Reproducibility test of the results of Farzanfar, D. & Walther, D. B. (2023)

Guo Kaitong[1], Deng Xinyu[1], Huo Yining[1], Shi Yinuo[1], & Zhang Yan[1]

[1] Group 7

Author Note

A chart on the division of labour among members is presented in the appendix at the end of this report.

The authors made the following contributions. Guo Kaitong: Reviewing the literature, HTML report presentation, Writing of written reports; Deng Xinyu: Visual image reproduction of experiment 1, 2 and 3, Partial inferential statistical reproduction, RMD collation and summary, Writing code for html reporting and papaja, Writing of written reports; Huo Yining: Reproduction of descriptive statistics of experiment 1 and 2, Partial inferential statistical reproduction, RMD collation, Writing of written reports; Shi Yinuo: Matlab model and image processing of experiment 1, 2 and 3, Partial inferential statistical reproduction, RMD collation, Writing code for html reporting and papaja, Writing of written reports; Zhang Yan: Literature screening, Descriptive statistial and inferred statistical reproduction of experiments 1 and 3, RMD collation, Writing of written reports.

Abstract

In order to explore the cognitive mechanisms behind aesthetic experience, researchers have developed a contour aesthetic model. The results show that contour manipulation directly affects subjective aesthetic judgments and supports the view that visual regularity is the basis of human ability to obtain pleasure from visual information. Reproducibility is an important cornerstone of scientific research. Previous research results may not be reproduced due to differences in experimental design, data collection and analysis methods, etc. Therefore, it is crucial to test the reproducibility of existing research results. This examination was designed to understand the reliability of the calculated results of the contour aesthetic model developed in the previous study by recalculating the contour properties and predicting the aesthetic value of the scene using the random forest model, using the same data sets and algorithms as the original study. The test results show that both the contour properties obtained by recalculating and the prediction results obtained by using the same algorithm are in agreement with the original research. However, the results of repeatability test also show that there may be some differences in the impact of individual contour attributes on aesthetic judgment, which suggests that we need to further study the interaction between different contour attributes and their comprehensive impact on aesthetic judgment. The results of the reproducibility test highlight the importance of rigor and transparency in the scientific process, and by making the data and code public, other researchers can more easily reproduce the findings and validate and extend them.

Keywords: visual aesthetics, contour attributes, aesthetic judgment, calculation, repeatability

Reproducibility test of the results of Farzanfar, D. & Walther, D. B. (2023)

1 Introduction

1.1 Selected Literature

This study is based on a paper by Farzanfar & Walther published in Psychological science in 2023 entitled “Changing What You Like: Modifying Contour Properties Shifts Aesthetic Valuations of Scenes”.

The study’s data, materials, and code are publicly available on the Open Science Framework (OSF) and can be accessed at https://osf.io/rb2wc/.

1.2 Literature Introduction

1.2.1 Background of literature research

Visual aesthetics is an important topic in the field of psychology and art, and it is generally believed that beautiful scenes or objects can trigger pleasant emotional experience(Biederman & Vessel, 2006; Chatterjee, 2022; Farzanfar & Walther, 2023; Palmer, Schloss, & Sammartino, 2013).The visual information provided by ordinary scenes, such as beaches and sunsets, can trigger the perceiver’s aesthetic feelings. It is precisely through such a phenomenon that artists can rationally arrange the visual features in artistic performance, so that people can obtain aesthetic pleasure. This pleasure may be related to the way human visual system processes information. Visual information is an important way for humans to obtain environmental information, and visual features carry a wealth of information, such as the direction, length, curvature and connection mode of contour lines, which can convey the spatial relationship between objects and scenes, help people identify objects and classify scenes, and evaluate the potential costs and benefits of interacting with different objects and environments. Research has shown that these visual features play an important role in aesthetic judgments, such as curved

shapes are often considered more attractive than straight shapes, and symmetry is also considered an important aesthetic feature. Furthermore, according to the extension of Marr's visual framework - component recognition theory (Marr, 1982), the categorical process by which people recognize objects in the environment occurs through a set of visual primitives, similar to geometric building blocks of shapes, which are viewpoint-invariant, i.e. non-accidental properties (Biederman, 1987). Non-accidental attributes are those visual features that help people identify objects and scenes and understand their spatial relationships, such as the way contours are connected. Studies have shown that non-accidental attributes play an important role in object recognition and scene classification.

However, exactly what visual features are responsible, and the cognitive mechanisms behind this aesthetic response, are still unclear. Therefore, the authors adopted an empirical approach to study how systematic changes in visual features in the scene would affect the aesthetic response of the observer, so as to test the possible relationship between visual information and aesthetic response, and further explore how visual features affect people's aesthetic judgment and reveal the neuropsychological basis behind it.

