and Facebook uses a "like" but not a "dislike" button. Despite the popularity and diversity of these rating scales, research offers little guidance for designers choosing between them.

This paper compares four different rating scales: unary ("like it"), binary (thumbs up / thumbs down), five-star, and a 100-point slider. Our analysis draws upon 12,847 movie and product review ratings collected from 348 users through an online survey. We a) measure the time and cognitive load required by each scale, b) study how rating time varies with the rating value assigned by a user, and c) survey users' satisfaction with each scale.

Overall, users work harder with more granular rating scales, but these effects are moderated by item domain (product reviews or movies). Given a particular scale, users rating times vary significantly for items they like and dislike. Our findings about users' rating effort and satisfaction suggest guidelines for designers choosing between rating scales.

## Categories and Subject Descriptors

H.5.2 [**Information interfaces and presentation**]: User Interfaces

## General Terms

Experimentation, Human Factors, Design, Measurement

## Keywords

rating scales, user studies, recommender systems

## 1. INTRODUCTION

Ratings help users explore huge information repositories. Digg features articles that users rate positively ("digged") but not negatively ("buried"). Facebook filters news feeds by analyzing what is "liked." Netflix recommends movies by analyzing one-to-five star ratings. Although ratings power each of these sites, each uses a different *rating scale.* How should a system designer choose the right scale?


*RecSys'11,* October 23–27, 2011, Chicago, Illinois, USA.


Researchers explain users' rating motivations economically [8]. Users pay a *cost* for each rating in the form of mental effort or time. Users *benefit* from the same rating; they may have fun, receive more accurate recommendations, or unlock a new feature of a website. According to the economic paradigm, users continue to provide ratings as long as they perceive that the benefits of a rating outweighs its costs.

Researchers evaluating rating scales have found evidence that finer-grained scales offer benefits to users. Users prefer them in some applications, and they may improve the accuracy of recommender systems [6]. However, little is known about the costs associated with different rating scales. This paper reports on an online user survey that measures the costs of popular rating scales. 348 survey participants rated items in two domains (movies and product reviews) using four rating scales: unary "like it," binary thumbs up / thumbs down, five star, and 100-point slider.

We frame our analysis using four research questions. Three explore costs and one measures benefits across different rating scales. First, we measure rating time:

*RQ1: Do different rating scales require different amounts of time?*

Time captures one element of mental effort, but even if two scales require the same amount of time, one may require users to "think harder." We measure users' cognitive load, which captures the "intensity of mental effort" [12] at some instant:

*RQ2: Do different rating scales elicit different levels of cognitive load?*

A rating system may also perform a cost-benefit analysis to choose whether or not to ask a user should to rate a particular item [1]. For example, recommender systems researchers have developed *item-selection algorithms* that carefully choose an item to rate to provide the greatest possible benefit to the system and user [15, 4, 2]. These algorithms assume the costs associated with rating a particular item are constant, but this may not the case. In particular, it may take a user longer to rate items they assign a high (or low) rating to. Item selection algorithms such as [4] that consider the possible ratings a user may assign may benefit from more accurate cost models. Therefore, we study:

*RQ3: Do different rating values require different amounts of time?*

Finally, in addition to measuring the mental costs of rating, we explore one benefit users derive from rating scales:

*RQ4: Do different rating scales elicit different levels of user satisfaction?*

The rest of this paper is laid out as follows. In Section 2 we

describe the characteristics of different types of items that are important to our study and discuss related work from psychologists and computer scientists. Section 3 describes the design of our online study. Section 4 analyzes and discusses each of the research questions in turn. Section 5 summarizes our findings and discusses their implications. Our findings about the costs and benefits of ratings suggest guidelines for designers choosing between rating scales and they provide insights into the psychology of how users rate.

# 2. BACKGROUND

## 2.1 Important item characteristics

Different rating systems ask users to rate different types of items: movies, music, product reviews, jokes, etc. The specific characteristics of these *item domains* may suggest a particular scale. We considered three characteristics closely related to the rating process when choosing the item domains we studied (movies and product reviews):

- *Experiential versus remembered*: Users may *experience* an item either on or off the page. Users recall their attitude towards a book they read in the past, but they read a book review on Amazon itself. If users spend a great deal of time experiencing an online item on a web page, the additional time required to actually rate the item may be negligible.
- *Rating distribution*: The overall distribution of rating values changes across domains. As evidence of this, YouTube recently switched from a five-star scale to a thumbs up / thumbs down scale because rating values of two, three, and four star values only made up 5% of ratings.[1]
- *Agreement*: Users may agree on their ratings for items in some domains more than others. If users generally agree on an item, less precision may be required to accurately assess the community's sentiment.

As described under the "Methodology" section, we select item domains that span these domain dimensions. Many other application characteristics may influence users' rating behavior such as the placement of the scale and the application powered by the ratings. We leave analysis of these characteristics as future work.

