

June 2010

# Cognitive Load of Rating Scales

E. Isaac G. Sparling  
*Macalester College, isaac.sparling@gmail.com*

Shilad Sen  
*Macalester College, ssen@macalester.edu*

Follow this and additional works at: [https://digitalcommons.macalester.edu/mathcs\\_honors](https://digitalcommons.macalester.edu/mathcs_honors)



Part of the [Graphics and Human Computer Interfaces Commons](#)

---

## Recommended Citation

Sparling, E. Isaac G. and Sen, Shilad, "Cognitive Load of Rating Scales" (2010). *Mathematics, Statistics, and Computer Science Honors Projects*. 17.  
[https://digitalcommons.macalester.edu/mathcs\\_honors/17](https://digitalcommons.macalester.edu/mathcs_honors/17)

This Honors Project - Open Access is brought to you for free and open access by the Mathematics, Statistics, and Computer Science at DigitalCommons@Macalester College. It has been accepted for inclusion in Mathematics, Statistics, and Computer Science Honors Projects by an authorized administrator of DigitalCommons@Macalester College. For more information, please contact [scholarpub@macalester.edu](mailto:scholarpub@macalester.edu).

# Cognitive Load of Rating Scales

Submitted to the Department of Mathematics,  
Statistics and Computer Science in partial  
fulfillment of the requirements for the degree of  
Bachelor of Arts

E. Isaac Sparling

Advisor: Prof. Shilad Sen  
Second Reader: Prof. Brooke Lea  
Third Reader: Prof. Daniel Kaplan

MACALESTER COLLEGE

May 11, 2010

## 1 Abstract

Why does Netflix.com use star ratings, Digg.com use up/down votes and Facebook use a “like” but not a “dislike” button? In this paper, we extend existing research on rating scales with findings from an experiment we ran to measure the cognitive load users experience while rating. In this paper, we analyze the cognitive load and time required by different rating scales. Our analysis draws upon 14,000 movie and product ratings we collected from 348 users through an online survey. In the survey, we measured the speed and cognitive load users experience under four scales: unary (‘like it’), binary (thumbs up / thumbs down), five-star, and a 100-point slider. We compare a variety of measures of cognitive load and rating speed, and draw conclusions for user interface designers based on our results. Our advice to designers is grounded in the responses from users regarding their opinions of scales, the existing research, and in the models we build of the data collected from the experiment.

## Contents

<b>1</b>	<b>Abstract</b>	<b>i</b>
<b>2</b>	<b>Introduction</b>	<b>1</b>
<b>3</b>	<b>Related Work</b>	<b>2</b>
3.1	Choosing Scales . . . . .	2
3.2	Cognitive Load . . . . .	2
3.2.1	Measuring Cognitive Load . . . . .	3
<b>4</b>	<b>Dimensions</b>	<b>5</b>
4.1	Rating Scale Dimensions . . . . .	5
<b>5</b>	<b>Methodology</b>	<b>5</b>
5.1	Scales . . . . .	5
5.2	Item Selection . . . . .	6
5.2.1	Product Reviews . . . . .	7
5.2.2	Movies . . . . .	7
5.3	Survey Overview . . . . .	7
5.4	Survey Screenshots . . . . .	8
<b>6</b>	<b>Experimental Design</b>	<b>17</b>
6.1	Secondary Stimulus . . . . .	17
6.2	Incentives . . . . .	17
6.3	Pilot Study . . . . .	18
6.4	Survey Implementation . . . . .	19
<b>7</b>	<b>Results</b>	<b>19</b>
7.1	Distribution of Ratings by Scale . . . . .	21
7.2	Metrics . . . . .	25
7.3	Correlations Between Measures . . . . .	26
7.4	Page Speed . . . . .	26
7.5	Rating Speed . . . . .	28
7.6	Inter-Rating Speed . . . . .	30
7.7	Secondary Measure Speed . . . . .	31
7.8	Item Duration . . . . .	33
7.9	User Satisfaction . . . . .	34
<b>8</b>	<b>Discussion</b>	<b>35</b>
8.1	RQ1 . . . . .	35
8.2	RQ2 . . . . .	35
8.2.1	Page Completion Times . . . . .	35
8.2.2	Rating Speed . . . . .	35
8.2.3	Mouse-over Duration . . . . .	36
8.3	RQ3 . . . . .	36

---

<b>9</b>	<b>Conclusions and Further Work</b>	<b>37</b>
9.1	Further Research . . . . .	38
<b>A</b>	<b>Histograms</b>	<b>39</b>

## 2 Introduction

Rating scales have become a pervasive element of the modern internet experience. As the internet has democratized access to information, content publishers have realized that those users were an untapped resource: across many domains, users willingly contribute information that content publishers then use for a variety of purposes. Readers of blogs comment extensively; buyers on Amazon.com review products and users on Netflix rate movies. Frequently, user-generated content involves rating an item – a movie, a product, a review. To date, choosing a scale has been an aesthetic choice made by the designer. Choices include unary (“I like this”), binary (thumbs up / thumbs down), five star and finely grained sliding scales.

This paper explores the cognitive load caused by four different scales (a unary scale, a binary scale, a five star scale and a 100 point sliding scale) in 2 different item domains (movies and product reviews). By more clearly understanding how rating scale choice affects cognitive load, user interaction designers can make more informed choices.

People’s rating contribution styles may vary. Some people may be high-quality contributors: they carefully consider each contribution. However, others may be high-quantity contributors. These two contribution styles conflict with each other: assuming a static amount of available time, as one increases the number of contributions, they must spend less time on each contribution. This is a trade off. Site designers and operators can build their sites in such a way to encourage one kind of interaction over the other. Deciding where to live on this quantity vs quality continuum is a choice that shouldn’t be made without a more complete comprehension of the affects of the rating scale.

We believe that there are several contributing costs which place users at particular points on the quality vs quantity continuum. Both cognitive load and time expenditures are costs to users. We ask three research questions about these features of user interactions:

*RQ1a: How does the cognitive load of a user vary with choice of scale?*

*RQ1b: Does cognitive load vary with item domain?*

*RQ2a: How does speed vary with choice of scale?*

*RQ2b: Does a particular scale cause faster completion times in one domain?*

With answers to these questions, we could predict how to best reach users as long as we made a critical assumption: users have no preferences. However, this is a faulty assumption, as users do have preferences, and those preferences may determine whether or not they will interact with a system. We pose another research question:

*RQ3: How satisfied are users with each scale?*

Are there specific conditions in which users are more or less satisfied with a particular scale? Understanding users' preferences as well as the carefully measured timing data will help us balance trade-offs in designing optimized user interfaces.

The methodology for answering RQ1 and RQ2 is similar: users take a web survey, and we log timing data. For the first two questions, we time users completing a secondary task in a *dual task* measurement system, as discussed below. We collect data for the second question by recording the time it takes a user to complete a page of ratings, the time it takes a user to rate an individual item, and the amount of time a user spends moused over an individual item. The third question is answered by polling users on their feelings about particular scales at the end of the survey.

## 3 Related Work

### 3.1 Choosing Scales

Our choice of scales was partially influenced by what appears in the wild. However, previous studies were important in coming to a final decision. Cosley et al. (2003) explore the differences between a variety of scales in terms of data utility and user satisfaction and found that scale doesn't significantly affect data utility, but users are more satisfied with higher granularity (though the finest scale they examined had only 6 options, and they did not examine a unary option). This is supported by Garner (1960) who concludes "it is clear that information cannot be lost by increasing the number of rating categories." Guilford (1938) states that the optimal number of response categories can be as high as 25.

### 3.2 Cognitive Load

Brunken et al. (2003) briefly summarize some key attributes of cognitive load, which expose why it is an interesting metric to consider. They note:

Sweller (1999) distinguished three types of load: one type that is attributed to the inherent structure and complexity of the instructional materials, and cannot be influenced by the instructional designer, and two types that are imposed by the requirements of the instruction and can, therefore, be manipulated by the instructional designer.

They continue to define these two types of load that designers can manipulate by differentiating them into *extraneous* cognitive load and *germane* cognitive load. Extraneous load is “cognitive load imposed by the format and manner in which information is presented and by the working memory requirements of the instructional activities...[it] is a form of overhead that does not contribute to an understanding of the materials.” Germane load, on the other hand, is “the load induced by learners’ efforts to process and comprehend the material.”

The cognitive load a rating scale causes a user falls directly into these latter two categories. Since humans have a finite cognitive capacity, tasks with a higher extraneous cognitive load cause users to have less available capacity to put towards the optimal germane cognitive load. (Brunken et al., 2003) Simply put, the increased brain power required to rate is preemptively taken from the pool available for thinking about the task at hand.

Harper et al. (2005) present an economic analysis of user interaction on websites. They model a user’s likelihood to rate an item as an amalgamation of costs against the benefits they will receive on completing a rating. Cognitive load is a cost to a user. Since  $value = benefits - costs$ , if value is negative (or the costs are too high), a user will choose not to rate. However, as Cosley et al. (2003) have shown, users relate to different scales differently, a fact which points out that the choice of scale can affect the benefits a user gets from rating, and agrees with Sen et al. (2007), who found that a binary system is more useful than either a negative (‘thumb down’) or a positive (‘thumb up’) unary system for reacting to tag quality.

Through a more faceted understanding of the cognitive load a user undergoes as related to a scale, designers might be able to exploit a scale’s properties to control how and when interactions occur by minimizing the chances that the *value* is not negative. However, designers could choose to do the opposite, and raise the barrier cognitive load provides. By doing this, and choosing a scale with a higher load, fewer people would interact, but presumably, each interaction would produce slightly higher quality data due to a more involved thought process.

### 3.2.1 Measuring Cognitive Load

Brunken et al. (2003) discuss a variety of methods to measure cognitive load. Table 1 discusses the main axes along which the methods fall. We discuss some of the key points covered in their paper.

We use two of these methods for measuring cognitive load in this survey. Our primary measure uses the dual task paradigm, and we also ask users to reflect on their experiences with each scale at the end of the survey. The dual task metric is quite useful for several reasons. First, it is an easy measurement technique to implement on a web survey: it doesn’t require external equipment



<i>Objectivity</i>	<i>Causal Relationship</i>	
	<i>Indirect</i>	<i>Direct</i>
Subjective	Self-reported invested mental effort	Self-reported stress level Self-reported difficulty of materials
Objective	Psychological measures Behavioral measures Learned outcome measures	Brain activity measures (e.g., fMRI) Dual-task performance

Table 1: Classification of Methods for Measuring Cognitive Load Based on Objectivity and Causal Relationship (Brunken et al., 2003)

or the presence of a researcher to code reactions. Second, the dual task measure also measures cognitive load in the moment, as opposed to after the fact, as a questionnaire does. Finally, interpreting the results is relatively easy: faster reaction times mean lower cognitive load. (Brunken et al., 2003)

The dual task paradigm is simple to implement. Users are instructed to carry out a primary task (in the case of this survey, rating items). Every so often, a stimulus occurs, which the user must recognize and react to. This is the secondary task – users must remain aware of it, but not stay focused on it.