1.2.2 Literature research problems

The main research question of this literature is to what extent does the human ability to perceive and extract meaning from visual features produce aesthetic experience? How do basic visual features (contour attributes and their spatial relationships) affect people's aesthetic evaluation of a scene? This paper further discusses the essence of these visual features and the role they play in aesthetic judgment. Based on the existing research background and the above research questions, the researchers conducted three experimental designs to verify three research hypotheses, namely: (1) There is a high correlation between the aesthetic value prediction of different scenes generated by the random forest model trained on the contour statistical characteristics and people's real aesthetic rating of different scenes; (2) Changing the contour characteristics of the scene on the basis of model prediction can change people's real aesthetic

rating in the expected direction; (3) When people's semantic processing of the scene is interrupted by flipping the contour in the scene, people's evaluation of the contour scene with high aesthetic value generated by the model will also be higher than that of the contour scene with low aesthetic value.

1.2.3 Literature research results and conclusions

By constructing contour models and conducting relevant experimental studies, researchers in this literature first show that contour attribute models can accurately predict people's aesthetic evaluation of scenes, whether it is line drawings or color photos. Secondly, modifying the contour attributes can lead to changes in people's aesthetic evaluation of the scene. The scene that retains the contour with high predictive aesthetic value is more popular than the scene that retains the contour with low predictive aesthetic value. While scene inversion reduces people's aesthetic evaluation of scenes, it does not change the aesthetic advantage of scenes with high-value contours, suggesting that feature manipulation directly changes aesthetic judgment. Finally, observers' negative emotions can predict their aesthetic evaluation of the scene, and observers with higher negative emotions may be more likely to give a positive evaluation of the line painting scene. Specifically, the results of experiment 1 show that there is a significant correlation between the aesthetic value predicted by the model and people's real aesthetic rating, indicating that the random forest model trained on scene contour characteristics can predict people's subjective aesthetic evaluation of the scene. The T-test results of experiment 2 show that when viewing contour modification scenes with high aesthetic grade and low aesthetic grade generated by the model, there is a significant difference between people's real aesthetic ratings of the two scenes, and the linear mixed effect model also shows a significant effect of high and low aesthetic conditions, indicating that modifying contour characteristics can change people's aesthetic evaluation of the scene. That is, there is a causal relationship between contour characteristics and people's aesthetic ratings. The ANOVA analysis results of experiment 3 show that the main effect of the semantic conditions of contour inversion and upright is significant. For both high aesthetic

conditions and low aesthetic conditions, the aesthetic rating of upright contour scenes is higher than that of inverted contour scenes. Moreover, the linear mixed effect model also shows significant effects of high and low aesthetic conditions, indicating that even if the semantic processing of scenes is interfered with, For example, reverse the contour in the scene, and the causal relationship between the contour characteristics in experiment 2 and people's aesthetic ratings still exists.

Therefore, the literature considers that the basic visual features (contour attributes and their spatial relationships) are important factors affecting people's aesthetic evaluation of the scene. By manipulating these features, one can change one's aesthetic evaluation of a scene, even if the semantic identity of the scene remains the same. Observers' individual differences (such as mood) also affect their aesthetic evaluation of the scene. These findings support the concept of a "perceptual reward system," in which people take information from their perceptual environment and experience pleasure. These findings provide important experimental evidence for understanding the cognitive mechanism behind visual aesthetics, lay a foundation for further research on the relationship between visual features and aesthetic experience, and provide some relevant guidance for artists and designers to create more attractive works.

2 Methods

2.1 Introduction to the original research methods

This literature designed three experiments to explore the impact of the contour features and their spatial relationships in natural scenes on aesthetic judgments.

In experiment 1, a random forest algorithm was used to construct a model based on the statistical properties of individual contours (including direction, length, curvature, and connection points), and to predict the aesthetic value for scenes of line drawings and color photographs. Based on the color photographs, the authors verified the predictive consistency of contour statistical properties and constructed a random forest model which included the statistical

properties of the color. In this experiment, the researchers established an aesthetic model based on contour properties to predict the aesthetic value of the scene. Firstly, they use MATLAB, extracting the color photographs and line drawings from the Toronto Scenes Dataset as stimulus materials, respectively including natural scenes and artificial scenes. The contour properties of the scene included direction, length, curvature, angle and type of contour junctions. A random forest regression model was trained based on these contour properties, and the properties were also used as features to generate aesthetic value predictions for different scenes (beaches, cities, forests, highways, mountains, offices). The representative decision trees and variable importance were given. In addition, researchers invited 75 participants to rate the aesthetics of the line drawings, while for color photographs, 121 participants were invited to rate (6 participants were excluded from the analysis). Two groups of participants were asked to view scene images and rate on a 5-point Likert scale. Then the researchers extracted the direction, length, curvature, and junction of the contours in each scene by using the random forest algorithm to establish a model, and predicted the aesthetic value of the scene based on the contour properties, and test the correlation between the predicted values of the model and the actual aesthetic score. R was used to present the aesthetic ratings of M and SD for two types of stimulus sets: line drawings and color photographs. M and SD were presented for two types of stimuli sets according to the six scenes. Subsequently, a correlation analysis was conducted between the predicted values of the model and the actual evaluation values.