## 2.2 Related work

**Cognitive load:** Researchers identify two types of cognitive load related our work: *intrinsic* load inherent to the task being performed, and *extraneous* load resulting from suboptimal task design [12] (a third type of cognitive load – germane load – is less clearly related to our study). In ratings, intrinsic load primarily results from the cognitive process of evaluating the item being rated, while extraneous load may result from a rating scale that is difficult to use. Cognitive load, as defined by psychologists, typically represents an instantaneous measure of mental effort. We measure both instantaneous cognitive load and total rating time to capture total mental effort.

Studies often use the *dual-task* paradigm when measuring cognitive load on websites [7] [5]. In the dual-task setup, a user's cognitive load is estimated by measuring their performance on a secondary task. For example, Brunken et al. asks online learners to click a letter when it changes color, and measure users' response times [5]. Our secondary task design is modeled on Brunken's and is described in "Methodology."

[1]http://youtube-global.blogspot.com/2009/09/five-stars-dominate-ratings.html


**Figure 1: The four rating scales we study in this paper: unary, binary (thumbs up / thumbs down), five-star, and slider.**

**Psychology of rating scales:** Psychologists have carefully compared different rating scales as survey response tools. In a meta-analysis, Churchill and Peter survey 108 different publications and find that rate reliability generally increases with rating scale granularity [10]. In more recent work, Preston and Colman study users' responses on scales ranging from two to 101 points [14]. They find that users are happiest and most consistent with a 5 to 10 point scale, but users state that the 101 point scale is best at expressing their feelings. Our work extends this research to the online setting, and it explores cost of rating in terms of mental effort.

**Rating on the web:** Researchers have recently extended decades of psychology research about rating scales to the Internet. Harper et al. construct an economic model explaining rating motivations [8]. They model ratings as a cost / benefit tradeoff where the users' benefits of rating are improved prediction quality, fun, and keeping track of movies, and the cost is time. Our work extends this work by experimentally measuring several costs (mental effort) and benefits (fun as measured by user satisfaction) of different rating scales.

In work closely related to ours, Cosley et al. investigate users' satisfaction, rating consistency, and recommendation accuracy when rating movies under three different different scales: a binary scale, a $\pm 3$ scale with no zero, and a five-star rating scale with half-star increments [6]. Users like the five star scale best, and they find evidence suggesting that as scale granularity increases, recommendation accuracy increases. Their work motivates our investigation of mental effort. We study whether the increase in recommendation accuracy comes at the price of greater mental effort. Furthermore, we extend their findings on user satisfaction to two new scales (unary and slider) and two item domains (movies and product reviews).

# 3. METHODOLOGY

Users accessed the survey through a publicly available website. We recruited users through a snowball strategy [13] via public advertisements on four channels: email, the MovieLens movie recommender website[2], Twitter, and Facebook. All timing data was recorded on the user's browser and logged on the server to mitigate network latency. In this section, we describe the details of the survey.

## 3.1 Scales

We study four scales capturing the variety of choices commonly used on the Internet: the unary "Like it" scale, the binary thumbs up/down scale, the five-star scale, and the 100-point slider scale (Figure 1). The unary scale has been popularized by Facebook, where users can note an interest in a particular wall-post or photo by clicking a "Like" button. The binary scale has been used widely in many social-news aggregators like Digg.com and on YouTube. The five star scale is used in many different situations, including recommender systems such as Netflix. We also included the 100-point slider to test the extreme end of the granularity spectrum. Users manipulated the slider either by typing a value into the box,

[2]http://movielens.org

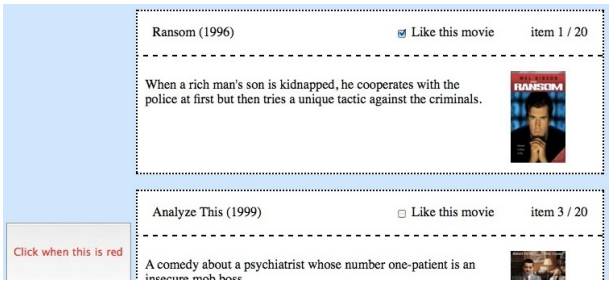


**Figure 2: Rating movies in the survey. The secondary stimulus in the lower left has changed color and begun growing. Users would normally see many items at a time (image is truncated to save space).**

