Faster Image Rating Using a CAPTCHA Inspired Task

Siyi Fan

The University of Chicago, Chicago, Illinois

Faster Image Rating Using a CAPTCHA Inspired Tasks

Abstract

Single-image rating (e.g., Likert scale) is labor-intensive and expensive when building a large-scale image database. We presented a new method that allows researchers to build image database faster, in which subjects viewed twelve images simultaneously and clicked on four images they preferred most. We call this method CAPTCHA-style rating, which was inspired by clicked-based image CAPTCHA tasks. We illustrated our method by evaluating the rating time and the internal consistency reliability of this new method compared to the Likert scale. A Generalized Linear Model (GLM) was conducted to control for different images and image positions simultaneously. The output of model suggested that the probability of an image being clicked depended not only on its relevance, but on its position on the screen. The experiment demonstrated that CAPTCHA-style rating was faster and yielded perfect consistency to the Likert scale assessment but without significantly higher reliability. Therefore, the CAPTCHA-style rating could be used as an alternative to the Likert scale for large-scale dataset construction.

Keywords: CAPTCHA, image rating, reliability, deep learning, machine learning

1. Introduction

With the development of high capacity supervised learning algorithms, there has been notable progress in the performance of image detection and classification. Lagging behind the growth in computational modeling, the number of parameters in many deep models exceeds the number of images in these datasets. Additionally, training images are spread across many categories, which results in the density of image examples in current datasets being quite low (Yu et al., 2016). In order to deal with this lack of data, a massive amount of training data with manual labels are required to produce. The input data with ratings provided by human raters would be especially useful for training and testing purposes since they can provide insights into whether the local processing is correctly modeled or whether further adjustments are warranted.

The bottleneck to building a supervised image dataset with high density is the high cost of manual labeling. Traditionally, manual image ratings have only been carried out on Likert scales rating a single image or ranking scales comparing two images at the same time, which involves an enormous time commitment for both raters and experimenters to create the image database.

The objective of our research was to propose a faster image rating method inspired by clicked-based image CAPTCHA and evaluate the reliability of this alternative method that does not require either a Likert scale or pairwise comparison. We designed an experiment in which subjects would view twelve images simultaneously and click on four images they prefer most. Our study is unique in that we took image position into consideration and found out that the probability of one image being clicked was influenced not only by its relevance, but on its position on the screen. Our primary hypothesis was that CAPTCHA-style image rating method would be faster than Likert scale rating method. Our secondary hypothesis was that CAPTCHA-style rating would reach the same performance as Likert scale rating did but with higher internal consistency reliability.

1.1 Image Rating Methods

While judging preferred images is a subjective task, it has been an area of active study for a long time and substantial progress has been made in evaluating those images.

The first line of image rating task is to display a single image on the screen at a time and have participants rate that image using a Likert scale. The Likert scale is a commonly used assessment tool, which evaluates how participants feel about stimuli and within a specific classification. Kaplan et al. (1972) studied whether images of natural or urban environments will influence preference by presenting slides depicting non-spectacular, relatively local places to participants. For each slide, the participants were asked to rate the image on complexity and preference using a 5-point rating scale ranging from “not at all” to “a great deal”. The evaluation instrument could also be a list of bipolar adjective pairs. Martindale et al. (1990) examined aesthetic preference by instructing participants to rate images on attributes using a 7-point scale, including: complex-simple, meaningless-meaningful, like-dislike, orderly-disorderly, nonphotographic-photographic, natural-unnatural, and static-dynamic. Some recent studies reduce the size of the Likert scale. Berman et al. (2008) compared the restorative effects on cognitive functioning of interactions with natural versus urban environments. Picture viewing lasted approximately 10 min, during which participants rated on a scale of 1 to 3 how much they liked each picture.