The purpose of the 2nd experiment is to verify the causal effect of contour properties on aesthetic judgment. Therefore, the researchers changed the contour properties and generated aesthetic value predictions for individual contours in a single scene based on the model established in Experiment 1. The contours were sorted according to the predicted values and were divided into two groups: high aesthetic value and low aesthetic value. Scenes containing different groups of contours were generated to test the aesthetic judgment of observers on different contour scenes. In this experiment, the researchers generated scenes with different contour properties using the contour aesthetic model constructed in Experiment 1 as stimulus materials. They used a

random forest model to predict the aesthetic value of a single contour and ranked the contours from lowest to highest based on the predicted aesthetic value. All scenes were divided into two groups of contours modified scenes with high and low aesthetic levels. 77 participants were recruited with the same standards and procedures in Experiment 1 to see the modified scenes and receive their aesthetic ratings. Information on age, gender, educational attainment, environmental-type familiarity, positive affect score, negative affect score, creativity score, years of artistic training and experience was collected to test whether the aesthetic value of the modified scenes was consistent with the predictions from the model. Paired sample t-test was used in R to test whether there was a statistically significant difference in average aesthetic value judgment when viewing modified scenes with high-level and low-level contours. Linear mixed effects model analysis was conducted using the lmer package to explore the impact of individual differences on aesthetic judgment.

The purpose of the 3rd experiment is to expand the model properties, and to explore the impact of spatial relationships and semantic content on aesthetic judgment, extend the properties in the contour model to the spatial relationships of adjacent contours (separation, parallelism, mirror symmetry), generate new contour scenes containing spatial relationship properties, and test the aesthetic judgment of independent observers. Scene reversal is used to manipulate the acquisition of semantic content in the scene, and the relative statistical contribution of individual contours and their spatial relationships in aesthetic reactions is clearly measured and compared in separate scene classification experiments. The stimulus material for this experiment is generated with the model which is constructed in Experiment 1, including spatial relationship properties in the scene and is processed through scene inversion. The researchers firstly used the same MATLAB program as in Experiment 1 to predict the aesthetic rating of a given scene in a new random forest model. The spatial relationship attributes of adjacent contours were added to the feature list to generate aesthetic value predictions for different scenes. Meanwhile, following the same stimulus generation process as in Experiment 2, scene images were generated for this experiment, with 950 presented as upright and 132 presented as reversed (i.e. rotated 180

degrees). Recruit two groups of participants simultaneously for aesthetic rating and scene classification experiments. In the aesthetic rating experiment, participants watched the modified scene and performed an aesthetic rating; In the scene classification experiment, participants watched scene images and performed scene classification. The researchers invited 77 participants for aesthetic rating; In another scenario classification experiment, an additional 60 participants were invited. They were randomly presented with semi split and inverted stimuli, and were asked to respond to the categories of scenes (beaches, cities, forests, highways, mountains, offices). Each participant classified 350 scenes and collected the same individual information as in Experiment 2. In terms of data analysis, R was used for correlation analysis, linear mixed effects model analysis, and the variance partitioning analysis was used to compare Model 1 based on contour properties with the Model 2 based on contour properties and spatial relationships, in order to test the impact of spatial relationships and scene reversal on aesthetic judgment, whether scene reversal changed the semantic content of the scene, to test the prediction differences between different models, and to compare the relative statistical contributions of individual contours and their spatial relationships in aesthetic response.

In the process of calculating the repeatability test, we reproduced the above analysis using MATLAB and R.

2.2 Replication Approach and R Packages

The team members reproduced the contents in the order of experiments I, II and III in the literature, including descriptive statistics, inferential statistical analysis and visual results of each experiment.

First of all, in terms of data preparation, the original literature provides the original data, processed data and data processing methods used in the research, as well as the code used in the research and the explanatory files about the code. The team members downloaded and decompressed the data set from https://osf.io/rb2wc/, and checked the format of the data set.