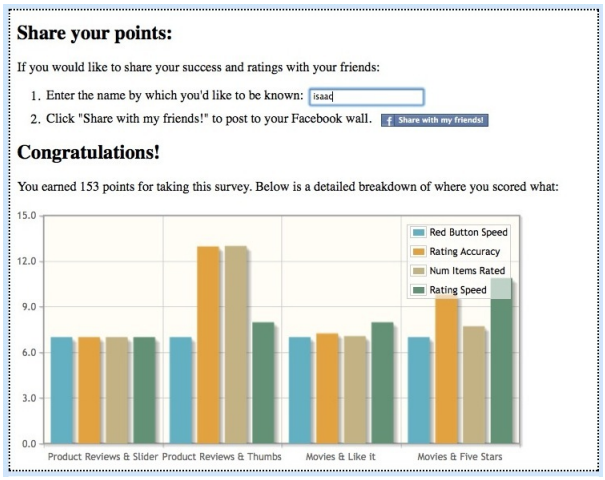


**Figure 3: Final point tabulation and prompt to share via Facebook.**


**Figure 4: Notification of survey participation posted to Facebook for users who chose to share their results.**

or by dragging the handle. All scales exhibited the conventional javascript mouse-over and mouse-click effects.

## 3.2 Item Selection

We asked users to rate movies from the Internet Movie Database, and product reviews from Amazon.com. The survey stressed that users should rate product review helpfulness, not the products themselves. We selected these two domains because they differed in a variety of dimensions we described previously. First, users experience product reviews by reading them on the page, but recall their past experience watching a movie. Second, users generally agree more often about product reviews than movies. Lui et al. found that human coders agreed on a four choice helpfulness rating of a product review 87% of the time [11], but based on our analysis of the MovieLens 1M dataset [9], raters of movies agree on a rating for a movie 34% of the time.

For each domain we selected *popular* items that people would recognize, and items with average ratings across the spectrum of possible values. For product reviews, we chose reviews from 3 relatively popular devices – a Sony HDTV, a third generation Apple iPod Touch and a Cuisinart blender. We selected a uniform sample of reviews across the range of helpfulness ratings. For movies, we chose imdb.com's Top 500 (All Time USA Box Office) list because of the popularity of the movies and relative diversity of average ratings. We sampled movies and product reviews whose average rating spanned the spectrum of rating values.

## 3.3 Survey Overview

The survey consisted of three main parts: an introductory set of instructions, a series of four sets of pages asking users to rate items (the bulk of the survey), and finally a followup questionnaire, where users reflected on their experience with the rating scales. The first section of the survey described the overall purpose of the study and asked users to provide background information: their email address, age, sex, and level of internet use. In the middle section of the survey, users completed four randomized experimental *treatments* corresponding to each of the four rating scales. Within a single treatment users either rated movies or products (two treatments each).

Every treatment in the middle section of the survey consisted of four pages. The first page described how users should use the rating scale associated with the treatment. The second page asked the user to rate 20 movies or 7 product reviews (Figure 2).[3] Each item contained text relevant to the item (a plot summary for movies, and the review text for the reviews), an image (a movie poster or an image of the item) and the rating scale being used for that item. The third page asked users to re-rate two items from the second page to ensure that they were rating honestly. The fourth page displays the incentive points users earned in the treatment, as described in Section 3.5.

In the last section of the survey (Figure 3) users responded to survey questions about each treatment, and optionally shared their results with their friends via Facebook. Users rated their satisfaction with each treatment's rating scale using a five point Likert scale. The survey then displayed the user's overall points and prompted them to share their points, ratings, and a link to the survey on Facebook (Figure 4).

## 3.4 Secondary Stimulus

As mentioned earlier, the survey measured cognitive load using a secondary stimulus similar to [5]. The rating page displayed a button with text "click when this is red" in the bottom left corner of the *viewable* browser window (as users scrolled, it remained visible). When the rating page loaded, the button would would wait for a random time between 0 and 20 seconds before slowly (over 4 seconds) turning red and even more slowly (6 minutes) growing vertically to fill the screen. We ensured that users understood and remembered the secondary stimulus button by displaying it on the instruction page and asking them to click it before continuing.[4]

[3] Although we control for item domain in our analysis, this choice was to ensure that page completion time would roughly similar.

[4] The secondary response times on the instructions page were for training purposes only

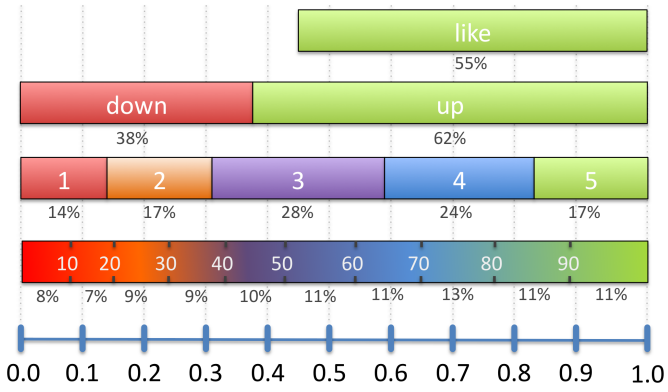


**Figure 5: Distributions of ratings for each scale. For example, thumbs down ratings accounted for 38% of all thumb ratings. The estimation of "like" ratings is described in the text.**