The measure has drawbacks. According to Brunken et al. (2003), the secondary task must be as cognitively challenging as the primary task. If it isn't, performance on the secondary task will be completely independent from performance on the primary. However, the task must be *simple*. If it isn't simple, it will be prioritized over the primary task. If this happens, the measure is no longer a secondary task and is rendered invalid. Finally, the secondary task must be able to fill all available free cognitive capacity. Brunken et al. (2003) examine these variables and conclude that a measure of reaction time is an effective way to meet these functional requirements of the secondary measure:

Because the monitoring task itself requires only few cognitive resources, it does not suppress the primary task; yet, when a reaction is necessary, the as-soon-as-possible condition consumes all available resources. This design minimizes the interference between the two tasks and maximizes the exhaustion of the free capacity. The performance on the secondary task (i.e., the speed of reaction) is directly related to the amount of free capacity, which itself depends on the amount of capacity required by the primary task.

Our reaction-based secondary stimulus conforms neatly to the design Brunken et al. (2003) propose for a multimedia environment.

## 4 Dimensions

### 4.1 Rating Scale Dimensions

A variety of rating scales are used across the web. We explore some aspects of these rating scales which might impact cognitive load and speed before zeroing in on several specific choices to implement in our study. Rating scales can be applied to an enormous variety of items and in a wide situations. The dimensions that might affect a user's interaction with a given scale include:

- **Experiential vs Instantaneous:** Is a user being asked to recall and rate a previous experience, or are they being asked to ingest content in the instant before rating?
- **Fact vs Opinion:** This resembles the objective/subjective relationship with an important difference: the user is being asked to rate the veracity of a fact as opposed to rating their opinion of an item.
- **Design:** Is the scale a drop down widget? Does the scale appear inline with items?
- **Location:** Where is the scale located? Do users have to navigate to a distinct rating page, or can they rate as they experience content? Do they rate multiple items at once, or each individually?
- **Scale Granularity:** How is rating presented? Two thumbs? Five stars? How many discrete options does the scale provide?

Choosing domains of items to explore requires careful consideration for a study like this. We felt that users interact in fundamentally different ways with objective and subjective items. The core difference between objective and subjective is a matter of agreement: a subjective item is one where ratings will fall across the spectrum, whereas an objective one's ratings will more readily agree with each other. We feel that product reviews and movies exemplify these two categories quite well.

## 5 Methodology

### 5.1 Scales

The four scales we chose to study are listed below, ordered by increasing granularity:

- **Unary :**  Like this movie

- Binary : 
- Five Star : 
- Slider :  

We chose these scales because we felt that they represented the variety of scales that appear in the wild. We chose these particular implementations because they effectively balanced ease of implementation in a survey environment and they were simple from a user's perspective. The only significant visual deviation from commonly accepted standards is our representation of the unary scale: typically, an icon which changes (Youtube's white heart changes to a red heart when you click "Favorite") and/or descriptive text changes (Facebook's "Like" changes to "Unlike," and a small thumb-up appears next to the rated post). We chose to use a simple checkbox as it captured the core of a simple, dynamic change that uniformly represents a rating across these various implementations.

The unary scale has been popularized by Facebook, where users can note an interest in a particular item (wall-post, photo, video, etc) by clicking a "Like" button. The unary scale is in use across many other sites as well, where people can mark items as "favorites" for later examination or quick retrieval. Overall, the unary scale provides an interesting counterpoint to the highly granular 100 point slider scale.

The thumbs (generically, a binary) scale is another clear choice: it has been used widely in many social-news aggregators (digg.com, reddit.com). In these sites, users rate both submissions (external links) and comments by other users. Frequently the thumb paradigm makes sense in such sites, as "thumbing up" correlates to moving an item higher on the page.

The five star scale was chosen as it is used in many different situations (eg: rating product reviews on amazon.com, rating movies on netflix.com or rating apps on the iTunes App Store). This makes it an interesting choice, simply because more fully understanding the scale will help us understand whether or not these companies have made an effective selection in their choices.

The 100 point slider was chosen as a logical extension of the findings of Garner (1960), who proposed that more options are not detrimental to the rating experience.

## 5.2 Item Selection

Due to the different provenances of the items being rated, they had to be collected differently. However, we aimed to keep the general method the same. We tried to select

1. Popular (or recognizable) items
2. Both high- and low-quality items.

This first criterion was to minimize any time users would have to spend understanding an item. If they could focus simply on evaluating it we would be able to take more accurate measurements. We also wanted to ensure an even distribution of good and bad items to ensure our findings weren't skewed towards one class of items over another. The second criterion is to ensure that our findings aren't skewed towards high or low quality items.

### 5.2.1 Product Reviews

For product reviews, we chose to take reviews from 3 devices – a Sony HDTV, a third generation Apple iPod Touch and a Cuisinart blender. These three products represent a variety of products that people use regularly–this meets our first criterion, for familiarity with the items. To meet the second criterion, we sorted the reviews by Amazon's quality (as determined by users' ratings) and selected every fifth review.

### 5.2.2 Movies

For movies, we scraped imdb.com's Top 500 (All Time USA Box Office) list. We chose to scrape this particular list because we felt that movies which had made money would be in the public consciousness (meeting the first of our criteria). Observation showed that not all movies which made large sums of money were considered good movies: the list contained both critically acclaimed movies (Star Wars: A New Hope (rated 8.8/10), the Lord of the Rings trilogy (rated 8.8/10)) and critically detested movies (Transformers: Revenge of the Fallen (rated 6.1/10)), which meets the second criterion.

## 5.3 Survey Overview

We presented users with an online survey which carefully logged most of their actions. The survey consisted of three main parts: an introductory set of instructions, a series of four treatments (the bulk of the survey), and finally a followup questionnaire, where users reflected on their experience with the rating scales.

The middle section was the most complex. In this section, we gave users 4 pseudorandomly generated pairings of scale and domain. We refer to this pairing as a *treatment*. Before each treatment, we showed a customized set of instructions, telling the user how they should complete the rating task at hand. These

instructions differentiated between rating *movies* and product *reviews* as well as how the rating scale should be used. They also reminded users about the secondary stimulus – during our alpha test, users would regularly forget the stimulus.

Every page followed the same format. Users were presented with either 20 movies or 7 product reviews to rate, laid out vertically. This choice was to ensure that page completion time would be of a similar order of magnitude – 20 product reviews would take significantly longer to rate than 20 movies. 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. Layout was consistent across all scales and domains.

The users were recruited simply: we composed several emails and sent them out to lists: Macalester’s Math/Computer Science list and GroupLens Research’s MovieLens.org research list. We recruited family and friends. Macalester College’s social networking office tweeted out a call to participate. All recruits were encouraged to forward the survey to their friends, to help widen our user base.

## 5.4 Survey Screenshots

We start where a user begins her experience of the survey: with the introductory demographics questionnaire, as seen in Figure 1.

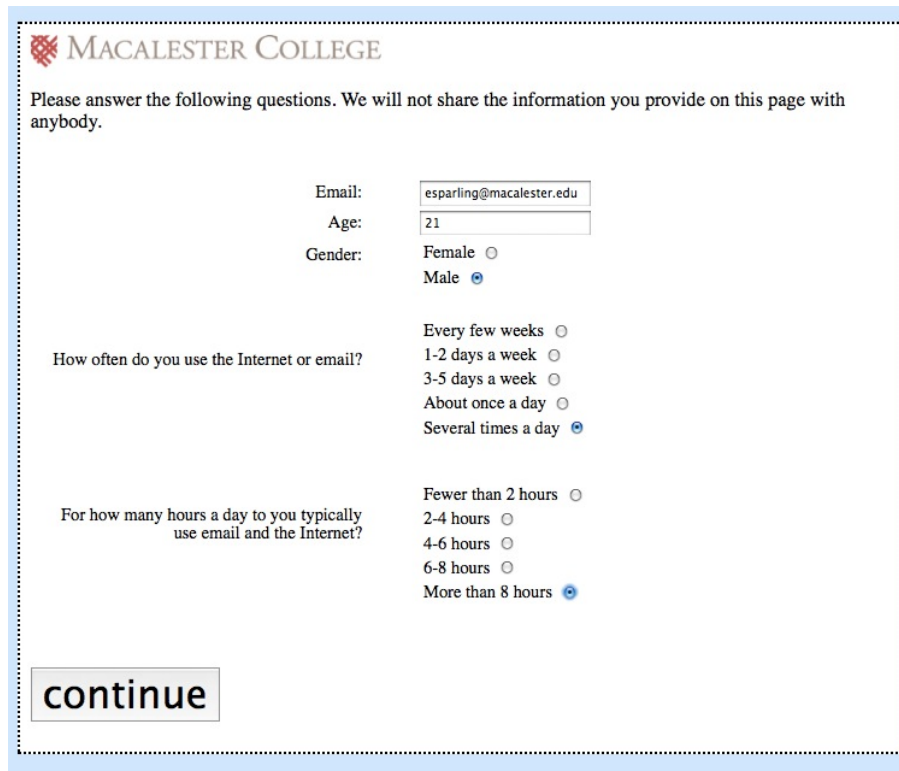
Users then see their first page of instructions (Figure 2), which both detail how to use the particular scale and how to rate for the particular domain, as well as reminding them to be aware of the secondary stimulus by requiring them to click a stand in for the stimulus. On clicking the stimulus stand-in, the upper box is changed to what is shown in Figure 3.

After reading through the instructions, users are brought to the page where they can rate items, as seen in Figure 4. This figure shows two important features of a rating page. Users are reminded once again, at the top of the page, how to most effectively interact with the survey. It also shows the secondary stimulus in the lower left, already turned red, and starting to grow to catch the user’s attention.

After completing a section of ratings, users are shown the points they acquired during that section via the interface shown in Figure 5.

These three elements (instructions, rating treatment, points display) repeat three more times; once per scale/domain pairing, before the user is presented with the final reflection questionnaire, as seen in Figure 6.

The final screen (Figure 7) that we showed all users was the points screen, where the four separate treatment points are tabulated and presented together. Users



MACALESTER COLLEGE

Please answer the following questions. We will not share the information you provide on this page with anybody.