Make sure it is in a format that R can read (such as CSV or Excel). Secondly, the original code was used to verify the data analysis results. Although the model results were not successfully run in R, the team members built the model through MATLAB using the Random Forest algorithm, taking contour features and color features as predictive variables and aesthetic scores as response variables. Use cross-validation to assess the accuracy of model predictions and analyze the importance of features; In addition, the “LD1.csv” and “CP.csv” data were imported according to the R code provided in experiment 1, and the unstandardized and standardized mean (M) and standard deviation (SD) of the aesthetic ratings classified according to the six types of scenes were reproduced. When the aesthetic score of the contour modification scene was reproduced, the modified contour scene was generated according to the method of experiment 2, and the “LD2.csv” data file was imported. The differences in aesthetic score of different contour scenes were analyzed using the paired sample T-test or the linear mixed effects model. The linear mixed effect model or correlation analysis was used to analyze the influence of individual differences on aesthetic scores. ANOVA or partial correlation analysis were used to compare the explanatory power of different models on aesthetic rating. The data files “LD3.csv” and “cross.csv” were imported during the scene classification experiment. The modified and reversed scenes should have been generated according to the method of experiment 3, and the difference in classification accuracy of different scene conditions was analyzed by ANOVA or linear mixed effects model. However, the original data file does not provide the variable of accuracy and the data cannot be reproduced.

Finally, when we reproduced the visual visualization results in the literature, we could not get the results consistent with the pictures presented in the literature by relying on the code provided by the original literature, and many problems occurred, such as the lack of chart elements, the mismatch between horizontal classification and conditions, and the mismatch between chart types, etc. Therefore, we rewrote the code for the visual part of the article. On this basis, the chart is embellished, such as adjusting the range of horizontal and vertical coordinates so that the content is not too crowded, adjusting the label spacing so that they do not overlap each

other, and so on, so that the results are similar or consistent with the pictures presented in the original literature.

The R packages used in the reproduction process are dplyr, tidyverse, ggplot2, ggstatsplot, ggpubr, summarytools, psych, randomForest, random Forest Explainer, effectsize, car, afex, DT, papaja, ggrepel, lme4, lmerTest, MuMIn, lmtest, randomForest, randomForestExplainer, ppcor. In the production of report courseware, xaringan, xaringanthemer, xaringanExtra, knitr and other R packages are used.

What's more, we used R (Version 4.4.0; R Core Team, 2024) and the R-packages *papaja* (Version 0.1.2.9000; Aust & Barth, 2023), and *tinylabels* (Version 0.2.4; Barth, 2023) for the generation of this report.

3 Results

3.1 Descriptive statistics

Table 1

Table 1: Descriptive statistics of the reproduced results of experiment 1

Category	Repeatability	M(LD)	SD(LD)	M(CP)	SD(CP)
Beach	Literature	0.7	0.42	0.58	0.34
	This study	0.7	0.42	0.58	0.34
	δ	0	0	0	0
	Level	100% consistent	100% consistent	100% consistent	100% consistent
City	Literature	0.36	0.35	-0.18	0.41
	This study	0.36	0.35	-0.18	0.41
	δ	0	0	0	0

	Level	100% consistent	100% consistent	100% consistent	100% consistent
Forest	Literature	-0.4	0.28	0.35	0.29
	This study	-0.4	0.28	0.35	0.29
	δ	0	0	0	0
	Level	100% consistent	100% consistent	100% consistent	100% consistent
Highway	Literature	0.09	0.3	-0.62	0.29
	This study	0.09	0.3	-0.62	0.29
	δ	0	0	0	0
	Level	100% consistent	100% consistent	100% consistent	100% consistent
Mountain	Literature	-0.24	0.32	0.71	0.26
	This study	-0.24	0.34	0.71	0.26
	δ	0	5.9	0	0
	Level	100% consistent	small deviation	100% consistent	100% consistent
Office	Literature	0.1	0.3	-0.84	0.31
	This study	0.1	0.3	-0.84	0.31
	δ	0	0	0	0
	Level	100% consistent	100% consistent	100% consistent	100% consistent

Table 2

Table 2: Descriptive statistics of the reproduced results of experiment 2

Repeatability	M(Top)	SD(Top)	M(Bottom)	SD(Bottom)
Literature	2.37	1.18	2.72	1.22
This study	2.37	1.18	2.72	1.22
δ	0	0	0	0
Level	100% consistent	100% consistent	100% consistent	100% consistent

Table 3

Table 3: Descriptive statistics of the reproduced results of experiment 3

Repeatability	M(Top)	SD(Top)	M(Bottom)	SD(Bottom)
Literature	2.35	1.26	2.11	1.23
This study	2.35	1.26	2.11	1.23
δ	0	0	0	0
Level	100% consistent	100% consistent	100% consistent	100% consistent

256 **3.2 Inferential statistics**

Table 4

Table 4: Inferential statistics of the reproduced results of experiment 2(t-test)

Repeatability	t	p	d	CI
Literature	15.73	<0.00001	0.69	[0.30, 0.39]
This study	15.73	<0.00001	0.69	[0.30, 0.39]
δ	0	0	0	0
Level	100% consistent	100% consistent	100% consistent	100% consistent

Table 5

Table 5: Inferential statistics of the reproduced results of experiment 2(LMM-1)