An alternative method is a ranking scale. It allows participants to compare individual elements to each other and rank the choices in order of preference. A common ranking scale approach in image evaluation is pairwise comparison, where an individual compares two images side by side and chooses which image they prefer. Gur et al. (1997) compared five-category ordinal ratings of relative quality to a two-alternative forced-choice approach by presenting pairs of laser-printed images (labelled as “A” and “B”) to participants. For the five-category ordinal scale, participants were asked to rank order the relative differences between the two images ranging from “Image ‘A’ is much better than image ‘B’” to “Image ‘A’ is much worse than image ‘B’”. For the two-alternative forced-choice scale, participants were forced to choose one image in each pair (“A” or “B”) as the better image. The researchers found a significantly lower ability to detect small differences in perceived image quality when using a five-category ordinal scale and participants are highly sensitive to image differences during a side-by-side review.

Phelps et al. (2015) compared test time (Experiment 1) and inter-rater reliability (Experiment 2) of pairwise comparisons with that of a Likert scale in biomedical image assessment. Using a pairwise comparison test, two images were presented side- by-side on one slide and subjects were required to determine which one is sharper. Using a Likert scale, a single image was displayed on each slide and subjects assigned a sharpness score between 1 (least sharp) and 10 (sharpest) for that image. The results of Experiment 1 revealed that participants took an average of 3.2 seconds to evaluate an image for the pairwise comparison test, which was 0.4 seconds shorter than the Likert test. Moreover, the results of Experiment 2 showed that the inter-rater reliability of pairwise comparison test was higher than the reliability of Likert test.

These classic image rating methods have been applied successfully in order to answer psychological questions so that fifty years later, researchers are still using single-image and pairwise comparison formats to gather image ratings. However, these tasks have two major limitations. First, participants are required to spend a lot of time and energy to rate the wide array of images. The rating time of Likert scale increases linearly with the number of images, whereas the total number of required comparisons (n*(n-1)/2) rises exponentially with the number of images to be compared. When the number of images required to be rated is massive, the cost of manual labeling using either Likert scale or pairwise comparison will be tedious and time consuming. Second, the reliability of both methods is doubtful. Underreporting of the extremes of a Likert scale known as central tendency bias has been described in the market research (Albaum, 1997). People may also respond too quickly in a Likert test, simply assigning an extreme score (i.e., highest or lowest) to each of the images without fully reading them. For pairwise comparison, people are forced to make a choice between two items when they may actually perceive them equally. Therefore, the present research attempts to design a new image rating method to improve the efficiency of human-in-the-loop annotation and make the rating process more reliable.

1.2 The Present Research

In light of the time-consuming rating procedure and disreputable reliability for the two conventional rating methods, we decided to create a faster image rating format inspired by clicked-based image CAPTCHA with humans who view several images at the same time and click on their preferred choices. CAPTCHA (Completely Automated Public Turing Test to Tell Computers and Humans Apart) is a program that generates and grades tests that most humans can pass, but current computer programs fail (Von et al., 2003). Such a program has many applications for website security, including online polls, email services, search engine bots, worms and spam, and preventing dictionary attacks. Click-based image CAPTCHA requires users to semantically understand the images first and then perform an image recognition task following a specific on-screen instruction, such as: click all of the images with an orange. By using the similar way for our study, participants were asked to view 12 images simultaneously displayed during each trial and click on 4 images they preferred most. For each image, one click assigned one point to it.

We speculated that CAPTCHA-style rating would be faster than Likert scale because it does not require the retrieval and integration of detailed information from memory when calling for a simple evaluation. We also expected that the internal consistency reliability of CAPTCHA-style rating would be higher than the reliability of Likert scale because it avoids the central tendency and the tendency to rate in an extreme way the Likert scale may involve.

2. Experiment

This experiment first examined whether CAPTCHA-style image rating method would be faster than Likert scale rating method. It also investigated whether CAPTCHA-style rating would reach the same performance as Likert scale rating did but with higher reliability. Since preference ratings for each image do not vary significantly among the scales, it is expected that CAPTCHA-style preference rating will positively correlate with Likert scale preference rating on the same image.

2.1 Methods

2.1.1 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 (N_{female} = 78, N_{male} = 48; M_{age} = 48.5 years, SD_{age} = 16.7 years, range = 20 – 81 years).

The majority of participants identified as White (*N* = 102), with the remainder identifying as African-American (*N* = 10), Asian (*N* = 8), or Other race (*N* = 6).