## 3.5 Incentives

We included a point-based incentive system in the survey to achieve three goals (Figure 3). First, we wanted users to find the survey fun enough to complete and share with their friends. Second, we wanted to make sure our timings were accurate; we did not want users to become distracted by other activities while rating. Third, we wanted to make sure users rated honestly. For example, we did not want them to randomly rate items in order to finish more quickly. In order to support these goals, we offered users points for rating quickly, rating accurately, clicking the secondary stimulus quickly, and completing the questionnaire at the end of the survey.[5] We designed a simple heuristic that assigned users points based on their performance in these categories. We also allowed users to share their point totals with friends to make the incentives more fun and compelling.

# 4. RESULTS AND DISCUSSION

The survey ran for the month of February, 2010. Of the 430 people who began the survey, 348 completed it. 43% of respondents were female, the mean age was 26 years old, and 91% of respondents used the internet every day. Respondents learned of the survey through email (43%), an advertisement on MovieLens (25%), Facebook (23%) and other sources (9%).

Users generated 12,847 ratings: 2,010 unary ratings, 4,163 binary ratings, 3,978 star ratings, and 4,426 slider ratings. The lack of a negative response in the unary scale led to lower numbers of unary ratings. Overall, users rated 63% of displayed items, but they rated fewer movies (56%) than products (81%). This makes sense - users could not rate movies they had not seen, but could rate every product review. However, since we showed more movies per page, movies accounted for 66% of all ratings.

As the scales increased in granularity, users rated fewer items. Users rated 73.9% of items with the thumb scale, 69.5% of items with the five star scale, and 67.9% of ratings with the slider scale ($p < 0.001$). Since the unary scale lacks a negative response, users rated significantly fewer items (39.4%) using it. Figure 5 shows the relative frequency of rating values for each scale. Since the unary rating only provided a single response value, we needed to estimate the percentage of items that were *ratable* if the user had other scale responses. We approximated this as the mean percentage of rated items across the other scales (70.4%).

[5]We measured accuracy using re-ratings.

| domain | overall | unary | thumbs | stars | slider | 95% conf. |
|---|---|---|---|---|---|---|
| movies | 4.08 | 3.12 | 3.91 | 4.09 | 5.13 | ±0.12 |
| review | 15.62 | 13.59 | 15.47 | 16.39 | 17.06 | ±0.56 |
| all | 9.87 | 8.45 | 9.92 | 10.05 | 11.00 | ±0.39 |

**Table 1: Mean rating time per item in seconds for each scale and domain. The final column lists 95% confidence intervals based on an ANOVA analysis.**

## 4.1 RQ1: Item rating time

Our first research question explores the time required by different scales:

*RQ1: Do different rating scales require different amounts of time?*

**Data collection:** To measure page ratings, javascript on a user's browser recorded the time a page was loaded and the time the user clicked the "continue" link at the bottom of the page. In total we collected 1,392 page duration timings (mean 95 seconds, median 82 seconds). The distribution of page completion times exhibited a long right tail. We hypothesize that some users forgot about the secondary response or left their browser window open. To account for these users, we truncated outliers at 290 seconds (1.2% of the data). To make the analysis more interpretable, we report the average rating time per item by dividing the page completion time by the number of items displayed on the page (7 for reviews, 20 for products)[6].

**Model:** We used an ANOVA to capture the relationship between rating scale and page completion time. Our analysis included two control variables. We hypothesized that each user would have a different natural rating speed, so we controlled for user. We also hypothesized that users would speed up as they gained speed and experience during the survey, so we controlled for the treatment number (1,2,3, or 4) of a rating page. The control variables did correlate with page completion. Different users exhibited different rating speeds (F value = 2.5, $p < 2e^{-16}$). Users also rated more quickly as they proceeded through the pages; the average rating time dropped from 12.5 seconds per item on the first page to 9.4 seconds, 8.7 seconds, and 8.4 seconds on pages two, three and four (F value = 37, $p < 2e^{-16}$).

**Results:** Table 1 shows the results from the analysis. Overall, the average rating time was 9.87 seconds. Users rated movies (4.08 seconds) much more quickly than product reviews (15.62 seconds).

Users' average rating time increases as the granularity of the scale increases. Overall, the slider required 30% more time than the unary scale (8.45 seconds versus 11.0 seconds). To identify significant differences between individual pairs of scales, we used the Tukey Honest Significant Difference test. All pairwise differences were significant at the 0.05 level or below except for the thumbs and five-star scale.