/	Repeatability	est	t	p	CI
High vs Low	Literature	0.33	12.36	<2*10-16	[0.27, 0.38]
	This study	0.33	12.36	<2*10-16	[0.275, 0.385]
	δ	0	0	0	/
	Level	100% consistent	100% consistent	100% consistent	rounding problem
Negative	Literature	0.052	2.59	0.01	[0.015, 0.09]
	This study	0.0528	2.6	0.01	[0.015, 0.09]
	δ	/	0.0038	0	0
	Level	rounding problem	small deviation	100% consistent	100% consistent
Age	Literature	0.012	0.98	0.33	[-0.011, 0.037]
	This study	0.0128	0.98	0.33	[-0.0117, 0.037]
	δ	/	0	0	/
	Level	rounding problem	100% consistent	100% consistent	rounding problem
Gender	Literature	-0.07	-0.45	0.65	[-0.37, 0.22]
	This study	-0.07	-0.45	0.65	[-0.376, 0.228]
	δ	0	0	0	/
	Level	100% consistent	100% consistent	100% consistent	rounding problem
Creativity	Literature	0.007	0.59	0.55	[-0.015,0.03]
	This study	0.007	0.59	0.559	[-0.0159,0.03]

Table 5

Table 5: Inferential statistics of the reproduced results of experiment 2(LMM-1) (continued)

/	Repeatability	est	t	p	CI
	δ	0	0	/	/
	Level	100% consistent	100% consistent	rounding problem	rounding problem
Artistic train	Literature	0.049	0.58	0.55	[-0.10, 0.20]
	This study	0.0499	0.59	0.559	[0.107, 0.207]
	δ	/	0.0169	/	/
	Level	rounding problem	small deviation	rounding problem	rounding problem
Experience	Literature	-0.009	-0.22	-0.22	[-0.088, 0.069]
	This study	-0.009	-0.22	0.82	[-0.0886, 0.0698]
	δ	0	0	1.268	/
	Level	100% consistent	100% consistent	large deviation	rounding problem
Environmental familarity urban	Literature	0.017	0.74	0.45	[-0.25, 0.59]
	This study	0.017	0.75	0.457	[-0.258, 0.599]
	δ	0	0.0133	/	/
	Level	100% consistent	small deviation	rounding problem	rounding problem
Suburban versus rural	Literature	0.2	0.87	0.38	[-0.24, 0.66]
	This study	0.207	0.87	0.386	[-0.247, 0.66]
	δ	/	0	/	/

Table 5

Table 5: Inferential statistics of the reproduced results of experiment 2(LMM-1) (continued)

/	Repeatability	est	t	p	CI
	Level	rounding problem	100% consistent	rounding problem	rounding problem
positive	Literature	0.036	1.53	0.12	[-0.007,0.08]
	This study	0.036	1.53	0.13	[-0.008,0.08]
	δ	0	0	0.0769	/
	Level	100% consistent	100% consistent	small deviation	rounding problem

Table 6

Table 6: Inferential statistics of the reproduced results of experiment 2(LMM-2)

Repeatability	R^2m	R^2c	SD
Literature	0.068	0.388	0.68
This study	0.0687	0.388	0.68
δ	/	0	0
Level	rounding problem	100%consistent	100%consistent

Table 7

Table 7: Inferential statistics of the reproduced results of experiment 3(ANOVA)

/	Repeatability	F	p	d	CI
condition	Literature	323.77	<0.001	0.1	[0.22, 0.29]
	This study	323.77	<0.001	0.1	[0.22, 0.269]
	δ	0	0	0	0.078

	Level	100% consistent	100% consistent	100% consistent	small deviation
manipulation	Literature	375.81	<0.001	NA	[0.35, 0.44]
	This study	2.35	<0.001	0.11	[0.359, 0.440]
	δ	0	0	/	/
	Level	100% consistent	100% consistent	/	rounding problem

Table 8

Table 8: Inferential statistics of the reproduced results of experiment 3(post-hoc)

/	Repeatability	Mdiff	padj	CI
Inverted: top vs. Bottom	Literature	0.14	0.002	[0.04, 0.23]
	This study	0.14	0.00169	[0.04, 0.239]
	δ	0	/	/
	Level	100%consistent	rounding problem	rounding problem
Top: upright vs. inverted	Literature	0.45	<0.00001	[0.38, 0.53]
	This study	0.458	<0.00001	[0.38, 0.53]
	δ	/	0	0
	Level	rounding problem	100%consistent	100%consistent
Bottom: upright vs. inverted	Literature	0.34	<0.00001	[0.26, 0.42]
	This study	323.77	<0.00001	[0.266, 0.417]
	δ	0	0	/

	Level	100%consistent	100%consistent	rounding problem

Table 9

Table 9: Inferential statistics of the reproduced results of experiment 3(LMM-1)