2.1.2 Stimuli and Tasks.

Materials. The 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 1** 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. Results from these questions are reported in the descriptives below but are not the focus of the main analysis. Finally, participants were given a random code that they had to submit an AMT to verify their completion of the experiment.

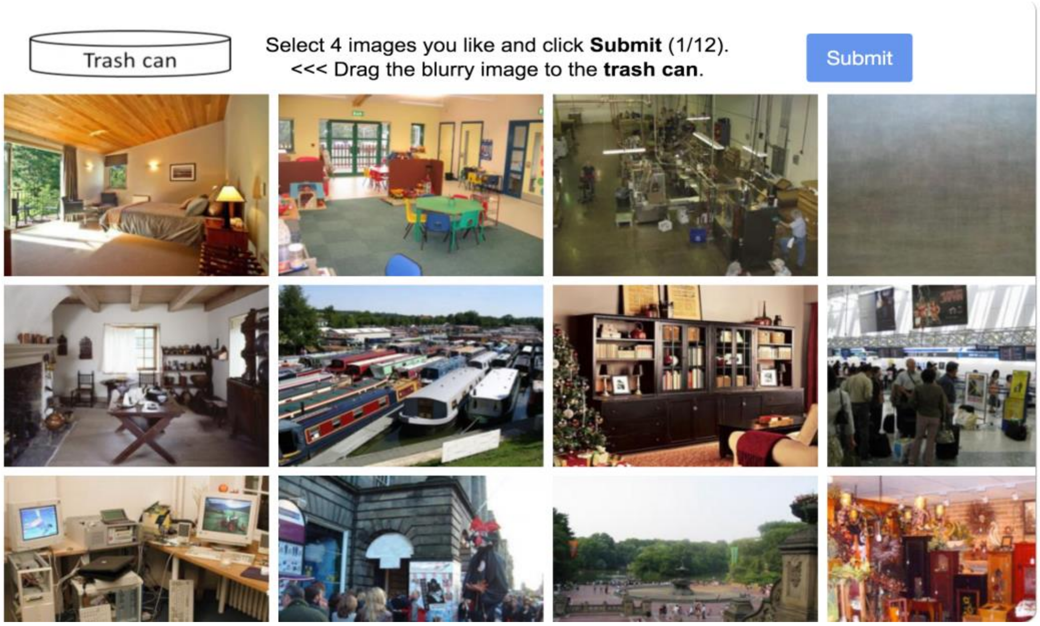
Select 4 images you like and click **Submit** (1/12).
<<< Drag the blurry image to the **trash can**.

Figure 1. ***An example of an attention check trial on preference rating.***

2.1.3 Data and Code Availability.

All statistical tests were performed using MATLAB R2016b, which were available at https://github.com/siyiii/fp-siyiii. 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.

2.2 Results

***Rating time*. Figure 2** compared the rating time of Likert rating and CAPTCHA-style rating across 126 participants. Results of a paired t-test indicated that the time used to rate each image by using CAPTCHA-style (*M* = 1.1705 , *SD* = 0.5445) is shorter than using Likert scale (*M* = 2.8944, *SD* = 2.0413; *t*(256) = -9.2677, $p < 0.05$). This indicates that, overall, CAPTCHA-style image rating method is faster than Likert scale rating method.

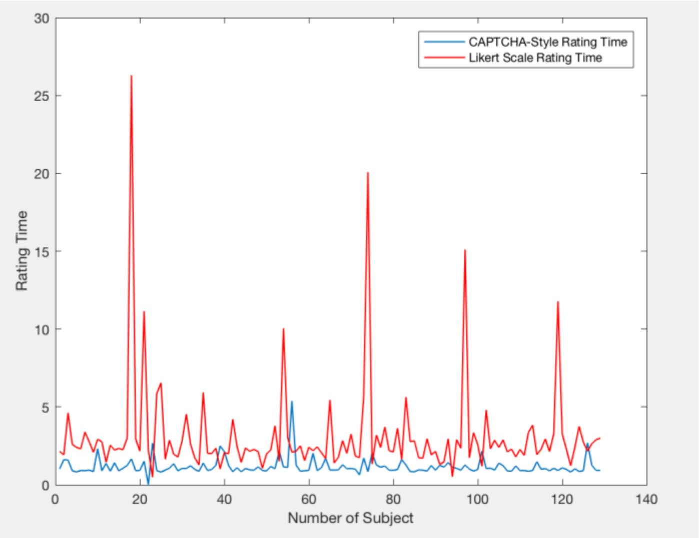

Figure 2. ***The rating time using Likert scale and using CAPTCHA-style.***

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 1**. 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 1.
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 was 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 could reach the same performance as Likert scale method did.

