Statistics 141SL: Final Report
Effect of Exposure on Perception of Beauty

Eustina Kim, Diana Pham, Ritesh Pendekanti, Cassandra Tai, Alex Trudel

I. Abstract

The main question we wanted to answer is if photo exposure is a significant factor in perception of beauty. The methods we used were Generalizability Theory (G-Theory) to compute the percent of variability that is contributed by photo exposure and a Randomized Block Design to control for nuisance variables in order to focus on our main effect of photo exposure. We found that exposure is not practically significant in perception of beauty. We also looked if the region of photo raters was a significant factor by running a Random Intercept Model which found that region is a statistically significant factor in perception of beauty, however, it was not practically significant. Our shortcoming was that we did not use gender and race of the photo in our final methods. We recommend repeating the same experiment using a filter on the photo itself so there is no variation between photos besides the exposure.

II. Description of Variables

We have four predictors:

- **Photo**: There are five images per participant, each with a different level of exposure. These images are then rated between 1-10 by a rater based on the rater's perception of the participant's beauty under the particular exposure level.
- **Rater**: Categorical variable of the individual rater who took the survey, between 1-200.
- **Region**: Categorical variable indicating whether the rater's IP address of taking the survey is in the US (1) or not (2).
- **Exposure**: Categorical variable of exposure applied to the photo between 1-5.
  - fstop 8
  - fstop 9
  - fstop 11
  - fstop 13
  - fstop 16

And the outcome variable:

- **Score**: Response variable with integer values between 0 and 10 to represent the rater's perception of beauty on the particular photo and exposure.

We use these predictors to create models to see if photo exposure level has a significant effect on perception of beauty. A summary of the schematic can be seen in Figure 1.

III. Description of Dataset

The dataset provided was attained through crowdsourcing from Amazon Mechanical Turk, in which 200 anonymous raters (represented as rows) graded 10 photos, each under 5 different exposures on a scale of 0 to 10. Information includes survey time taken, IP address, duration, latitude and longitude (converted into region), and their score 50 photos: 10 participants each under 5 different exposures. We then separated the column headers representing the individual photo ID into their own columns representing the individual in the picture (Photo) and the exposure applied (Exposure). Since each rater graded 50 pictures in total and each picture is now on its own row, this enlognates our new dataset into one with 10,000 rows, with the rating of the picture by each rater having its own column labeled Score. This became the long format dataset for our G-Theory analysis which can be seen below.

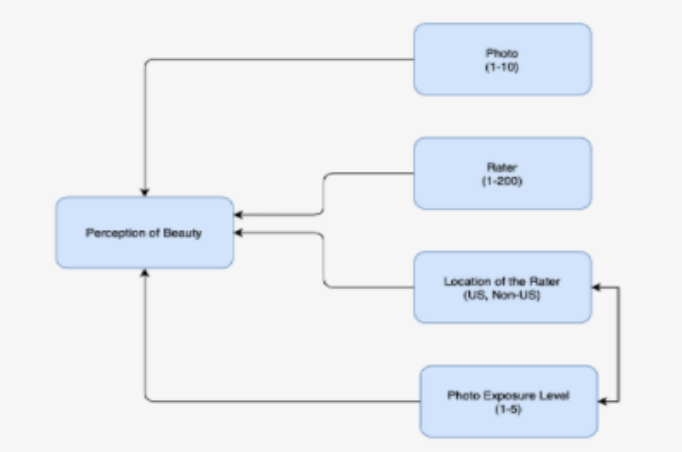


Figure 1. Schematic of variables.

Region	Photo	Exposure	Gender	Rater	Score
1	1	1	F	1	3
1	1	2	F	1	7
.	1	3	F	1	3
2	1	4	M	1	3
2	1	5	M	1	2

Figure 2. Long dataset format.

For the Randomized Block Design and Random Intercept Model, we transformed that data again. In order to eliminate the effect of photo on the scores, we averaged the scores of all 200 photos under each exposure per rater, which ended up

as a dataset with 1000 rows and columns for Region, Rater, Exposure, and Mean score (Mean). This file can be found as "mixed_data".

Region	Rater	Exposure	Mean (Across all photos for that rater/exposure combination)
1	1	1	Mean(R1, E1)
1	2	1	Mean(R2, E1)
1	3	1	Mean(R3, E1)
...	...	...	Mean(R4, E1)
...	...	1	...
2	200	1	Mean(R200, E1)

Figure 3. Mixed dataset format.