/	Repeatability	β	t	p	CI
Top VS bottom	Literature	0.33	12.36	<2*10-16	[0.27.0.38]
	This study	0.245	13.3	<2*10-16	[0.21, 0.28]
	δ	0.347	0.0707	0	0.3214
	Level	large deviation	small deviation	100% consistent	large deviation
Negative affect	Literature	-0.014	2.33	0.02	[-0.003, 0.03]
	This study	-0.0148	2.336	0.02	[-0.003, 0.026]
	δ	/	/	0	/
	Level	rounding problem	rounding problem	100% consistent	rounding problem
Positive affect	Literature	-0.01	-2.21	0.03	[-0.02, 0.002]
	This study	-0.01	-2.21	0.03	[-0.02, 0.0019]
	δ	0	0	0	/
	Level	100% consistent	100% consistent	100% consistent	rounding problem
Age	Literature	0.004	0.33	0.74	[-0.018, 0.026]
	This study	0.004	0.33	0.74	[-0.018, 0.026]
	δ	0	0	0	0
	Level	100% consistent	100% consistent	100% consistent	100% consistent
Gender male	Literature	0.32	2.06	0.04	[0.03, 0.62]

Table 9

Table 9: Inferential statistics of the reproduced results of experiment 3(LMM-1) (continued)

/	Repeatability	β	t	p	CI
	This study	0.325	2.066	0.04	[0.03, 0.617]
	δ	0	/	0	/
	Level	100% consistent	rounding problem	100% consistent	rounding problem
Gender other	Literature	-0.41	-0.6	0.54	[-1.67, 0.84]
	This study	-0.41	-0.6	0.548	[-1.67, 0.84]
	δ	0	0	/	0
	Level	100% consistent	100% consistent	rounding problem	100% consistent
CB	Literature	0.034	1.23	0.21	[-0.10, 0.20]
	This study	0.0345	1.239	0.219	[-0.018, 0.087]
	δ	/	/	/	2.645
	Level	rounding problem	rounding problem	rounding problem	large deviation
IAE	Literature	0.05	1.82	0.07	[-0.10, 0.20]
	This study	0.05	1.82	0.07	[-0.0006,0.101]
	δ	0	0	0	84.5
	Level	100% consistent	100% consistent	100% consistent	large deviation
AA	Literature	0.009	-0.56	0.57	[-0.10, 0.20]
	This study	0.0099	-0.56	0.57	[-0.04,0.02]
	δ	/	0	0	1.12

Table 9

Table 9: Inferential statistics of the reproduced results of experiment 3(LMM-1) (continued)

/	Repeatability	β	t	p	CI
	Level	rounding problem	100% consistent	100% consistent	large deviation
Region suburban	Literature	0.2	0.7	0.38	[-0.12,0.06]
	This study	-0.57	-1.685	0.096	[-0.121,0.065]
	δ	1.35	1.415	2.958	/
	Level	large deviation	large deviation	large deviation	rounding problem
Region urban	Literature	0.17	0.74	0.45	[-1.25,-0.03]
	This study	-0.64	-1.987	0.05	[-1.25,-0.035]
	δ	1.27	1.37	8	/
	Level	large deviation	large deviation	large deviation	rounding problem

Table 10

Table 10: Inferential statistics of the reproduced results of experiment 3(LMM-2)

Repeatability	R^2m	R^2c
Literature	0.159	0.382
This study	0.1595	0.3825
δ	/	/
Level	rounding problem	rounding problem

Table 11

Table 11: Inferential statistics of the reproduced results of experiment 3(Model Comparison

/	Repeatability	F	p	R^2	Y1
Model 1	Literature	3193	<0.001	0.87	0.35
	This study	3193	<0.001	0.87	0.359
	δ	0	0	0	/
	Level	100% consistent	100% consistent	100% consistent	rounding problem
Model 2	Literature	2922	<0.001	0.86	
	This study	2922	<0.001	0.86	
	δ	0	0	0	
	Level	100% consistent	100% consistent	100% consistent	

3.3 Assessment of the reproducibility of the literature

Table 12

Table 12: Evaluation table for reproducibility of all data calculations

Repeatability	Quantity(N*)	Proportion(%)
100% consistent (δ = 0%)	106	56.99
small deviation (0% < δ < 10%)	7	3.76
large deviation (δ > 10%)	12	6.45
rounding problem	36	19.35
not replicable	25	13.44

Based on the table above, it is evident that the original literature contains a total of 186 data

results that need to be replicated, of which 25 cannot be subjected to reproducibility tests. The issues are primarily concentrated in two areas: First, the correlation (r) and mean squared error (MSE) of two random forest models (RF) mentioned in the literature could not be reproduced in R. This is because, although the researchers loaded the necessary packages ("random Forest" and "random Forest Explainer") in the provided R code, they did not specify the actual code needed for execution in subsequent parts of the R script. As a result, it was impossible to obtain data results demonstrating a significant correlation between the predicted aesthetic value based on cross-validation and the observed aesthetic ratings. Second, the semantic content control experiment data results from Experiment 3 could not be reproduced in R either, because the data files provided by the researchers on the OSF platform did not include the variable for accuracy. Thus, it was not possible to replicate the results, and therefore their consistency remains uncertain.