The results for page completion time varied significantly between movies and product reviews. For movies, users completed unary pages 62% faster than slider pages (3.12 vs 5.13 seconds). For product reviews, users completed unary pages 25% faster than

[6]We divided by the number of items displayed on a page instead of the number of items rated on a page for consistency across scales. Although the survey asked users not to rate movies they did not know, with the unary scale a lack of rating does not imply that the user did not know the movie - they may just not have liked it.

| domain | overall | unary | thumbs | stars | slider | 95% conf. |
|---|---|---|---|---|---|---|
| movies | 5.95 | 5.57 | 5.76 | 5.86 | 6.39 | $\pm$0.27 |
| review | 5.84 | 5.40 | 5.96 | 6.08 | 5.89 | $\pm$0.20 |
| all | 5.89 | 5.47 | 5.89 | 5.98 | 6.13 | $\pm$0.17 |

**Table 2: Mean secondary response time per item in seconds for each scale and domain. The final column lists 95% confidence intervals based on an ANOVA analysis.**

slider pages (17.06 vs 13.59 seconds). However, the absolute time difference between unary and slider was greater for product reviews (2.01 seconds for movies vs 3.47 seconds for products).

The higher relative difference in rating times for movies (62% vs 25%) suggests users may apply a two-staged rating process. In the first stage, users recall or experience an item. In the second stage, users evaluate and rate that item. The first stage varies less with rating scale, and will often be much longer for experiential items such as product reviews. For movies, the first stage (remembering a movie) is relatively short and the rating time dominates. Based on this hypothesis, different scales may lead to the largest differences in total rating time for ratings systems that ask users to rate many quickly recalled items, such as movie, music or book recommenders.

## 4.2 RQ2: Cognitive load

Our second research explores cognitive load, a measure of instantaneous mental effort:

*RQ2: Do different rating scales elicit different levels of cognitive load?*

**Data collection:** As mentioned earlier, we measured cognitive load by timing user's response time to a secondary stimulus: a button that turns red and grows. In total we collected 7,644 secondary timings (mean 5.89 seconds, median 4.25 seconds). Like page completion times, the distribution of secondary response times exhibited a long right tail. We hypothesize that some users forgot about the secondary response or left their browser window. To account for these users, we truncate outlying secondary response times at 40 seconds (1.7% of the data).

**Model:** As with rating time, we analyzed secondary response time using an ANOVA between response time and scale, controlling for treatment number and user (Table 2). The ANOVA found both control variables to be significant. Different users responded more quickly (F-value = 10.4, $p < 2.2e^{-16}$). Users also responded more quickly as the survey progressed; response times dropped from 7.0 seconds in the first treatment to 5.3 seconds in the last treatment (F-value = 56.5, $p < 2.2e^{-16}$). These findings may indicate that the rating tasks vary in difficulty for different users and treatment numbers, or they may indicate that the actual secondary response task itself varies in difficulty.

**Results:** Table 2 shows the results for RQ2. Overall, the relationship between scale and cognitive load is inconclusive. Although the ANOVA finds that secondary response times vary significantly with scale (F-value = 5.8297, $p = 0.0005$) the results between reviews and movies are inconsistent. For movies, response times monotonically increase by a small non-significant margin with scale granularity: from 5.57 seconds for unary to 6.39 seconds for slider. For reviews, response times increase from unary to thumbs to five-star scale, but response times for slider fell below those of all other scales but unary. A Tukey HSD post-hoc test indicates that the pairwise differences between the unary scale and other scales are significant ($p < 0.05$), but the pairwise differences between other scales are not.

In summary, we find evidence that users work less hard in an instantaneous sense under the unary scale, but the results for the other scales are inconclusive. This should make it easy for practitioners choosing between non-unary scales to measure mental effort, as they can simply measure rating time. It is possible that a more accurate test of cognitive load, such as pupillary response [12], could identify meaningful differences for the other scales. We leave this for future work.

## 4.3 RQ3: Time for different rating values

Our third research question examines the relationship between time and a specific rating value such as two-stars or thumbs-up.

*RQ3: Do different rating values require different amounts of time?*

**Data collection:** To measure rating times for different rating values, we needed to calculate rating times for individual items on a page, not the page as a whole. Since the survey displayed multiple items per page, we could not reliably tell when a user started evaluating a particular item for two reasons. First, users may rate items out of order. Second, users may evaluate some items and ultimately choose not to rate them. In an earlier version of the study, we instrumented the survey to fade out items that were not at the center of the page, enabling the system to know which item a user was evaluating. However, users strongly disliked this interface manipulation, and we ultimately used a more traditional rating page.

To reduce the effects of unrated items and items rated out of order, a rating for item $k$ is only included if 1) the user rated item $j$ displayed immediately above it, and 2) no other items were rated in between $j$ and $k$. If both criteria are met, the rating time for $k$ is calculated as the elapsed time between the ratings for $j$ and $k$. Because of the constraints imposed by this procedure, rating times for the items displayed first and last on a page are not included. Since item ordering is randomized, this should not affect our results.