IV. Question We Are Answering

The main question we wanted to answer was if exposure is a significant factor in the perception of beauty. While working on this question, we found a few other sub-questions to look into such as:

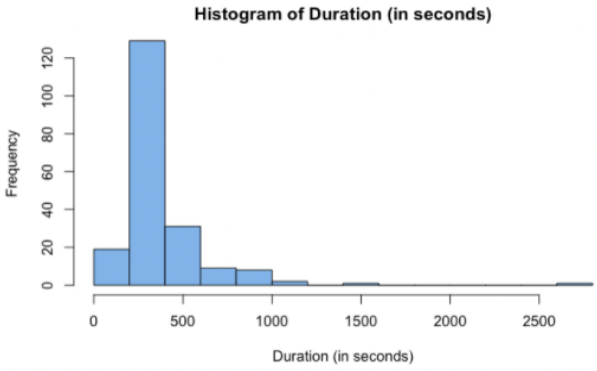
- When removing the effect of the photos, is exposure a significant factor in the graders' perception of beauty?
- Does the grader's location have an impact on their ratings for each exposure level?

V. Visualization and Exploratory Data Analysis

Outliers:

- Outliers were raters who were deemed to have answered the survey questions incorrectly. Methods of answering the survey incorrectly include:
  - Taking too little time (using a bot/computer program to choose answers).
  - Choosing answers randomly.
  - Putting the same score for all responses.
  - Having scores vary for an arbitrary reason (male/female, race, etc.).
- Several points were flagged as being outliers, but only 4 points were defined as truly "incorrect" survey responses.
- We plotted several histograms to detect outliers, which are shown below.

A. Histogram of Duration

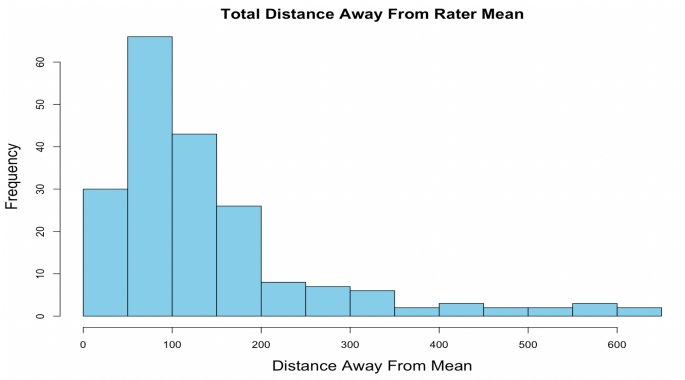


We did not remove outliers based on duration because none of the raters took too little time to finish the survey. Those who took significantly longer did not have suspicious data by any of our other metrics, and therefore are not outliers.

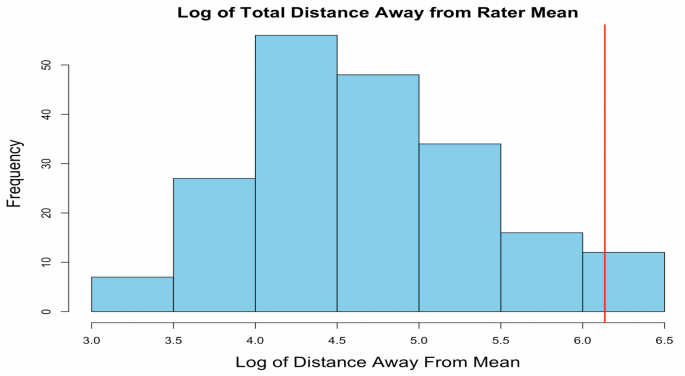
- Explanation of Statistical Terms:
  - **Sum of Squares**: The sum of squares is the sum of the square of variation, where variation is defined as the spread between each individual value and the mean
  - **Monte Carlo**: A method of simulating the distribution of an arbitrary function by repeating a series of randomized tests.

B. Histograms of Distance Away From the Mean

Sum of squares (mean = average score across all raters). For each photo exposure combination, subtract a rater's score from the average score across all raters then sum the squares. Raters with a significantly large sum of squares (after taking the log of the results) are suspected of answering questions randomly and being an outlier. The histogram below shows the distribution of the sum of squares.