Additionally, among the data results that could be replicated, 55 showed inconsistencies. In R, the data with small deviations compared to the original findings mainly included the standard deviation (SD) of aesthetic ratings for "Mountain" in Experiment 1 descriptive statistics, and the t-values for "Artistic train" and "Environmental familiarity urban" in the linear mixed models of Experiment 2. In contrast, results with large deviations were mainly found in Experiment 3 within the linear mixed-effects models assessing the impact of individual factors on aesthetic ratings of stimuli. Inconsistencies due to rounding issues were primarily observed in Experiments 2 and 3 in the linear mixed-effects models, where researchers did not standardize the number of decimal places retained nor the rounding principles when presenting the data. It is worth noting that the original literature reported confidence intervals (CI), but team members encountered difficulties in obtaining these results when running the R code. Upon further inspection, it appeared that the researchers might have commented out the CI calculation code by adding a "#" because it required a lengthy execution time. Therefore, the code did not produce CI results during the initial runs. After altering the code to remove the comment, the CI results could be generated. However, after modifying and rerunning the code, some of the CI results still differed from those reported in the original literature, likely due to rounding issues.

Lastly, the original code from this study generated four visualization result graphs, which differed from those presented in the original document. We made modifications to the code; however, because the reproducibility delta (δ) value could not be calculated, these were not included in the aforementioned table.

4 Discussion

The computational reproducibility of a research study is of paramount importance in ensuring the integrity and reliability of scientific discoveries, representing the ability for other researchers to replicate the study's computational methods and results, thereby validating the study's conclusions and contributing to the advancement of science. The members of our group conducted a test and analysis on the computational reproducibility of the research results from Farzanfar, D. and Walther, D. B. (2023). The validation process employed the original study's computational methods, including the selection of software, R packages, and parameters, and followed the data processing and analysis procedures as outlined in the original literature.

4.1 Analysis of the results of the computational reproducibility test

In our replication, we found that the data in the original literature is open, transparent, and well-documented, the computational methods are reliable, and clear explanations and code are provided for other researchers to repeat the tests. Therefore, our replication test largely reproduces the results and findings of the original literature, enhancing the credibility and reliability of the research conclusions. However, there are still some differences between our replication results and the original study. As shown in Table 15, the proportion of data results that are completely consistent is 56.99%. Although the replication rate is not particularly high, a significant portion of the inconsistencies are due to rounding issues. For the reasons behind other inconsistencies, we need to consider multiple factors that may lead to differences in the replication test results. Here are some speculations on important reasons that may cause differences, with each point explained in detail.

Firstly, the general issue of open access. The minor discrepancies in several results may be due to printing or copy-pasting errors. For example, in the analysis process of Experiment 3's mixed linear model, the original literature yielded two completely identical confidence intervals (for CB and IAE, the confidence intervals were both [-0.10, 0.20]). However, in our replication test, the results were significantly different from the original ones. We believe this discrepancy might be attributed to the original literature copying and pasting the same data into two different results.

Secondly, the specific issue with OSF open access. The description of the correlation analysis methods used in Experiments 1 and 3 is unclear, leading to initial difficulties in completely reproducing the original results using R language. Subsequently, replication was achieved using Matlab and validated. Additionally, the data and code files on OSF do not always correspond to each other. For example, in Experiment 3, the results obtained from the Control experiment for semantic content section are not found in the data files, making it impossible to use the codes to operate on the data and repeat the validation of the experiment's results. Furthermore, the data provided on OSF is not entirely original but is based on ratings for each image. The process of how to rate for each participant are aggregated is not explained in the literature or the codes.

Third, the issue of code integrity. The vast majority of the codes provided by the original researcher were able to reproduce the results of the study, but there still existed some codes that yielded results after running that differed significantly from the original literature, and the statistical plots derived by the researcher from the three experiments in the literature were completely different from those derived from the original code runs, whereas most of the mating plots in the original literature were reassembled and processed from the plots derived from the runs performed by the R code. Based on this situation, members of this group modified and recoded the original code, and eventually obtained similar and more detailed statistical plots as those in the original literature. In response to this reproduction result, we speculate that the original authors' graph visualizations in the literature have been embellished by other software.

Fourth, differences in computational methods and reporting. Even for the same dataset, the use of different research methods may lead to different results, and the use of different statistical methods, different parameter settings, or different data preprocessing steps may lead to differences. For example, this review found that when repeating some of the results of Experiment 3, the original study's settings for the data were not clear, which led to our inability to restore all the results of the study completely. At the same time, there may be differences in the process of data rounding, for example, for many of the results of the original study, this reproduction is not quite the same data, after analyzing and comparing the results, we found that the reason for this discrepancy is that many of the data in the original study have rounding problems, and most of them chose to round off the back of the numbers and retain them directly when they should be rounded up.

