Faster Image Rating using a CAPTCHA Inspired Task

Siyi Fan

The University of Chicago, Chicago, Illinois

Introduction

The objective of this study is to propose a CAPTCHA-style image rating method and validate this new rating method against the conventional image rating methods. Participants were presented with twelve images simultaneously each trial and asked to click on “top 4” images in terms of preference. **Our research question is: Could CAPTCHA-style image rating method replace conventional image rating methods?** Specifically, we have two hypotheses:

Hypothesis1: CAPTCHA-style rating method would remain a consistency to the Likert scale rating method. That is, CAPTCHA-style image ratings will positively correlate with single-image style ratings collected previously by Likert scale.

Hypothesis2: CAPTCHA-style rating method would be more reliable than conventional Likert scale rating method.

Method

Simulation

To determine the required number of subjects (P*) and ideal number of images (N*) subjects should click on of the 12 images displayed on the screen at a time, the following computer simulation was conducted. First, we calculated the inter-rater correlation for the actual dataset, which was created by Environmental Neuroscience Lab at the University of Chicago. Next, we generated a simulated dataset with P (P=30, 60, 120) participants rating 1030 images and computed the inter-rater correlation for this simulated dataset. To mimic the idiosyncrasies of actual preference rating, “sbjweight” was added to the simulated preference rating for each image. We modified the “sbjweight” so that the inter-rater correlation for this simulated dataset was close to the inter-rater correlation in the actual dataset (simulated preference = actual preference + sbjweight * randn(size(actual preference))). Then, we shuffled the order of all images and sorted

N (N=2, 3, 4) images out of 12 images with the highest simulated preference ratings (simulated preference) displayed on each trial. Based on the simulated preference ratings, those selected N images would be clicked on in CAPTCHA-style. Therefore, we assigned point 1 to each sorted image and point 0 to all unsorted images, generating a new dataset with 1030 preference rating. Finally, we calculated (**Table 1**) and plotted (**Figure 1**) the correlations between the preference rating in the new dataset and actual preference rating. The correlation increased as the number of subjects increased, and the number of images clicked on each trial increased, reaching the highest value with P = 120 and N = 4. In this way, we could approximate the best parameters (P* = 120, N* = 4) and decrease sampling error.

Table 1.
Correlations of Simulated Preference Rating and Actual Preference Rating

	N = 2	**N = 3**	**N = 4**
P = 30	0.88	0.92	0.95
P = 60	0.90	0.94	0.96
P = 120	0.92	0.95	0.97

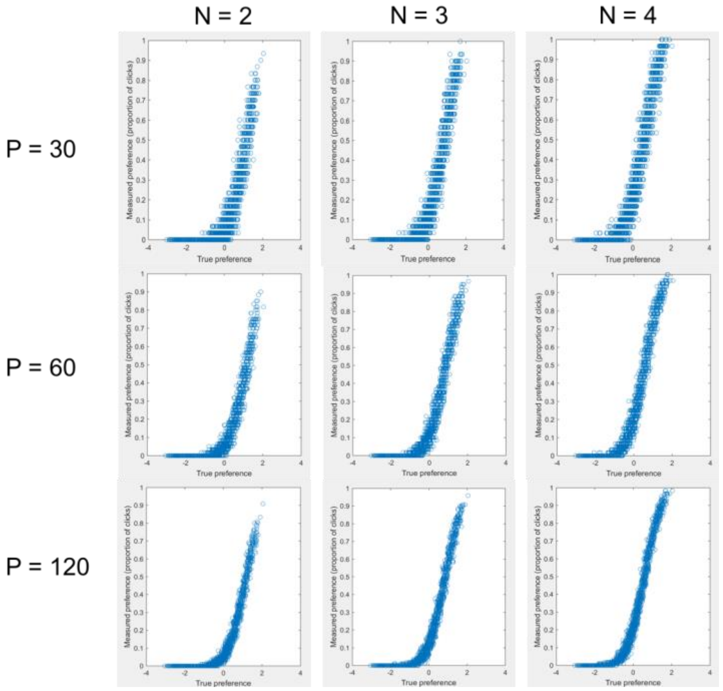


Figure 1. **The correlation between actual preference rating and simulated preference rating.** The true preference indicates the ratings using Likert scale. The measured preference indicates the average simulated ratings across all participants P.