Email:

Age:

Gender:  Female  
 Male

How often do you use the Internet or email?  
 Every few weeks  
 1-2 days a week  
 3-5 days a week  
 About once a day  
 Several times a day

For how many hours a day to you typically use email and the Internet?  
 Fewer than 2 hours  
 2-4 hours  
 4-6 hours  
 6-8 hours  
 More than 8 hours

Figure 1: Introductory demographic questionnaire.

are also prompted to share their points and ratings via Facebook. Figure 8 shows what users saw when the link was posted to their Facebook wall.

Users who chose to share their points and ratings were then taken to the public page, which showed the same point graph they had previously seen, and allowed the general public to peruse their ratings, in a format similar to how items were presented within the survey proper. Anybody who navigated to this page was prompted to take the survey. Figure 9 details this final page.

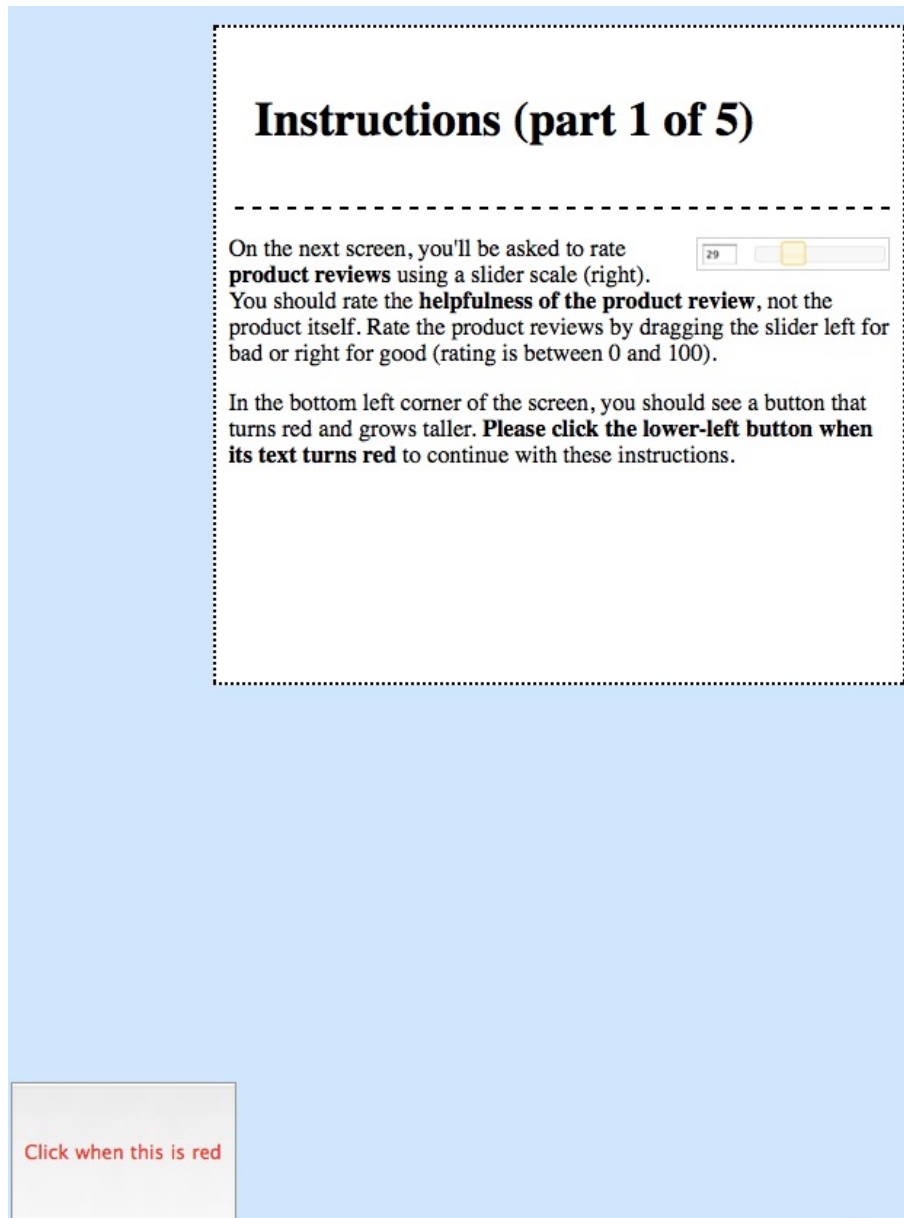


Figure 2: First set of instructions presented to users

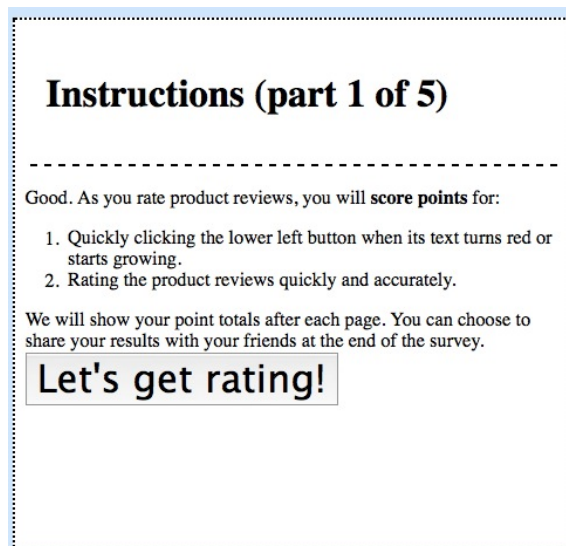


Figure 3: Second part of instructions presented to users before each treatment.



**Rate these movies**

Please rate the following 20 movies. Remember:

- You get points for clicking the lower left button whenever its text turns red or starts growing.
- You get points for rating the movies quickly and accurately.
- Skip any movies you don't like or don't know.

Ransom (1996)  Like this movie item 1 / 20

The 40 Year Old Virgin (2005)  Like this movie item 2 / 20

Analyze This (1999)  Like this movie item 3 / 20

Click when this is red

Figure 4: Movies for users to rate using the unary scale. Movie posters and plots redacted. Please see the hard copy in the Dewitt Wallace Library at Macalester College for original screenshot.

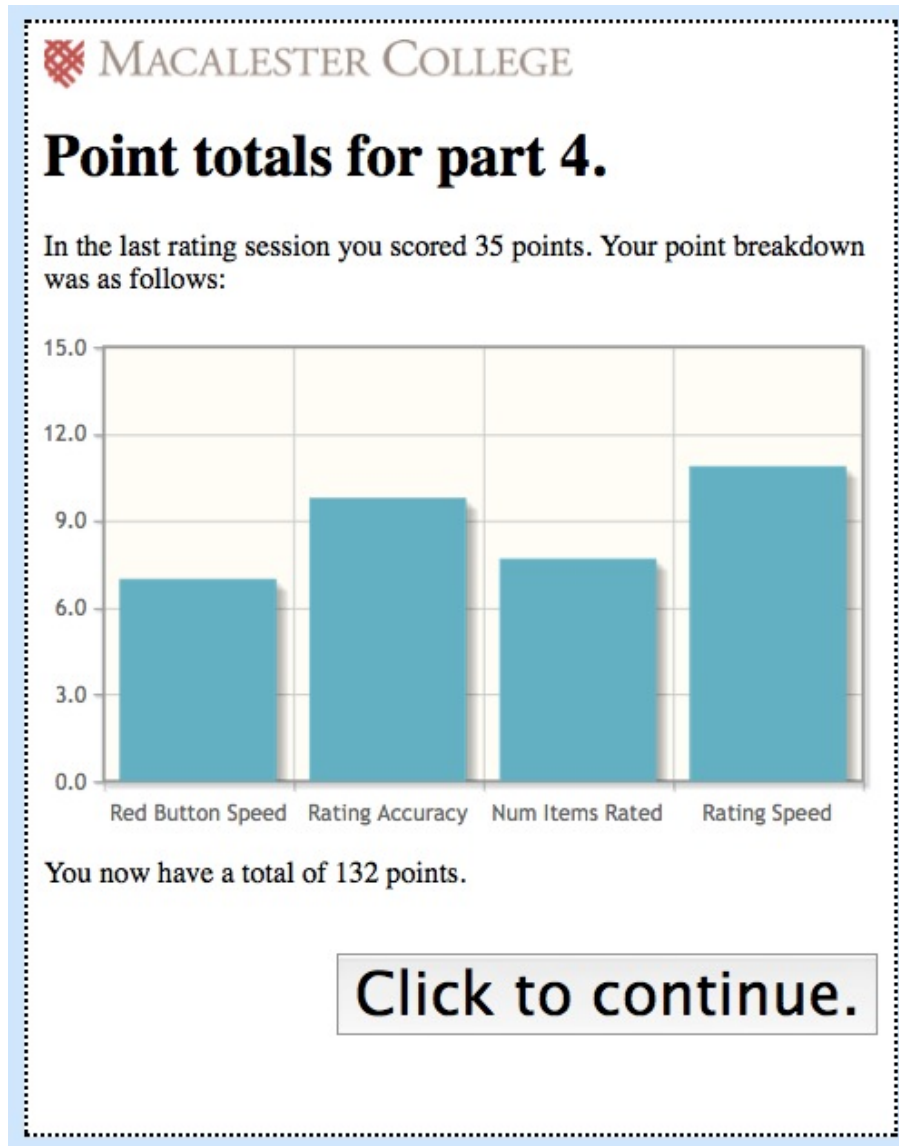



Figure 5: Inter-treatment point visualization.

### Survey Questions

We would like to ask you questions about each of the four pages of ratings you completed.

---

1. In the first page you rated product reviews using the slider scale: 


Please rate your agreement with the following statement:  
**Overall, I liked using the slider to rate product reviews .**

strongly disagree  disagree  neutral  agree  strongly agree

Do you have any other comments about using the slider to rate product reviews ?

Too many options!

---

2. In the second page you rated product reviews using the thumbs up / thumbs down scale: 

Please rate your agreement with the following statement:  
**Overall, I liked using the thumbs up / thumbs down to rate product reviews .**

strongly disagree  disagree  neutral  agree  strongly agree

Do you have any other comments about using the thumbs up / thumbs down to rate product reviews ?

Some product reviews are good; some are bad.

---

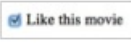

3. In the third page you rated movies using the unary (like it) scale: 

Figure 6: Reflection questionnaire, taken after rating on all four scales.