In summary, the results of the computational reproducibility tests provide valuable insights into the studied computational methods and contribute to the development of the field. The results show that the studied computational methods are largely reproducible, reliable, transparent and efficient. The reproduction of the validation results of this test on different computational environments and datasets shows that the results and findings of the study are reliable and are likely to be applicable to other similar studies, enhancing the credibility and reliability of the study's conclusions.

4.2 Other thoughts

Data analysis is a complex and multidimensional process involving multiple steps and tools. Firstly the purpose and objectives of the analysis need to be defined, a good problem definition can help to determine the scope and importance of the analysis. Secondly in data cleaning and screening, cleaning the data to ensure its quality and usability, this includes dealing with missing values, outliers, duplicates, etc. The original literature mentions the use of a variety of methods to eliminate invalid data, such as consecutive identical responses, failure of attention checking, completion time anomalies, and so on, and these are more common methods of

excluding invalid data used in the study, but it is necessary to consider whether these methods are reasonable for this study and if there are other invalid data that need to be taken into account. reasonable and whether there are other types of invalid data that need to be considered; the sample size in the experiment was large, but consideration needs to be given to whether this is sufficient and whether it needs to be adjusted based on the statistical power analysis; in addition, the original literature mentions converting aesthetic scores to standardized scores, but it is debatable whether this applies to all the analyses and whether there is a need to consider other methods of standardization. In terms of model selection and assessment, the random forest model was used in the experiment for prediction, and other models or methods can also be considered to explain the model's prediction results, such as feature importance analysis, model visualization, etc., as well as comparative assessment of the models and consideration of the model's generalization power. In terms of statistical methods, linear mixed-effects models were used in the experiment to analyse the effects of individual differences, but things like multiple regression and structural equation modelling can also be taken into account, and also because of the large sample sizes used by the researchers in the original literature, it is necessary to calculate the statistical test power based on the sample sizes and the effect sizes in order to ensure the reliability of the statistical results, e.g. Cohen' d, η2 p, in order to interpret the significance of the results more intuitively. Finally, in the experimental design and hypothesis testing of the study, it is necessary to consider whether the experimental design is reasonable, such as whether it is necessary to control other variables, whether it is necessary to balance the scenario categories, etc.; at the same time, it is necessary to specify the method and significance level of the hypothesis test, and to carry out the interpretation of the results of the hypothesis test.

This computational reproducibility test for the results of an existing literature study is based on the fact that the data, materials and code of the study are available on OSF. Making the data and code publicly available can facilitate other researchers to verify the reliability of the research results and to reproduce the experimental process, which can help to improve the transparency and reproducibility of the research, and to facilitate the knowledge sharing and communication

among researchers, but it also needs to pay attention to the measures taken to protect the intellectual property rights, data security, and privacy protection issues.

References

Aust, F., & Barth, M. (2023). *papaja: Prepare reproducible APA journal articles with R Markdown*. Retrieved from https://github.com/crsh/papaja

Barth, M. (2023). *tinylabels: Lightweight variable labels*. Retrieved from https://cran.r-project.org/package=tinylabels

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*(2), 115–147. https://doi.org/10.1037/0033-295X.94.2.115

Biederman, I., & Vessel, E. A. (2006). Perceptual pleasure and the brain: A novel theory explains why the brain craves information and seeks it through the senses. *American Scientist*, *94*(3), 247–253.

Chatterjee, A. (2022). An early framework for a cognitive neuroscience of visual aesthetics. In A. Chatterjee & E. R. Cardillo (Eds.), *Brain, beauty, & art: Essays bringing neuroaesthetics in focus* (pp. 3–7). Oxford University Press.

Farzanfar, D., & Walther, D. B. (2023). Changing what you like: Modifying contour properties shifts aesthetic valuations of scenes. *Psychological Science*, *34*(10), 1101–1120.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press.

Palmer, S. E., Schloss, K. B., & Sammartino, J. (2013). Visual aesthetics and human preference. *Annual Review of Psychology*, *64*(1), 77–107.

R Core Team. (2024). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

414

Appendix

Division of labor among group members

Leader	Guo Kaitong		
Members	Deng Xinyu; Huo Yining; Shi Yinuo; Zhang Yan		
Division of labor			
Data Analysis	Deng(25%);Shi(25%);Zhang(20%); Huo(17%);Guo(13%)	PPT production	Deng(40%);Shi(30%);Zhang(15%); Huo(15%)
Text report production	Deng(20%);Shi(20%);Zhang(20%); Huo(20%);Guo(20%)	PPT show	Guo(100%)

* The same student can be responsible for multiple parts; If more than one student is responsible for a content, you can indicate the percentage of contributions.

Figure 1. Division