Participants

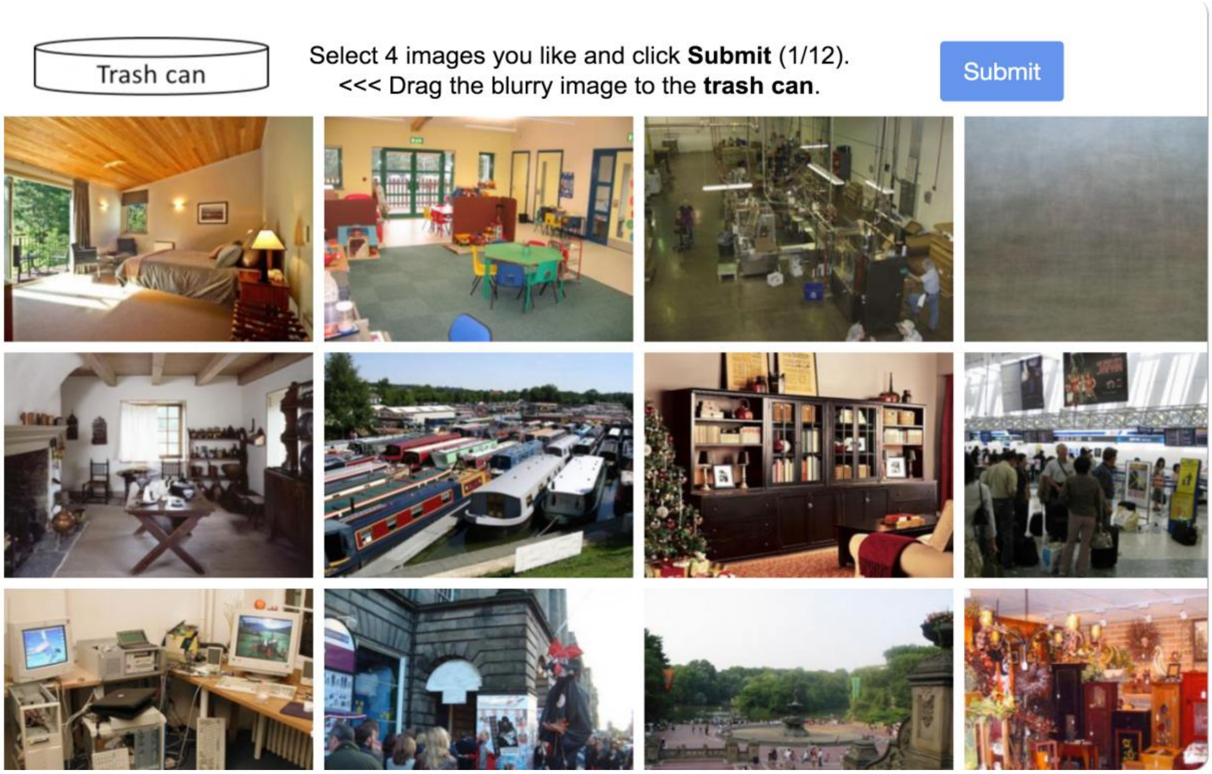
Volunteers participants were recruited using TurkPrime to complete the online study via the online labor market Amazon Mechanical Turk (AMT), in which workers perform human intelligence tasks (HITs) for requesters. The study was conducted during weekdays only and participation was allowed only between 8 a.m. to 8 p.m. CST. If participants failed more than 3 attention checks, the study session was terminated automatically. Participants were paid $2.00 upon completion of the task. AMT worker qualifications included location in the United States, a HIT approval rate greater than or equal to 90%, and the number of HITs approved greater than or equal to 5,000. Informed consent was administered by the Institutional Review Board (IRB) of the University of Chicago.