 MACALESTER COLLEGE

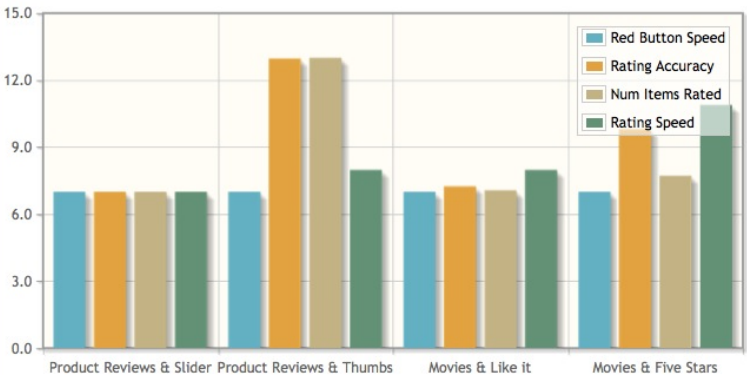
**Share your points:**

If you would like to share your success and ratings with your friends:

1. Enter the name by which you'd like to be known:
2. Click "Share with my friends!" to post to your Facebook wall. [Share with my friends!](#)

**Congratulations!**

You earned 153 points for taking this survey. Below is a detailed breakdown of where you scored what:



Section	Red Button Speed	Rating Accuracy	Num Items Rated	Rating Speed
Product Reviews & Slider	7.0	7.0	7.0	7.0
Product Reviews & Thumbs	7.0	13.0	13.0	8.0
Movies & Like it	7.0	7.0	7.0	8.0
Movies & Five Stars	7.0	10.0	7.0	11.0

Once you have shared your score, or if you would rather not share it, please click below to sign out.

[Logout](#)

Figure 7: Final point tabulation and prompt to share via Facebook.

 **Sebi Cohn**

**Isaac's Honors' Survey**  
poliwiki.macalester.edu

Sebi scored 193 on this survey. Click to see a breakdown of the points and take the survey!

[February 4 at 2:14am](#) · [Comment](#) · [Like](#) · [Share](#)

Figure 8: Link to the survey, shared via a Facebook profile

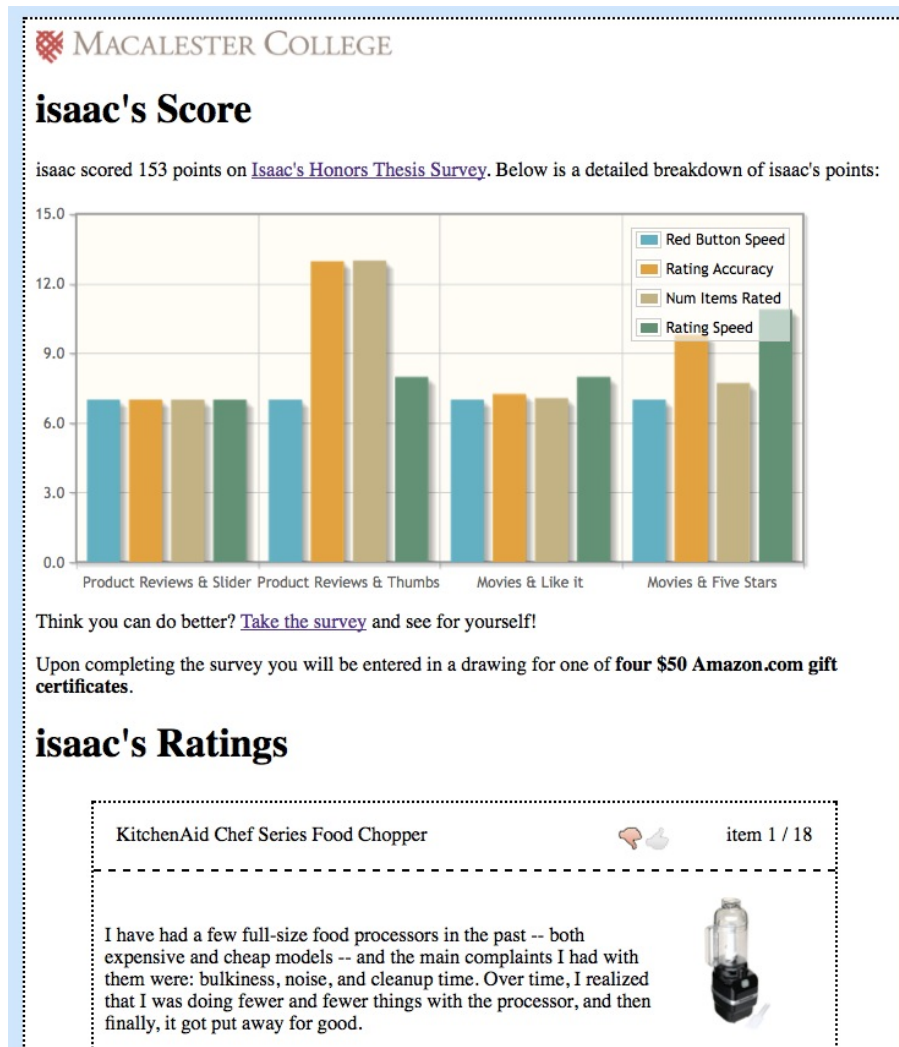


Figure 9: Optional public display of points and ratings

## 6 Experimental Design

### 6.1 Secondary Stimulus

We placed the button used to measure cognitive load in the bottom left corner of the screen. This button had several key features:

- It was uniformly located across all treatments.
- It contained instructional text which read "click when this is red."
- Every 0-20 seconds the button's text would slowly (over 4 seconds) turn red and then even more slowly (6.6 minutes) grow vertically to fill the screen.

The survey logged a timestamp when the button started morphing, and then again once the user clicked the button, restoring it to its original state. We ensured that users couldn't create a habit of clicking the button by randomizing (between 0 and 20 seconds) when the button would start changing again. We hoped to force users to avoid learning such a habit so we could maximize the efficacy of the secondary stimulus as a measure of cognitive load.

### 6.2 Incentives

Incentives are commonplace in the online environment. By providing an incentive to users, we could more accurately simulate the reality of online rating systems. However, equally, if not more, importantly, we could also more effectively challenge users to recruit their friends to take the study. If users felt that they were competing with either the amorphous public or their concrete friends, they would be incentivized to do a better job on all counts – rate more accurately, focus harder (and thus, provide a more effective cognitive load measure).

Assigning points was a relatively simple procedure. Users were given points for:

- Quickly rating items
- Accurately rating items
- Clicking the secondary measure in a timely fashion
- Completing the page
- Completing the questionnaire at the end in which they described their feelings about each scale

Assigning points for accurately rating an item (in addition to simply rating it) was more difficult than any of the other point-related tasks: how does one

measure accuracy for something which has no absolutely correct answer? After every treatment, users were asked to re-rate two items they had previously rated: the more closely the second rating matched the first, the more points they received. For the unary re-rating treatment, users were presented with one item they had rated and one they hadn't, so that they would have to choose to determine which they had previously liked.

After each treatment, users were presented with a graph showing the points they had accumulated and how they had accumulated those points. Once the survey was fully completed, users were given the option of posting their score to their Facebook wall, sharing their score with their friends and acquaintances. This was a key part of our strategy for recruiting a variety of new users to the survey, as well as fueling users' egos.

Additionally, both the point system and Facebook sharing brought in a fun factor. By making the survey more fun and interesting, we aimed to keep users attention more closely.

### 6.3 Pilot Study

We ran a pilot study with 3 colleagues versed in both rating scales and experimental design. We asked that they complete the survey and reflect on the experience of the *survey*. Was it effective? Frustrating? Did the content marry with the presentation?

The pilot study contained a system which we had hoped to use to control and measure exactly what a user was considering at any given point in time. All items on the page were grayed out save the one currently containing the mouse. This was meant to be a stand-in for the more complicated (and vastly more expensive) method of tracking a user's attention: an eye-tracking machine. Our reviewers unanimously hated this system: we had hoped that this mouse-over highlighting would be a natural interaction for users, but were clearly wrong. We did, however, continue logging the mouse-in and mouse-out times for all items for use in relation to rating time, as well as the potential to recreate a user's mousing experience.

We had initially presented users with 10 product reviews, a number that we reduced when our pilot study members all said that the survey became tedious at times. We reduced this number to 7 reviews. The notes about tedium also inspired us to incorporate points, in an attempt to make the survey more engaging.

## 6.4 Survey Implementation

The survey was implemented on a Groovy-on-Grails stack with dynamic content written in Javascript (primarily using the jQuery<sup>1</sup> library). Point related graphs were generated on the fly using a the jQuery plugin jqPlot<sup>2</sup>. The five star scale was an implementation built by Fyneworks<sup>3</sup>.

## 7 Results

Of the 430 people who began the survey, 348 completed it. The survey ran for the month of February, 2010 – the first user took it on the 2/1/2010, and the last analyzed user took it on the 2/28/2010. Figure 10 shows the ages of survey takers, as well as the distribution between male and female takers. Figure 11 shows users' frequency of internet activity. This final statistic is remarkable, as the majority of survey takers have significant experience online: of 430 users, 253 use the internet between 2 and 6 hours per day. Only 19 go online one or fewer times per day.

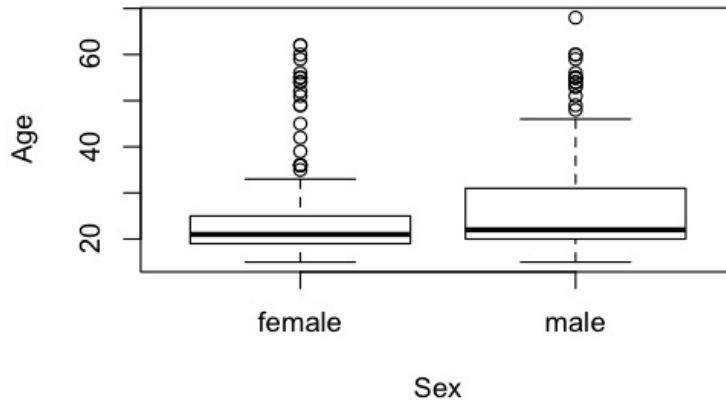


Figure 10: Sex and age distribution of survey takers. Overall mean=25.8 years, overall media=21 years.

<sup>1</sup><http://www.jquery.com>

<sup>2</sup><http://www.jqplot.com>

<sup>3</sup><http://www.fyneworks.com/jquery/star-rating/>



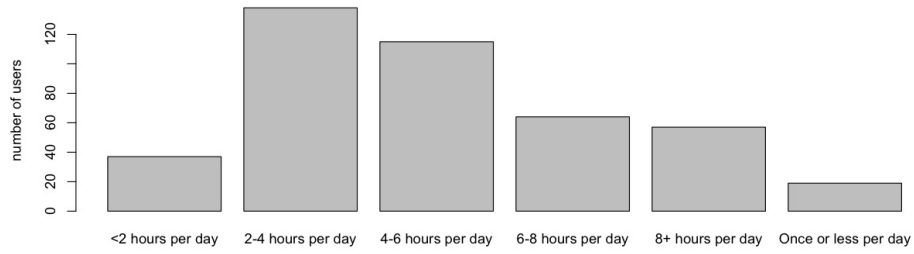


Figure 11: Frequency of internet use of users

Scale	Total Ratings	Positive	Negative	Percent Diff pos/neg
Slider	4227	2344	1883	24%
Five Star	3821	1574	1207	30%
Thumbs	4027	2508	1519	65%
Unary	4025 <sup>4</sup>	1894	2131	-12%

Table 2: Distribution of positive vs negative ratings.