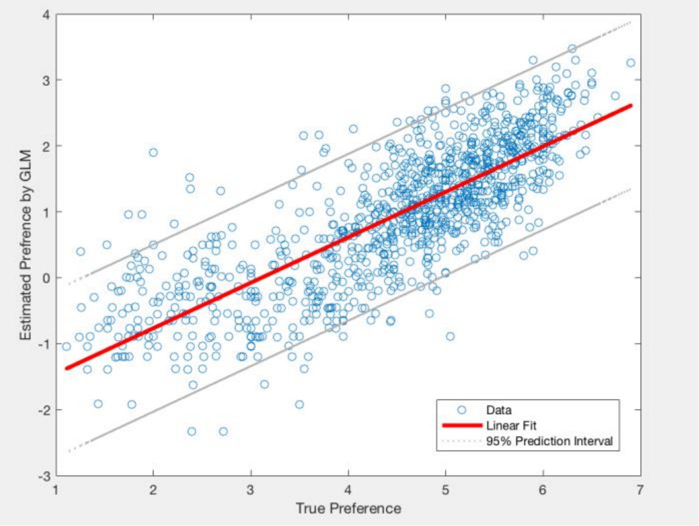

Figure 3. **Relationships between preference rating using Likert scale and predicted preference rating using CAPTCHA-style method.** The true preference indicates the ratings

using Likert scale. The estimated preference indicates the predicted ratings using CAPTCHA-style method generated by GLM. Circle indicate each image.

Reliability. In order to examine the internal consistency reliability of CAPTCHA-style rating and Likert scale rating, we randomly divided ratings for each image into two groups and then calculated the correlation between the two groups of average ratings. There was an insignificant result found, where the internal consistency reliability of CAPTCHA-style rating (r = 0.9309, 95% CI [0.9237, 0.9374], $p < 0.05$) was similar to the reliability of Likert scale rating (r = 0.9311, 95% CI [0.9189, 0.9416], $p < 0.05$).

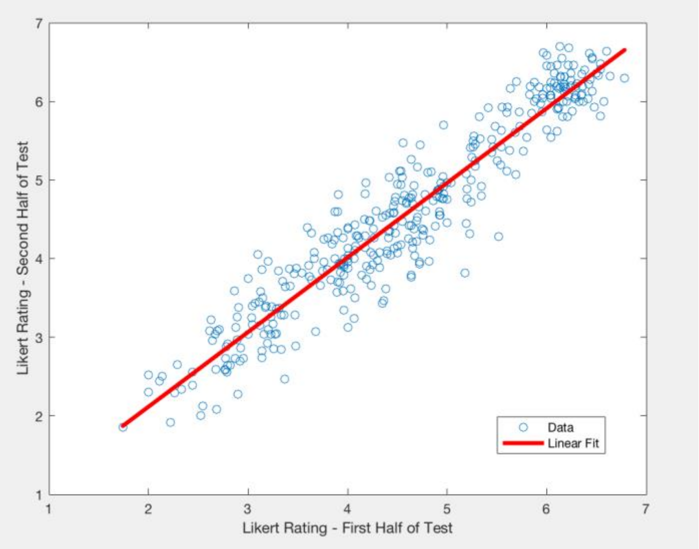

Figure 4. **Scatterplot of Likert ratings on the second half of the test plotted against ratings of the first of the test.**