The target sample size was 120 participants, the optimal number of participants in the initial sample based on the results of a simulation. A total of 155 participants began the study online via AMT, but 29 participants did not complete all trials. Thus, a final sample of 126 participants was obtained.

Stimuli and Tasks.

Materials. 1030 images utilized in the current study were real-world environmental scenes selected from the Scene Understanding (SUN) image database (http://vision.princeton.edu/projects/2010/SUN/; Xiao, Hays, Ehinger, Oliva, & Torralba, 2010), which contains 899 categories and 130,519 images, diverse in semantic content (e.g., indoor, urban, and nature). In the online experiment, the 1030 full-color 280 * 210 pixel photographs of real-world, indoor and outdoor scenes were used.

Procedure. Participants viewed 1030 scene images in total during 87 trials, including 14 attention check trials. On each trial, 12 different scene images arranged in 3 discrete rows with random item placement were presented on a plain white background. Twelve was the optimal set of images that could be displayed simultaneously to the participants without overcrowding the display. After scanning 12 images on the screen, participants were instructed to click on the "top 4" images in terms of preference. Once an image got clicked on, it was assigned 1 point. Otherwise, it was assigned 0 points. The prompt "Select 4 images you like and click submit" was written above 12 images on each trial. For the attention check trials, subjects were told to drag the blurry image to the trash can before selecting 4 images they preferred. **Figure 2** shows the scheme of one attention check trial. Following the image rating task, participants responded to two surveys (1) Big Five Personality Survey and (2) Nature Intuitions Survey. Finally, participants were given a random code that they had to submit an AMT to verify their completion of the experiment.



Select 4 images you like and click **Submit** (1/12).
<<< Drag the blurry image to the **trash can**.

Figure 2. ***An example of an attention check trial.***

Data and Code Availability.

All statistical tests were performed using MATLAB R2016b. To investigate the factors affecting an image to be clicked or not, we conducted the Generalized Linear Model (GLM) analysis using the *fitglm* function from the Statistics and Machine Learning Toolbox in MATLAB. In the model, the binary dependent variable was "Click" (1) or "Not Click" (0), and the binomial distribution was utilized to specify this dependent variable. The *corr* function was used to calculate the effect size (95% CI) of Spearman's correlations, the *corrcoef* function was used to calculate the effect size (95% CI) of Pearson's correlations, and the *polyfit* function was used to calculate the 95% confidence band of a regression.

Results

Position bias. To examine what factors could influence an image to be click or not, we conducted a GLM analysis, in which 1030 images and 12 image positions were added to the model as predictors. In this analysis, the binary dependent variable was one image clicked or not (1: click, 0: not clicked). The predictors were the image (1, 2, …, 1030) and image position during each trial (1, 2, …, 12).

The model (*df* = 1027) explained 16.63% of the variance (adjusted R^2), and the results of image positions as predictors on preference are presented in **Table 2**. Of the 12 positions, only position 6 (the second row and the second column) significantly predicted preference (p < 0.05), which suggests that subjects tended to click on the image displayed on the second row and the second column no matter what that image was.

Table 2.
Image Positions as Predictors of Preference Rating

Image Position	**Coefficient**	**SE**	***t*-value**	***p*-value**
Intercept	-1.645	0.256	-6.434	1.243
2	-0.002	0.031	-0.062	0.951

3	-0.059	0.031	-1.903	0.057
4	-0.048	0.031	-1.530	0.126
5	-0.181	0.031	-5.743	9.292
6	-0.077	0.031	-2.479	0.013
7	-0.147	0.031	-4.687	2.769
8	-0.220	0.032	-6.986	2.823
9	-0.305	0.032	-9.619	6.637
10	-0.288	0.032	-9.106	8.547
11	-0.251	0.032	-7.959	1.741
12	-0.321	0.032	-10.118	4.606

Correlation. For the purpose of this study, we assessed whether CAPTCHA-style rating would remain a consistency to the Likert scale rating. Therefore, we correlated the estimated ratings (i.e., beta coefficients) generated by GLM and ratings collected previously by Likert scale. The correlation value between two rating methods were 0.7955 ($p < 0.001$), suggesting the GLM could be utilized in predicting scene preference for a large variety of images, and more importantly, CAPTCHA-style rating method reaches the same performance as Likert scale method does.