## 7.1 Distribution of Ratings by Scale

Users had different distributions of ratings on each scale, a finding which agrees with Cosley et al. (2003). Generally, users tended towards the positive. Table 2 shows the breakdown of positively leaning reviews. We consider three out of five stars to be neutral, and we count it neither towards positive nor negative totals. Calculating the percent of positive unary ratings against negative unary ratings isn't a readily apparent procedure: one of the major drawbacks of the unary scale is that there is no way to tell disinterest from dislike. We can, however, extrapolate from the other three scales to make an estimate. We average the total number of ratings for the other three scales to get an estimate of the total ratings for unary and subtract off the number of positive ratings we have empirically measured. This brings us to an estimate of the number of dislike ratings (differentiated from disinterest). We know 45 of these ratings are sure: in 45 separate cases, users had checked "Like," and went back to uncheck it, indicated a dislike of the item.

Figures 12 through 15 detail distributions of each discrete rating event for each scale. Figure 16 shows where each rating maps to on the 100 point slider. This was calculated by taking a cumulative distribution of ratings on each scale and marking where it intersected with the cumulative distribution function for the slider. To estimate for unary, we used the above analysis to estimate the number of "dislike" ratings.

---

<sup>4</sup>The total number (positive and negative) of unary ratings is, by nature of the scale, an estimate. It follows that the number of negative ratings, and therefore, the percent difference between positive and negative are all estimates as well.

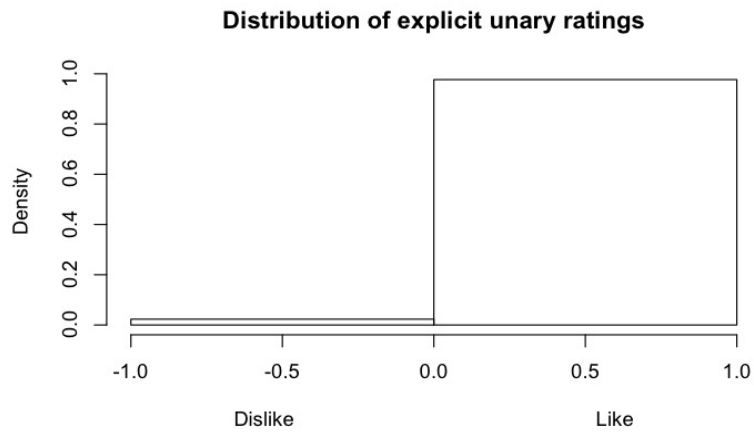


Figure 12: Distribution of unary ratings

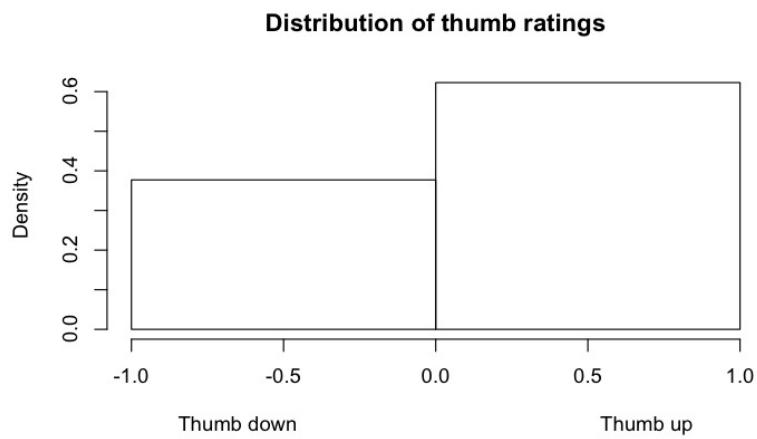


Figure 13: Distribution of thumbs ratings

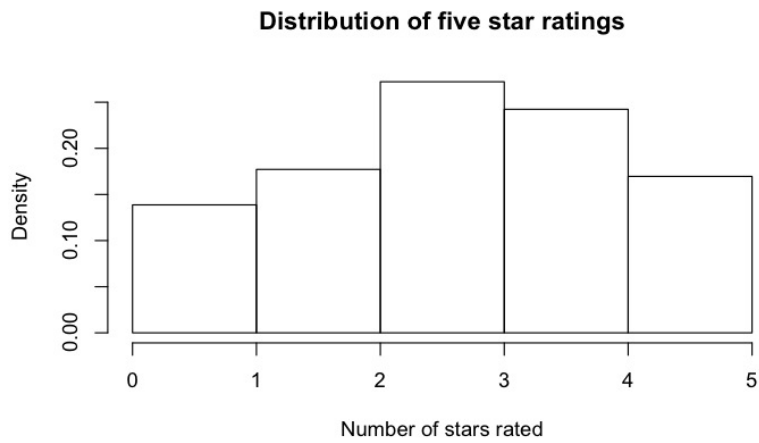


Figure 14: Distribution of five star ratings

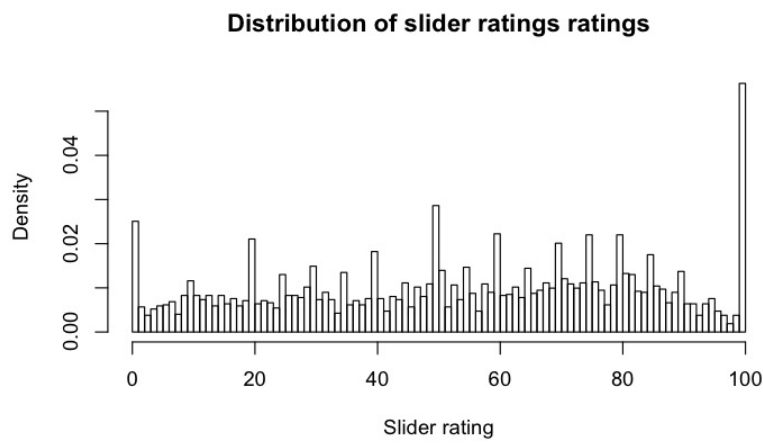


Figure 15: Distribution of slider ratings

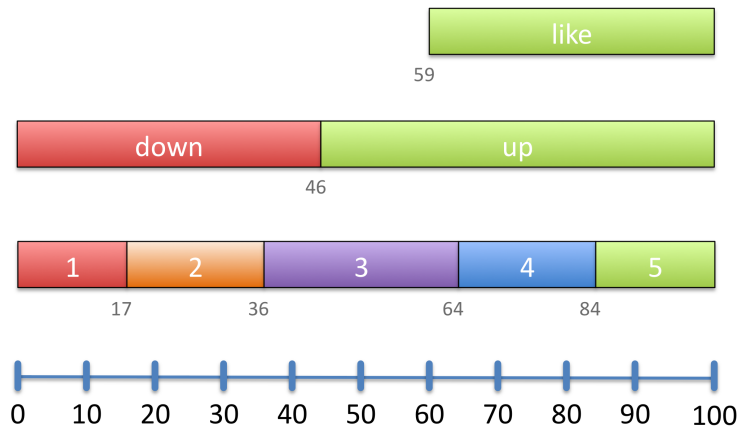


Figure 16: Relative distributions of ratings, showing where the cutoffs for each rating live relative to the 100 point scale.

## 7.2 Metrics

The data generated falls into four main categories:

- Page completion time: the total time a user spends rating a single page of items.
- Rating times: the time a user spends rating individual items. This is measured in two ways: time from mouse-in to rating and time between ratings (or "inter-rating time").
- Mouse-over durations: the time between mouse-in and mouse-out for a particular item.
- Secondary stimulus response times: the amount of time it takes a user to respond to the secondary stimulus.

The first three measures all measure, in different ways, the speed with which a user rates an item or a collection of items. The main reason we chose these measures is that we felt each measured an important aspect of the rating experience. When taken together, the three measures allow us to get a multifaceted understanding of where users were fast and where they were slow. The last measure is focused entirely on capturing the cognitive load of a user.

Data collected was noisy: it was clear that several users had simply gotten up and left their computer part of the way through a page (as evidenced by 15+ minute page completion times). These outliers were removed by defining a threshold over which all data would be trimmed. Table 3 clarifies how much data was removed as outliers.

Since the data we collected were noisy, we had to minimize the effect of the outliers by removing them. Table 3 shows the amount of data removed from our collected samples to remove the most drastic outliers. A goal of removing 1.75% of data was aimed for, though if reasonable models could be built with less data removal (as the case with page completion times and rating-rating times), we aimed to remove less data. We used different thresholds for the objective and subjective domains with the inter-rating speed measure, as the amount of time we intended for users to spend rating a particular item was heavily dependent on the domain of the item. Removing outliers for the slower (objective) wouldn't clean up the data for the faster (subjective).

We started our data analysis with a simple approach by building a linear model describing observed times minus the global mean in terms of both scale and domain for each measure. We then built a second model which allowed for random effects in both the treatment ordering and users.

This is a logical step to take, as some users will read reviews or recall movies faster than others. Once rough significance is shown, these per-user effects

Measure	Threshold (ms)	Points Removed	Total Points	Percentage
Page	290000	18	1456	1.2
Secondary	39500	128	7414	1.7
Item Duration	30500	881	51129	1.7
Rating	49000	250	14014	1.8
Inter-rating (Subj)	22500	159	9042	1.8
Inter-rating (Obj)	62000	86	4895	1.8

Table 3: Outliers removed; 1.8% of data or less removed for every measure.

should be minimized. The same goes for treatments, though with a slightly different line of logic. As a user progresses through the survey, they learn the particulars, and can complete tasks more quickly.

We followed this approach for all measures, first generating a simplistic model and then elaborating on it. The primary results of both models are reported and discussed in following sections.

### 7.3 Correlations Between Measures

Table 4 shows the general correlations between measures: we see that the secondary measure is the least correlated with the other measures. The correlation between page completion times and inter-rating speeds is among the highest correlations we see (rating speed-mouseover times is comparably high).