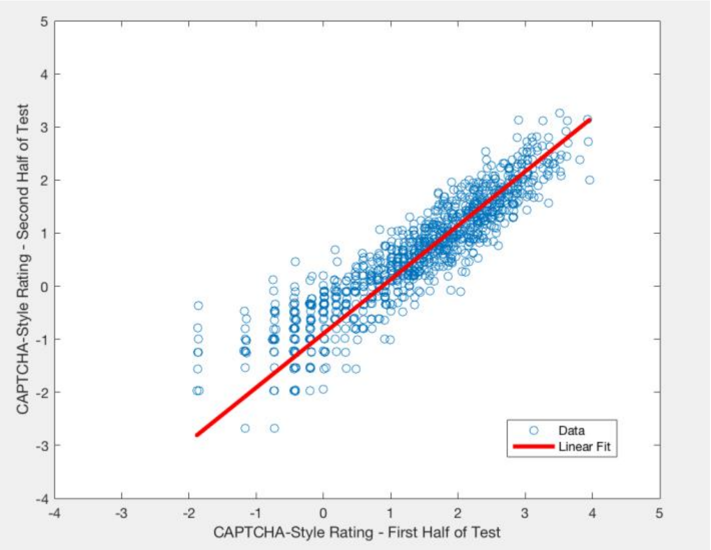

Figure 5. **Scatterplot of CAPTCHA-style ratings on the second half of the test plotted against ratings of the first of the test.**

3. General Discussion

The goal of our study was to validate an alternative to the Likert scale and pairwise comparison for image rating. Our primary hypothesis was support. CAPTCHA-style image rating method was nearly two times faster than Likert scale rating method. Our second hypothesis was not fully supported. The CAPTCHA-style image rating method yielded perfect consistency to the Likert scale assessment but without significantly higher reliability. There are two main reasons that could explain this finding. First, the two experiments were conducted under different conditions. The Likert scale ratings were collected from 200 participants rating 375 scene images, which was not the same as our study with 126 subjects viewing 1030 images. Second, we evaluated the internal consistency using split-half correlation. The data type for Likert scale, however, was ordinal which was not the same as the dichotomous rating in CAPTCHA-style. Therefore, additional examination of reliability would be valuable, as would replication studies with larger number of subjects.

The main limitation of this study is related to position bias. The probability of an image being clicked depends not only on its preference, but on its position on the screen. In the current study, images on the first and second row received more clicks than the images on the third row, particularly for the images on the second row and the second column. We accounted for this bias by performing GLM analyses, which ensured that image position was one independent variable that might affect whether an image was clicked. This model generated a sequence of predicted ratings for each image, which was significantly correlated with the ratings achieved with Likert scale. This positive correlation suggests that CAPTCHA-style image rating could yield the same results as the Likert scale. Given that reasons for this bias are still unknown, future studies may conduct eye-tracking experiments to get a comprehensive record of the regions of the image that are being selected and explore the underlying mechanisms responsible for this tendency. Understanding the nature of position bias will be an important step to study how to model and correct for it. We expected to compare the rating time of pairwise comparison and CAPTCHA-style rating, yet data related to image rating using pairwise comparison was not available in the recent literature.

We envision that the usefulness of CAPTCHA-style image rating method is not necessarily limited to the ratings on environmental scenes. Not only are there likely to be image rating tasks in other domains of laboratory experiment, such as face rating, but also this method could be applied to construct large image datasets, which generally contain the set of images along with corresponding ratings or labels from a representative pool of human subjects. This built large scale image database could then be used as training and validation set in deep learning.

In summary, we conclude that Likert scales and CAPTCHA-style can be both used for image rating, but the CAPTCHA-style method improves the efficiency of human-in-loop annotation and lowers the costs of manual labeling. Therefore, the CAPTCHA-style image rating method, as an alternative to the Likert scale, could be applied to enormous image studies and large scale dataset construction.

ACKNOWLEDGMENTS

We thank Dr. Benjamin Soltoff and Dr. Kyoung Whan Choe for their guidance and advice throughout this project.

References

Albaum, G. (1997). *The Likert scale revisited. Market Research Society. Journal., 39*(2), 1-21.

Berlyne, D. E. (1963). Complexity and incongruity variables as determinants of exploratory choice and evaluative ratings. *Canadian Journal of Psychology/Revue canadienne de psychologie, 17*(3), 274.

Berman, M. G., Jonides, J., & Kaplan, S. (2008). The cognitive benefits of interacting with nature. *Psychological science, 19*(12), 1207-1212.