**Model:** We created an ANOVAs measuring the relationship between rating value and time for each of the six treatments ([movies, reviews] $\times$ [binary, five-star, slider]). Unary is not included because it only corresponds to one rating value. To make our results more interpretable, all rating scales are translated to the [1,5] range. Thumbs down and thumbs up are encoded as 2 and 4 respectively. Slider ratings for ranges [0, 20), [20, 40), [40, 60), and [80,100] are grouped together as 1, 2, 3, 4, and 5 respectively.

As with the previous ANOVAs, user and treatment number are treated as control variables. We also added review length as a control variable for product reviews because we thought that longer reviews may receive higher ratings. The times reported are differentials after control variables are removed. Because of this, all the rating time differentials reported by one ANOVA will sum to 0.

**Results:** Figure 7 shows the times associated with each rating value. The figure displays a different line for each rating scale and domain. The dotted lines represent review ratings, while the solid lines show movie ratings. Each color corresponds to a difference scale: blue lines correspond to slider, red correspond to five-star, and green corresponds to the thumbs scale. The figure displays the offset in average rating time after controlling for user, treatment number, and length (for product reviews only). For example, the dotted red line at x-position 3 indicates that users who rate a review three out of five stars generally rate one second faster than the average star rating for a review.

The ANOVAs indicate that rating value is sig nificantly related to rating time at the $p < 0.05$ level or below for all treatments but

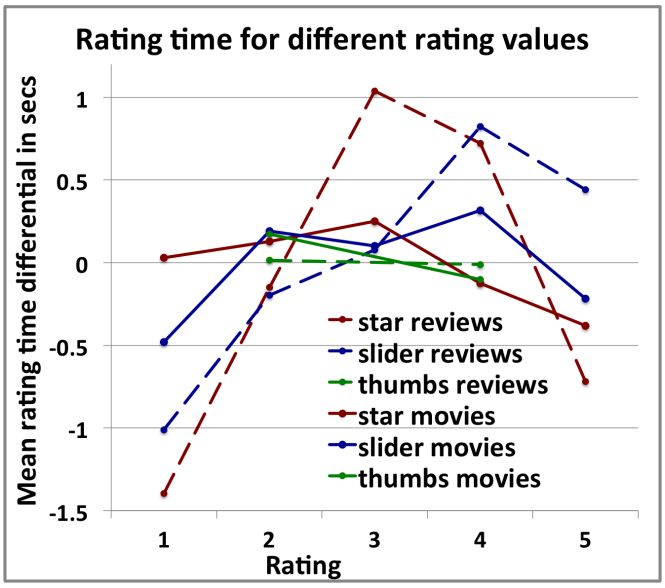


**Figure 6: Times for items rated different values (e.g. three-stars vs. four-stars). Each line represents a different rating scale. Review rating times are shown with dotted lines and movie rating times are shown with solid lines. Ratings for all scales have been translated to [1,5]. Times are reported as a differential from the mean time for all ratings represented by a line after controlling for user, treatment number, and review length.**

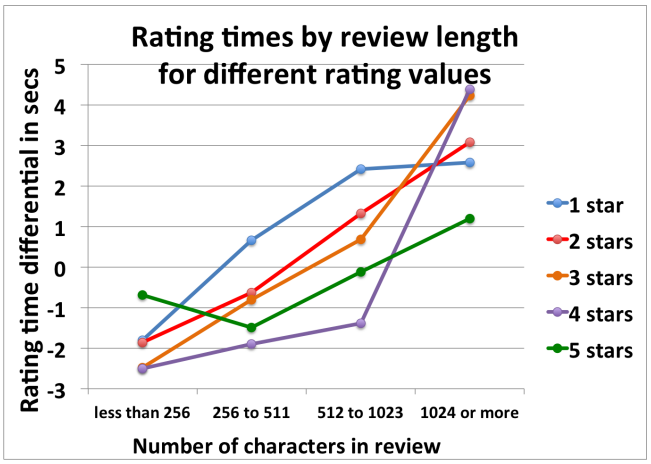


**Figure 7: Rating times for reviews of different lengths. Each line represents times for a particular rating value. Times are reported as a differential from the average for all ratings associated with the line. For all rating values, rating times generally increase with review length. Rating times vary with length most strongly for reviews rated three or four stars and least strongly for reviews rated five and one stars.**

binary ratings for product reviews. In general, it seems that users rate on the endpoints of a scale more quickly than in the middle. One explanation for this is that users may easily recall items they react strongly to. The relationship between rating value and rating time differs by domain, with larger effects for reviews. Under the five-star scale, the reviews rated three stars require about 2.5 seconds longer than those rated one star. The effects may not be as large for the thumbs scale because the scale does not support more extreme rating values.