All measures found that the most significant differences to be seen were the differences between the slider and the unary scale. The slider tended to be the slowest scale to rate with, while unary tended to be fastest. This comes as no surprise. Both slider and unary tended to be significantly different from thumbs and five star, as well as each other, but thumbs and five star were never significantly different from each other.

For all measures, the second model shows that some users are significantly faster than others. This also comes as no surprise, as given the broad demographics of the survey, we are examining people with different levels of familiarity with rating scales as well as cognitive capacities.

### 7.4 Page Speed

Figure 17 shows the distribution of page completion times, measured in milliseconds.

	Page	Mouseover	Rating	Inter-rating
Page Duration	–	–	–	–
Mouseover Duration	0.3164	–	–	–
Rating Speed	0.3157	0.5882	–	–
Inter-rating Speed	0.5911	0.2902	0.2881	–
Secondary Speed	0.02662	0.23054	0.1472	0.0476

Table 4: Correlations between measures. Redundant data not shown.

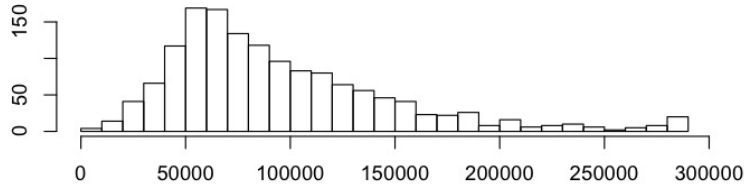


Figure 17: Distribution of page completion times in ms.

Examining the first, simpler, model, we found that page completion times are affected significantly by both scale and domain. When rating in the subjective domain, users complete a page approximately 27 seconds faster than in the objective domain ( $p < 0.01$ ). Slider, five star, and thumbs all cause the user to be slower than the average page completion speed (approximately 30, 15 and 12 seconds respectively), whereas unary allows users to complete a page more quickly (approx. 3 seconds). The unary decrease in speed was shown to be not significant ( $p = 0.28$ ). Slider, five star, thumbs and subjective’s  $p$ -values were all  $< 0.01$ . See Table 5 for more details.

The second model brings to light a different view: treatment number has a significant impact on page completion times. Slider, five star and thumbs all had similar (though more pronounced) effect—they cause completion times to be 59, 47 and 43 seconds longer respectively, and with lower significance. Unary turned around, and, like the other scales also showed an increasing page completion speed, but with higher significance than the simple model gave it. Subjective was

Scale/Domain	Estimate	Std Err	p-value
Slider	29820	2998	$< 0.01$
Five Star	15319	3024	$< 0.01$
Thumbs	12616	2984	$< 0.01$
Unary	-3268	3012	$= 0.28$
Subj	-27603	2691	$< 0.01$

Table 5: Model Summary without random effects for page completion times (ms) after subtracting the global mean (global mean = 95188ms)



Scale/Domain/Treatment	Estimate	Std Err	p-value
Slider	59000	19000	< 0.01
Five Star	46000	19000	= 0.02
Thumbs	42000	19000	= 0.03
Unary	29000	19000	= 0.13
Subj	-26000	2000	< 0.01
Treatment 2	-19200	2800	< 0.01
Treatment 3	-29700	2800	< 0.01
Treatment 4	-34600	2900	< 0.01

Table 6: Model Summary with random effects for page completion times (ms) after subtracting the global mean (global mean = 95188.05ms)

	Slider	Five Star	Thumbs
Five Star	66.4%	–	–
Thumbs	60.3%	44.2%	–
Unary	74.5%	63.8%	65.3%

Table 7: Percentage of page durations where the page duration on the y-axis is lower than the page duration the x-axis. Redundant data not shown.

had both a similar effect and significance. Treatments were important though: we see clear learning present, as users complete each treatment faster than the previous one, with high significance. The second treatment was completed approximately 19 seconds faster than the first. The third, approximately 10 seconds faster than the second, and the fourth approximately 5 seconds faster than the third. Table 6 shows this in detail.

Because the page measurement looked promising, we explored further, and performed a pairwise analysis on it, where we paired different scales and ignored domains. Table 7 lists the results of this analysis. We compared all users who had rated with two particular scales on the same domain (ie, a user who had done unary obj and slider obj). We then took the difference each user’s average page completion time for each two such paired scales, and calculated the percentage of users who were faster on one scale versus the other. This test does two things: it controls for differences in domain (as each user examined used the same domain for the scales examined) as well as controlling for differences between users. This analysis confirms that the most significant difference is between unary and slider, and that thumbs and five star are hardly different.

## 7.5 Rating Speed

The first model shows that mouse-in-to-rating speed might be significantly modeled by scale and domain. However, like with page completion times, incorporat-

ing random effects for both users and treatment ordering show that treatment ordering is significant. With this more complex model, the only scale that had a significant effect was the 100 point slider: it slowed rating times by 4 seconds, with  $p = 0.01$ . Users are approximately 2.3 seconds faster at rating a subjective item than an objective one ( $p < 0.01$ ). The largest jump in times on a treatment-by-treatment basis is between the first and second treatments, where users decrease rating times by 0.7 seconds ( $p < 0.01$ ). For the following two treatments, speeds decreased by a further 0.2 seconds each time ( $p < 0.01$ ). See table 8 and table 9 for complete data.

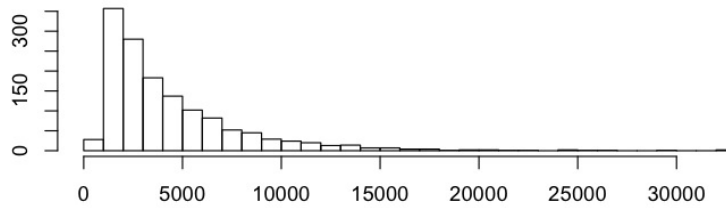


Figure 18: Distribution of rating times measured in milliseconds.

Figure 18 shows the distribution of rating speeds in milliseconds.

Scale/Domain	Estimate	Std Err	p-value
Five Star	865.82	206.62	< 0.01
Slider	3104.92	197.59	< 0.01
Thumbs	682.44	197.47	< 0.01
Unary	-41.51	200.28	= 0.84
Subj	-2323.10	179.14	< 0.01

Table 8: Model summary without random effects for rating times (ms) after subtracting the global mean (global mean = 4393.439ms)

Scale/Domain/Treatment	Estimate	Std Err	p-value
Slider	3900	1548	= 0.01
Five Star	1800	1550	= 0.24
Thumbs	1500	1548	= 0.32
Unary	800	1550	= 0.59
Subj	-2300	167	< 0.01
Treatment 2	-700	233	< 0.01
Treatment 3	-900	235	< 0.01
Treatment 4	-1100	236	< 0.01

Table 9: Model summary with random effects for rating times (ms) after subtracting the global mean (global mean = 4390ms)

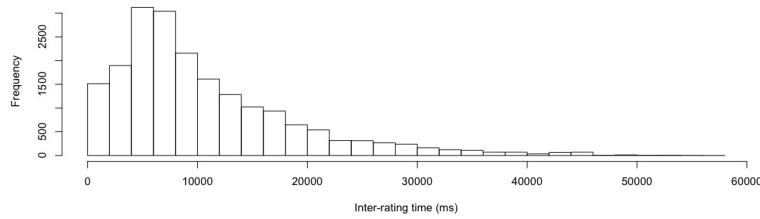


Figure 19: Distribution of inter-rating times measured in milliseconds.

Scale/Domain	Estimate	Std Err	p-value
Slider	3950	97	< 0.01
Five Star	3680	101	< 0.01
Thumbs	2920	96	< 0.01
Unary	8010	98	< 0.01
Subj	-9190	87	< 0.01

Table 10: Model summary without random effects for inter-rating times (ms) after subtracting the global mean (global mean = 10776ms)

## 7.6 Inter-Rating Speed

The simple model for inter-rating times leads us to believe that all scales and domains cause significant difference from the mean inter-rating time. All scales cause an increase in speed, but there is a notable difference from the other measures: unary has the highest inter-rating time (at 8.0 seconds ( $p < 0.01$ ), is about 4 seconds slower than the next fastest measure, slider, at 3.9 seconds ( $p < 0.01$ ). Reasons for this are discussed below, in section 8.2.

The model which took into account random effects by user and treatment shows this in a different light. It shows that five star and slider are almost indistinguishable, as both are 5.4 seconds slower than the global average ( $p < 0.01$  for both). Thumbs is slightly different, clocking in at only 4.5 seconds slower than the global average ( $p < 0.01$ ), and unary is the only significant effect among the scales, causing a slowdown of 11 seconds ( $p < 0.01$ ). This model retains the difference between objective items and subjective items as being significant: subjective items had an inter-rating time approximately 9.1 seconds shorter ( $p < 0.01$ ). Again, treatment order was significant. The second treatment saw a 2.1 second decrease in inter-rating times from the first treatment ( $p < 0.01$ ). The third saw an a 0.8 second decrease over the second treatment ( $p < 0.01$ ) and the fourth treatment had an additional 0.6 second decrease ( $p < 0.01$ ). See table 10 and table 11 for complete data.

Scale/Domain/Treatment	Estimate	Std Err	p-value
Five Star	5700	610	< 0.01
Slider	5600	610	< 0.01
Thumbs	4700	610	< 0.01
Unary	9900	610	< 0.01
Subj	-9200	66	< 0.01
Treatment 2	-2100	91	< 0.01
Treatment 3	-2900	92	< 0.01
Treatment 4	-3500	93	< 0.01

Table 11: Model summary with random effects for inter-rating times (ms) after subtracting the global mean (global mean = 10776.93ms)

## 7.7 Secondary Measure Speed

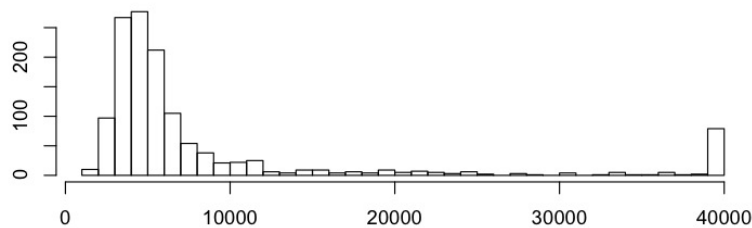


Figure 20: Distribution of secondary times measured in milliseconds.

Figure 20 shows the distribution of secondary measure speeds, measured in milliseconds.

The first model doesn't give much hope for showing that cognitive load varies based on scale and domain. It shows the general ordering of scales (in order of decreasing time to click the secondary measure: slider: 256ms above the mean; thumbs: 161ms above the mean; five star: 118ms above the mean; unary: 210ms below the mean). Subjective items cause secondary times to decrease by 164ms as compared to objective items. None of these values had significant  $p$ -values (the lowest, for the slider, was  $p = 0.66$ ). The standard errors were over twice the size of the estimate for all samples, again pointing towards little (if any) meaningful correlation. See table 13 for detailed data.