Berman, M. G., Hout, M. C., Kardan, O., Hunter, M. R., Yourganov, G., Henderson, J. M., ... & Jonides, J. (2014). The perception of naturalness correlates with low-level visual features of environmental scenes. *PloS one*, *9*(12), e114572.

Day, H. (1968). The importance of symmetry and complexity in the evaluation of complexity, interest and pleasingness. *Psychonomic Science, 10*(10), 339-340.

Gur, D., Rubin, D. A., Kart, B. H., Peterson, A. M., Fuhrman, C. R., Rockette, H. E., & King, J. L. (1997). Forced choice and ordinal discrete rating assessment of image quality: a comparison. *Journal of digital imaging, 10*(3), 103-107.

Kaplan, S., Kaplan, R., & Wendt, J. S. (1972). Rated preference and complexity for natural and urban visual material. *Perception & Psychophysics, 12*(4), 354-356.

Kardan, O., Demiralp, E., Hout, M. C., Hunter, M. R., Karimi, H., Hanayik, T., ... & Berman, M. G. (2015). Is the preference of natural versus man-made scenes driven by bottom–up processing of the visual features of nature?. *Frontiers in psychology*, *6*, 471.

Kardan, O., Henderson, J. M., Yourganov, G., & Berman, M. G. (2016). Observers' cognitive states modulate how visual inputs relate to gaze control. *Journal of Experimental Psychology: Human Perception and Performance, 42*(9), 1429.

Khosla, A., Raju, A. S., Torralba, A., & Oliva, A. (2015). Understanding and predicting image memorability at a large scale. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2390-2398).

Kotabe, H., Kardan, O., & Berman, M. G. (2016). Can the High-Level Semantics of a Scene be Preserved in the Low-Level Visual Features of that Scene? A Study of Disorder and Naturalness. In CogSci.

Kotabe, H. P., Kardan, O., & Berman, M. G. (2016). The order of disorder: Deconstructing visual disorder and its effect on rule-breaking. *Journal of Experimental Psychology: General*, *145*(12), 1713.

Kotabe, H. P., Kardan, O., & Berman, M. G. (2017). The nature-disorder paradox: A perceptual study on how nature is disorderly yet aesthetically preferred. *Journal of Experimental Psychology: General*, *146*(8), 1126.

Martindale, C., Moore, K., & Borkum, J. (1990). Aesthetic preference: Anomalous findings for Berlyne's psychobiological theory. *The American Journal of Psychology*, 53-80.

Phelps, A. S., Naeger, D. M., Courtier, J. L., Lambert, J. W., Marcovici, P. A., Villanueva Meyer, J. E., & MacKenzie, J. D. (2015). Pairwise comparison versus Likert scale for biomedical image assessment. *American Journal of Roentgenology, 204*(1), 8-14.

Schroeder, H. W., & Anderson, L. M. (1984). Perception of personal safety in urban recreation sites. *Journal of leisure research*, *16*(2), 178-194.

Seresinhe, C. I., Preis, T., & Moat, H. S. (2017). Using deep learning to quantify the beauty of outdoor places. Royal Society open science, 4(7), 170170.

Seresinhe, C. I., Moat, H. S., & Preis, T. (2018). Quantifying scenic areas using crowdsourced data. Environment and Planning B: Urban Analytics and City Science, 45(3), 567-582.

Thayer, R. L., & Atwood, B. G. (1978). Plants, complexity, and pleasure in urban and suburban environments. *Environmental Psychology and Nonverbal Behavior, 3*(2), 67-76.

Torralba, A. (2009). How many pixels make an image?. *Visual neuroscience*, *26*(1), 123-131.

Von Ahn, L., Blum, M., Hopper, N. J., & Langford, J. (2003, May). CAPTCHA: Using hard AI problems for security. In *International Conference on the Theory and Applications of Cryptographic Techniques* (pp. 294-311). Springer, Berlin, Heidelberg.

Wohlwill, J. F. (1968). Amount of stimulus exploration and preference as differential functions of stimulus complexity. *Perception & Psychophysics, 4*(5), 307-312.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010, June). Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 3485-3492). IEEE.

Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., & Xiao, J. (2015). Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*.