The shorter rating times for extreme ratings hints toward a sweet spot for rating systems wishing to intelligently learn a user's preferences. Extreme ratings are often precisely the ratings that offer the greatest insight into a user's taste profile. A system that can frequently ask a user about items she loves or hates may not only learn more a user's profile well, it may learn it more quickly.

**Length vs rating value:** We wanted to investigate if and how users change their reading style based on the quality of a review. Perhaps they read a small portion of a low quality review before forming a quick judgement, but carefully read an entire review before deciding to rate it five stars. If this is the case, the rating times for one-star reviews would not vary with the length of a review, but the five-stars reviews would vary.

Figure 7 shows the relationship between rating time and the number of characters in a review. The methodology mirrored that for the previous analysis. We choose to limit our analyses to five-star ratings because they provided an appropriate level of granularity for our analysis. We grouped reviews by their characters length: [0, 256), [256, 512), [512, 1024), and [1024, $\infty$)[7]. Each line on the graph represents one rating value's relationship between review length and rating time. For example, the blue line shows the average rating time for different length reviews that are rated one star.

At all rating levels, users spend more time rating longer reviews, as shown by the upward trends on the graph. However, a pattern emerges in the differential between the time to rate a short versus long review. Rating times for short and long reviews varied most for reviews rated three or four stars. Contrary to our hypothesis, reviews rated five stars vary the least with length, followed by reviews rated one star. This suggests that users read more quickly once they have identified a review may be of very high or low quality.

In summary, we find that users rate items they assign extreme ratings to more quickly. In addition we find that users do not just rate an item faster after they experience it; our analyses of review length and rating time suggests that users actually experience (i.e. read) faster when they assign extreme ratings. It is possible that these results were shaped by the artificial task imposed by our survey. Our users may not have cared to learn about a Sony Television or Cuisinart blender. Therefore, once they had determined a product review would be definitely assigned a particular rating, they may have lost interest in reading the remainder of the review. Future researchers may test this by measuring rating times during an actual field trial of a real rating system.

## 4.4 RQ4: User satisfaction

In the final section of the survey, we asked user to rate their satisfaction to answer:

*RQ4: Do different rating scales elicit different levels of user satisfaction?*

**Data collection:** At the end of the survey we asked users to rate their agreement using a five point Likert response to the statement "Overall I liked using the {slider} to rate {movies}." We encode the responses numerically using the natural 1 to 5 encoding.

**Results:** Users' mean satisfaction response was 3.2 out of 5.0. Figure 8 show users' satisfaction with each scale. Users preferred the scales in the middle of the granularity spectrum (thumbs and

[7]This binning yielded roughly equal counts per bin

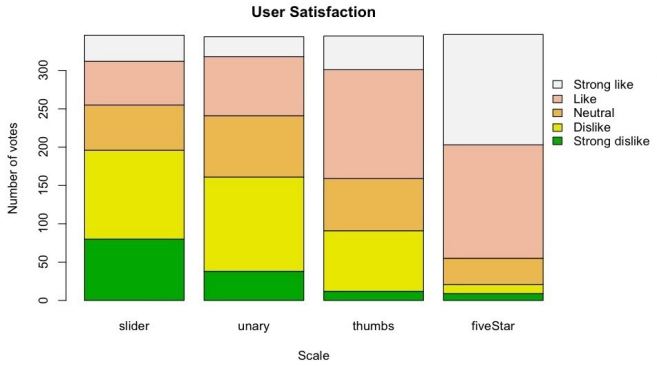


**Figure 8: User satisfaction across all scales, disregarding domain. Each vertical segment is the proportion of ratings that a given scale received.**

five star), but liked the star scale best. Users disliked the scales at the ends of the granularity spectrum (slider, unary). Users mean satisfaction responses differed significantly: 2.78 for unary, 3.48 for thumbs, 4.17 for five star, and 2.56 for sliders ($p < 0.05$ for pairwise two-tailed t-tests). 83% of users liked the five-star scale compared to 54% for the thumbs scale. Users liked the slider least (57% dislike), and only liked the unary scale slightly more (47% dislike).

Users' absolute satisfaction scores vary moderately between movies and product reviews for some scales. When comparing movies to product reviews, users report a higher satisfaction for the five star scale (4.32 vs 4.0), and a slightly lower satisfaction for the thumbs scale (3.25 vs 3.48). Other differences were not significant. We hypothesize that users may prefer finer-grained scales more for subjective item domains, but more research is necessary.

# 5. CONCLUSION

This paper reports on a survey that measure the costs and benefits associated with different rating scales. To summarize our findings:

- Users' rating costs increase as they have more rating choices. All scales show similar cognitive load. However, page rating times increase significantly with finer-grained scales - despite users leaving more items unrated with those scales.
- User rating times between the unary and slider vary more in relative time for movies (62% vs 25%), but vary more in absolute time for reviews (about 2 vs 3.5 seconds).
- Users spend more time assigning ratings at the middle of a scale, such as three or four stars on a five-star scale.
- Users prefer the five-star scale overall, although the thumbs scales comes in as a relatively close second choice for product reviews.