Additional analysis offers some more insight: by modeling speed simply in terms of rating scale, we see a rough ordering of speeds: unary response times are faster than average, and slider response times are slower. Thumbs and five star fall in the middle. None of these values are significant, but do suggest that with a more careful measurement of cognitive load, significance might be shown. See table 12 for data.

Scale	Estimate	Std Error	$p$ -value
Five Star	8520	520	< 0.01
Slider	8660	523	< 0.01
Thumbs	8560	526	< 0.01
Unary	8190	526	< 0.01

Table 12: Model summary with just scales for secondary response times (ms)

Scale/Domain	Estimate	Std Err	$p$ -value
Slider	256	580	= 0.66
Five Star	118	590	= 0.84
Thumbs	161	580	= 0.782
Unary	-210	580	= 0.719
Subj	-164	520	= 0.754

Table 13: Model summary without random effects for secondary response times (ms) after subtracting the global mean (global mean = 8480)

The more complex model, which takes into account random effects for users and treatments, shows significance in decreasing times being most strongly correlated with treatments: as a user progresses through the survey, they get slightly faster at recognizing the secondary scale, by about 0.7 seconds each time. While the  $p$ -values of the scales indicate significance, a more careful understanding shows that each is important in describing a difference from the global mean, but no scale is significantly different from the other. Subjective items have a slightly faster response time, but again, not significantly so ( $p=0.54$ ). See table 14 for detailed data.

Scale/Domain/Treatment	Estimate	Std Err	$p$ -value
Five Star	9510	3423	< 0.01
Slider	9590	3419	< 0.01
Thumbs	9420	3419	< 0.01
Unary	8910	3421	< 0.01
Subj	-232	384	= 0.54
Treatment 2	-5170	539	< 0.01
Treatment 3	-5830	542	< 0.01
Treatment 4	-6440	546	< 0.01

Table 14: Model summary with random effects for secondary response times (ms) after subtracting the global mean (global mean = 8486.848)

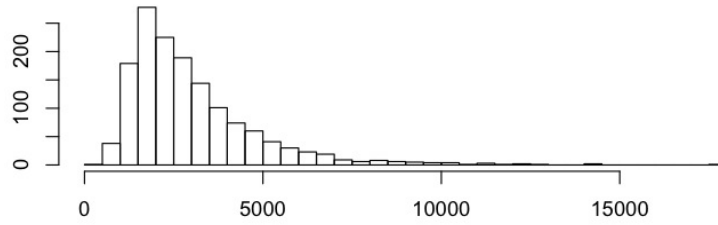


Figure 21: A brief overview of mouseover durations.

Scale/Domain	Estimate	Std Err	p-value
Slider	1150	99	< 0.01
Five Star	703	100	< 0.01
Thumbs	616	98	< 0.01
Unary	312	99	< 0.01
Subj	-1404	89	< 0.01

Table 15: Model summary without random effects for item durations (ms) after subtracting the global mean (global mean = 3004.428ms)

## 7.8 Item Duration

The first model for item duration follows the established pattern: both scales and domain appear to be significant, with slider being the slowest (1.1 seconds,  $p < 0.01$ ), unary being the fastest (0.3 seconds,  $p < 0.01$ ), with five star (0.7 seconds,  $p < 0.01$ ) and thumbs (0.6 seconds,  $p < 0.01$ ) falling between, and the subjective domain being faster (1.4 seconds,  $p < 0.01$ ) than the objective domain. Table 15 contains the complete data.

The second model shows similar results, but, again, also shows that treatments have an effect as well. When rating with the slider, users spend 2.1 seconds ( $p < 0.01$ ) moused over an item. Five stars are next slowest, with users spending 1.7 seconds ( $p = 0.01$ ), followed by thumbs, where users spend 1.6 seconds ( $p = 0.02$ ) and finally unary, at 1.4 seconds ( $p = 0.05$ ). When rating subjective items, users spend 1.3 fewer seconds ( $p < 0.01$ ) moused over an item. In the second treatment, users spend 0.4 seconds ( $p < 0.01$ ) less moused over an item than they spent in the first treatment. The second treatment sees a speedup of an additional 0.4 seconds ( $p < 0.01$ ), and the third an additional 0.1 seconds ( $p < 0.01$ ). Table 16 contains the complete data.

Figure 21 shows the distribution of mouse-over durations, measured in milliseconds.

Scale/Domain/Treatment	Estimate	Std Err	p-value
Slider	2090	684	< 0.01
Five Star	1750	684	= 0.01
Thumbs	1610	684	= 0.02
Unary	1360	684	= 0.05
Subj	-1340	72	< 0.01
Treatment 2	-480	100	< 0.01
Treatment 3	-800	102	< 0.01
Treatment 4	-902	102	< 0.01

Table 16: Model summary with random effects for item durations (ms) after subtracting the global mean (global mean = 3004.428ms)

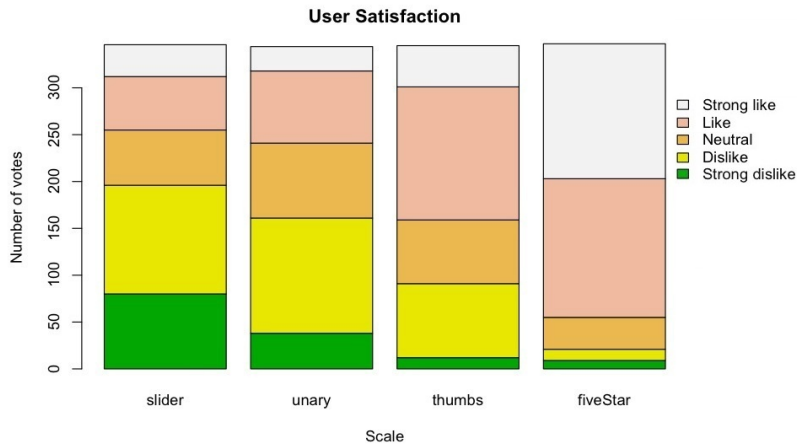


Figure 22: User satisfaction across all scales, disregarding domain. Each vertical segment is the proportion of ratings that a given scale received.

## 7.9 User Satisfaction

Users were most satisfied with the five star scale, and were least satisfied with the slider. Figure 22 shows how user's reactions relate to scale, where '0' is a strong dislike of the scale and a '5' is a strong like of the scale. The y-axis is broken down into percentages of total votes for a particular level of liking.

## 8 Discussion

### 8.1 RQ1

The data does not support RQ1 (“*Does cognitive load vary with scale, domain, or both?*”). A simple box-and-whisker plot (Figure 26, page 41) visually shows that there doesn’t appear to be any significant difference in cognitive load between treatments. Building a linear model of the times as relying on scale, domain and allowing for random effects based on both user and treatment number showed none of the independent variables significantly affect the speed with which users react to the secondary measure.

This model did show that some users were simply faster than other users, but there was no general relationship shown.

### 8.2 RQ2

While RQ1 was not supported, RQ2 (“*Does speed vary with scale, domain, or both?*”) is supported in some ways. The different speed metrics, and how they do and do not support RQ2, are discussed below.

#### 8.2.1 Page Completion Times

Examining the more complex model shows that, while scale might affect the page completion speed somewhat, which treatment the user was on had similar impact. We believe that this is due to the user learning how to more effectively take the survey as they progress through it. This leads us to believe that where speed is critical, designers should try to avoid over-granular scales. This feeling is echoed in users’ reflections about each scale. One user says about the slider, “There are far too many options, unless you are so picky you honestly feel differently at 75 and 76,” and they are far from alone in expressing this sentiment.

#### 8.2.2 Rating Speed

Whether or not the rating measure supports RQ2 depends on how we build the rating measure: “rating time” is taken to be the time from mouse-in of an item to rating, the data supports RQ2. However, calculating time between ratings leads us to believe that inter-rating time doesn’t vary at all. Most notable of this data (particularly as visualized in fig 25) is the spike in times for the unary scale and objective (product review) treatment. We feel that this spike is present for the objective items but not the subjective ones is due to the fact



that a thorough user reads the whole product review (and potentially sinks a large amount of time into doing so) before rating. If they decide that they don't like the review, that time is lost (and can't be accounted for, as the point when a user is cognitively done with an item isn't easily measurable with the unary scale). This doesn't affect the subjective domain as heavily, as users have a faster, more visceral reaction to movies. They don't have to spend time evaluating a movie to decide whether or not they liked it, whereas they have to read a significant portion of a product review before deciding to rate it or move on. Additionally, fewer product reviews were presented to the user: each one took a longer portion of their attention on the page, and so not rating it (i.e. not like it) causes a larger portion of their time to go unaccounted for.

Measuring rating as the difference in time between mouse in and the moment of rating does not show any significant relationship beyond the treatment relationship: as users progress through the survey, they get faster at completing the tasks. Insignificantly though, we observe that the typical order still holds, with the 100 point slider as the slowest, and unary as the fastest. This makes sense, as the slider requires both more mousing finesse to accurately move it to the desired position as well as more internal debate as to what, precisely, that position should be. The unary scale is simple: there is one point to click.

### 8.2.3 Mouse-over Duration

In examining the mouse-over durations, again, the treatment effect is clear. However, duration and  $p$ -value significance are correlated with decreasing scale granularity, again most notably in the edge cases, that is, for the slider and unary scales. This means that designers who want users to rate quickly should be quite wary of highly granular scales such as the slider. Moving downwards in granularity is generally a good idea, though mostly because moving *upwards* in granularity is a decidedly bad one.

## 8.3 RQ3

Our data shows that users feel strongly about the scales they use. Both the unary and slider are polarizing scales; some users love them and some hate them. Generally, users who liked the unary scale appreciated that it was simple and fast. Users who liked the slider felt they could accurately and precisely express their opinions with the scale. Those who like thumbs and five star liked them for a wide variety of reasons, including familiarity, perceived simplicity, and the ability to differentiate between good and bad discretely. This answers RQ3 rather simply: users like the five star scale best, if for a variety of reasons. This is backed up by users comments about the five star scale, which were generally positive, ranging from "Yay stars!" to "A manageable number of distinct values and a neutral value. What's not to like :-)".

In addition to examining how users felt about particular scales, we also consider the distribution of ratings users create with a particular scale, as seen in table 2 on page 21. Most notable in this table is that with the unary scale, we estimate that people rate positively *less* frequently than they would with other scales. This presents an interesting dilemma for sites like Facebook, which are looking to emphasize positivity among their users. By choosing a scale that only allows for positive expression, they actually decrease the positive contributions users might otherwise create.