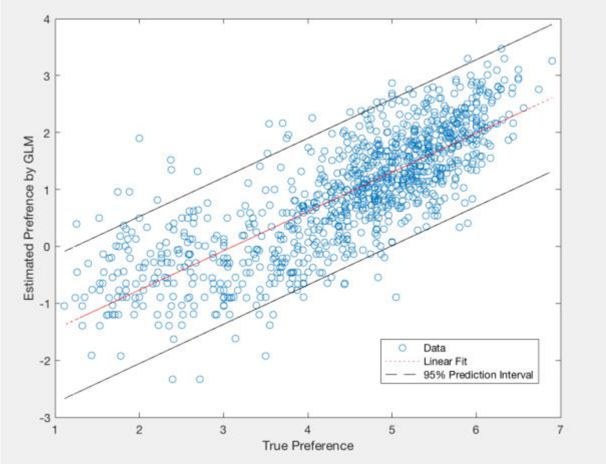


Figure 3. **Relationships between preference rating using Likert scale and predicted preference rating using CAPTCHA-style method.** The true preference indicates the ratings using Likert scale. The estimated preference indicates the predicted ratings using CAPTCHA-style method generated by GLM.

Reliability. (Likert scale, r = 0.9380, p = 3.1599 (not significant???); CAPTCHA-style method, r = 0.9312, p = 0). More analysis required…

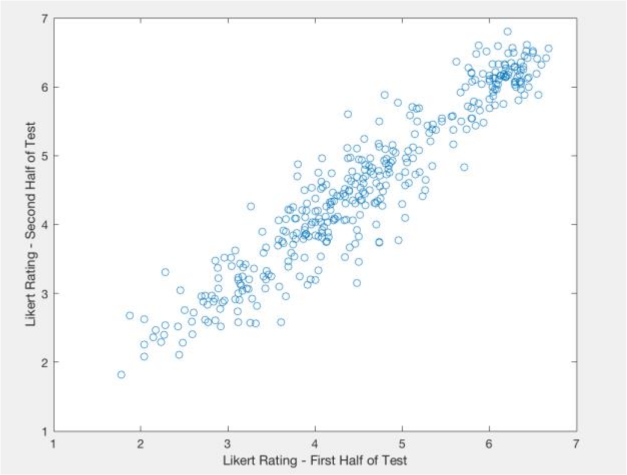


Figure 4. **Scatterplot of Likert ratings on the second half of the test plotted against ratings of the first of the test.**

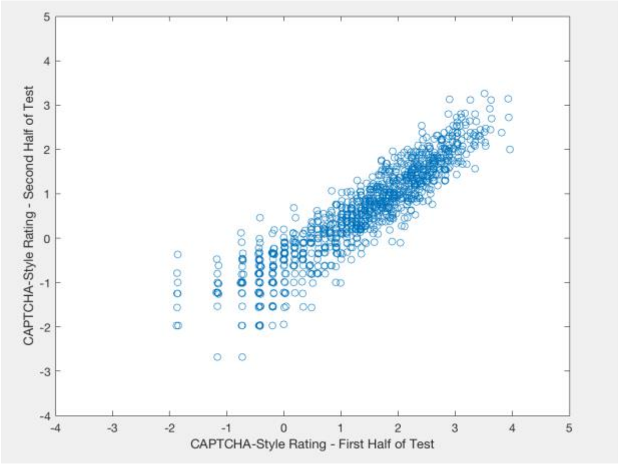


Figure 5. **Scatterplot of CAPTCHA-style ratings on the second half of the test plotted against ratings of the first of the test.**