Although we studied item domains that spanned item characteristics, it is difficult to know whether our results will generalize to other domains. For example, when users rate YouTube videos, they cannot "watch more quickly" in the same way a user can read more quickly. However, users may stop watching a video they dislike and rate it thumbs down. Overall, we believe that many of our findings will apply to different domains, but there will be exceptions.

We believe our survey is unique in its use of point incentives, often called "game mechanics" [3], to shape user behavior. To our knowledge, our survey is the first to do so. During a pilot of the survey, testers found it difficult to remember all the guidelines we gave them (rate quickly, rate accurately, quickly click the secondary response). Users said that the point system combined with the social sharing incentives motivated them to achieve all three objectives, even though they said that they did not understand exactly how the survey calculated points.

In addition to the findings for designers outlined earlier, two results suggest that designers should carefully evaluate different scales before deploying them on a site. First, based on the results for RQ1, systems asking users to rate many quickly recalled items, such as book recommenders, are most strongly affected by the choice of rating scale. Second, our results for RQ3 suggest that the choice of scale may change the way a user experiences an item. This effect may be surprising to designers.

Based on our findings, researchers should investigate systems that adapt their rating scale to the context of a rating. Early in a user's lifecycle a system may prefer lots of coarse-grained feedback about a user's taste profile. As the system gains learns more about a user, it may desire a few highly-detailed pieces of information. If this is true, recommender systems might shift form a low-granularity scale for new users to a higher-granularity scale for more experienced users.

Our work points toward a variety of future research directions. Due to sample size limits, we only tested two item domains. We hypothesize relationships between characteristics of those domains and our results, but our findings must be verified across more item domains. Researchers should verify that our results hold across other rating scale designs with similar granularity. For example, our results for binary scales should be tested with the vertically oriented layout used by Digg.

# 6. 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 transactions on knowledge and data engineering*, pages 734–749, 2005.

[2] G. Al Mamunur Rashid and J. Riedl. Learning preferences of new users in recommender systems: an information theoretic approach. *ACM SIGKDD Explorations Newsletter*, 10(2), 2008.

[3] A. Bader-Natal. Incorporating game mechanics into a network of online study groups. In *AIED 2009: 14 th International Conference on Artificial Intelligence in Education Workshops Proceedings*, page 109. Citeseer, 2009.

[4] C. Boutilier, R. Zemel, and B. Marlin. Active collaborative filtering. In *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, pages 98–106. Citeseer, 2003.

[5] R. Brunken, J. L. Plass, and D. Leutner. Direct measurement of cognitive load in multimedia learning. *Educational Psychologist*, 38(1):53–61, 2003.

[6] D. Cosley, S. K. Lam, I. Albert, J. A. Konstan, and J. Riedl. Is seeing believing?: how recommender system interfaces affect users' opinions. In *Proceedings of SIGCHI*, pages 585–592. ACM New York, NY, USA, 2003.

[7] J. Gwizdka. Distribution of cognitive load in web search. *Arxiv preprint arXiv:1005.1340*, 2010.

[8] F. M. Harper, X. Li, Y. Chen, and J. A. Konstan. An economic model of user rating in an online recommender system. In *User Modeling 2005*, pages 307–316. 2005.

[9] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl. An

algorithmic framework for performing collaborative filtering. In *Proceedings of SIGIR*, page 237. ACM, 1999.

[10] G. A. C. Jr and J. P. Peter. Research design effects on the reliability of rating scales: a meta-analysis. *Journal of Marketing Research*, XXI(1):360–375, November 1984.

[11] J. Liu, Y. Cao, C. Y. Lin, Y. Huang, and M. Zhou. Low-quality product review detection in opinion summarization. In *Proceedings of EMNLP-CoNLL*, page 334âĂŞ342, 2007.

[12] F. Paas, J. E. Tuovinen, H. Tabbers, and P. W. M. V. Gerven. Cognitive load measurement as a means to advance cognitive load theory. *Educational psychologist*, 38(1):63–71, 2003.

[13] J. Patrick, R. Pruchno, and M. Rose. Recruiting research participants: a comparison of the costs and effectiveness of five recruitment strategies. *The Gerontologist*, 38(3):295, 1998.

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

[15] A. Rashid, I. Albert, D. Cosley, S. Lam, S. McNee, J. Konstan, and J. Riedl. Getting to know you: learning new user preferences in recommender systems. In *Proceedings of the 7th international conference on Intelligent user interfaces*, pages 127–134. ACM, 2002.