This leads us to conclude that for general use, the unary scale is potentially quite dangerous from a content provider's perspective: providers forgo having concrete knowledge about users' dislikes, and decrease the number of ratings users provide. This is less critical if the unary scale is being used as a way to mark favorites, in conjunction with other more detailed scales being used for more typical rating tasks. By using the unary scale in conjunction with others, site operators also get past users' main gripe with the scale: an inability to express dislike.

## 9 Conclusions and Further Work

We have shown that the cognitive load hypothesis was not supported by the collected data, and that the speed hypothesis is vaguely supported by the data. The data has shown that the five star scale is the best liked scale.

To provide a reasonable recommendation of a best-practice for site operators looking to deploy a new rating system, we marry our findings regarding cognitive load, speed and user satisfaction as well as the findings of (Cosley et al., 2003).

Our three findings combine to let us compile a single recommendation for designers. Since we have found that there is no significant relationship between cognitive load and rating scale, we don't have to consider it. All we must do is balance our speed findings with our user satisfaction findings. The unary scale lacks the ability to collect as much data as the other scales do, and the slider takes significantly longer to use than the other scales. Additionally, users are polarized about these scales, and by choosing one or the other, a site operator would risk alienating a potential user base. This leaves us to decide between using a thumbs and five star system. Since we found there to be little statistically significant difference between the two scales in terms of speed, and user satisfaction was higher (in the two studied domains), we recommend the five star scale for use rating product reviews and/or movies.

Our recommendation aligns quite closely with that of Cosley et al. (2003), who observed that users were more satisfied with finer granularity (explaining why they prefer the five star scale to the thumbs). We have shown, however, that

this has a limit: users are quite dissatisfied with a highly granular scale such as the 100 point slider. We feel that the unary scale should probably be avoid for reasons involving both user satisfaction and implicit data collection. Its numerous drawbacks simply aren't worth the relatively minor decrease in rating times it affords.

## 9.1 Further Research

These results bring to mind several questions for further research.

- How do sites users frequent affect their familiarity with scales? e.g.: if a user's most visited site is Facebook, do they have different characteristics when interacting with the unary scale? (reddit/digg to the binary, netflix, apple to the five star, and some example to the slider). This would require another (or a followup) survey.
- Does internet-savviness affect cognitive load and/or speed? This could be calculated from collected data; significant further analysis would be required.
- Of the scales examined, we have found that five star is the best for rating product reviews and movies. What other rating domains does this extend to? Specifically, how would a five star scale fare in the social new sphere?

Exploring these questions might help site operators more effectively choose a scale for specific subsets of users.

## A Histograms

This appendix contains box and whisker plots showing 95 of the data, broken down by individual treatment for each measure discussed.

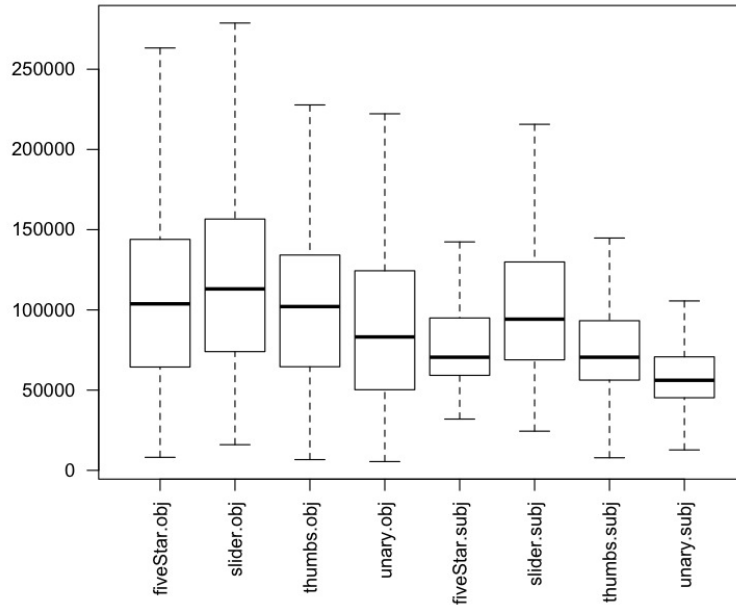


Figure 23: Page completion times across scale/domain interaction, outliers ignored (ms)

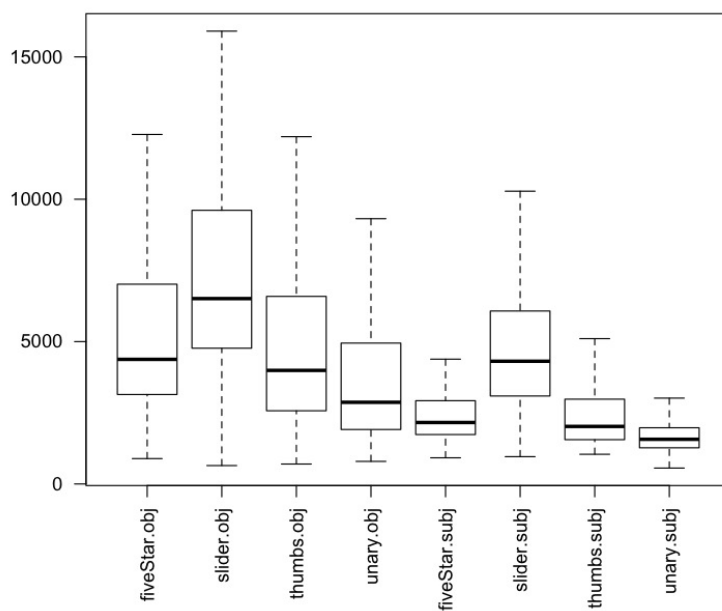


Figure 24: Time from mouse in to rating across scale/domain interaction, outliers ignored (ms)

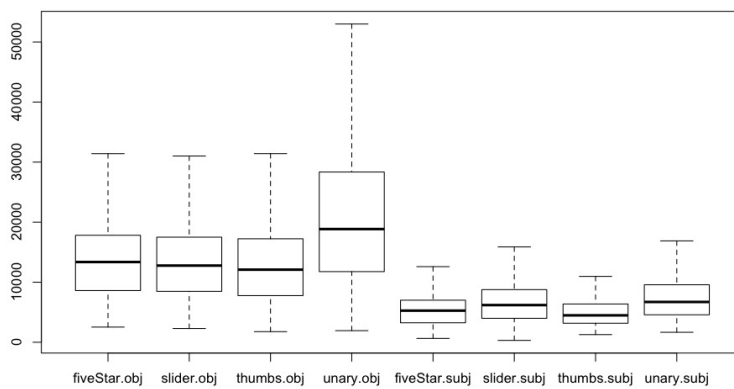


Figure 25: Time between ratings across scale/domain interaction, outliers ignored (ms)

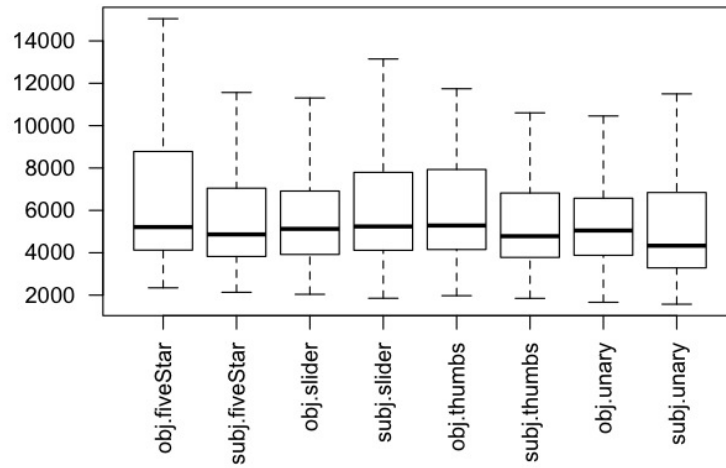


Figure 26: Secondary times across scale/domain interaction, outliers ignored (ms)

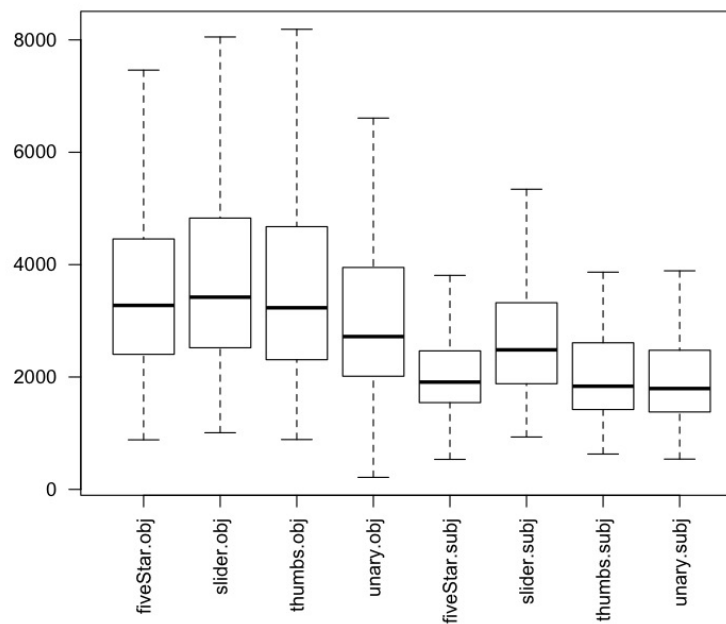


Figure 27: Time between mouse-in and mouse-out across scale/domain interaction, outliers ignored (ms)

## References

- Brunken, R., J. L. Plass, and D. Leutner (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist* 38(1), 53–61.
- Cosley, D., S. K. Lam, I. Albert, J. A. Konstan, and J. Riedl (2003). Is seeing believing?: how recommender system interfaces affect users' opinions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 585–592. ACM New York, NY, USA.
- Garner, W. R. (1960, November). Rating scales, discriminability, and information transmission. *Psychological Review* 67(1), 343–52.
- Guilford, J. P. (1938). The computation of psychological values from judgements in absolute categories. *Journal of Experimental Psychology* (22), 32–42.
- Harper, F. M., J. A. Konstan, X. Li, and Y. Chen (2005). User motivations and incentive structures in an online recommender system. In *Proceedings of Group 2005 Workshop on Sustaining Community: The role and design of incentive mechanisms in online systems*. Citeseer.
- Sen, S., F. M. Harper, A. LaPitz, and J. Riedl (2007). The quest for quality tags. In *Proceedings of Group 2007*. ACM New York, NY, USA.
- Sweller, J. (1999). *Instructional Design in Technical Areas*. Camberwell, Australia: ACER Press.