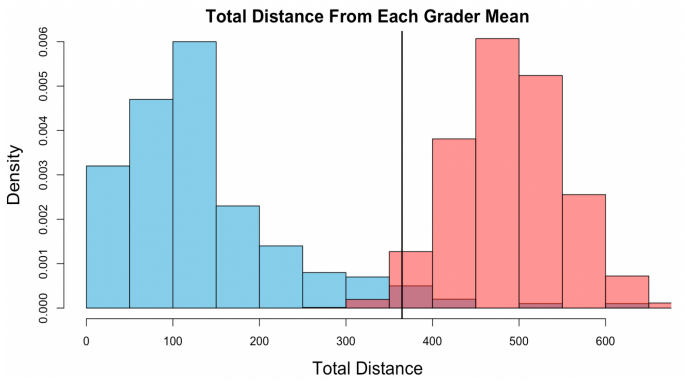


We detected outliers by doing log transformation then considering points outside of 2 standard deviation, which is represented by the red line in the graph below.



Sum of squares (mean = average score of each rater). For each photo exposure combination, subtract the rater's score from their own overall average. If the RSS was at or near 0,

then it meant that the rater scored each photo the same. Raters with high sum of squares values that intersect with the Monte Carlo (n = 100,000) simulation of randomness are suspected of randomly putting scores, and therefore being outliers.The red histogram above is the distribution of sum of squares found using Monte Carlo simulation. Anything that falls to the right of the black line (2 SDs away from the mean) is suspected of being an outlier.



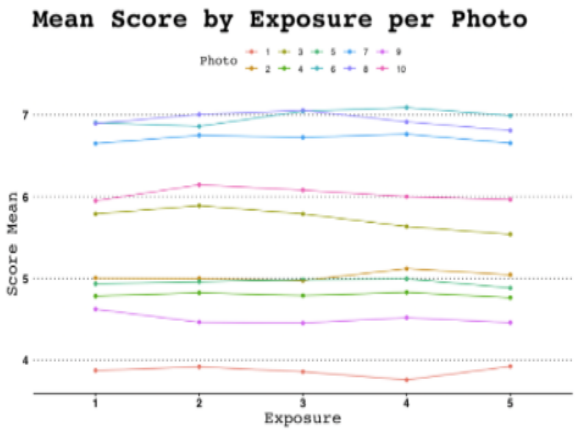
The four points we removed include rater 62, 70, 103, and 151, however we performed our analysis on both the dataset with and without these outliers and obtained the same result.

C. Frequency Table of Regions

Region	Count
1	129
2	71

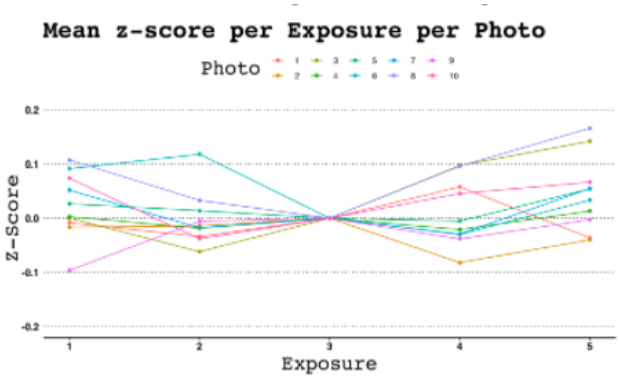
The frequencies of the different regions are adequate since neither of them are extremely low.

D. Plot of Mean Scores Based on Exposure level For Each Photo



Each line represents a photo. Overall, there is no significant pattern in the change in mean score between different exposure levels for all photos.

E. Plot of Mean Z-Scores Based on Exposure Level for Each Photo



Above is a plot of average z scores with exposure 3 as the baseline. Each line represents a photo. There's still very little difference in the scores across different photos because the z-scores are between -0.1 and 0.2. From this plot, we can hypothesize that maybe regardless of photo, the average perception of beauty across all exposures are the same.

VI. Statistical Analysis

A. G-Theory

G-Theory studies the reliability/reproducibility of the outcome by analyzing different sources of variation by quantifying the amount of inconsistencies caused by each source and their interactions. For our analysis, it helped us explain how much of the variability in the rating of the photos came from exposure.

- **Assumptions**: Randomly parallel tests sampled from the same population.
- **Strengths**:
  1) Able to estimate multiple sources of inconsistencies and identifying which sources of error are the most serious.
  2) Allows you to examine all of the different sources of variability in a single analysis, including the interactions between sources.
- **Shortcomings**: Accuracy of the estimates may be unreliable if the sample size is not large enough.

Sources of Variation	Percent Total Variance
Photo	29.139%
Exposure	0.009%
Rater	25.172%
Photo:Exposure	0.095%
Photo:Rater	35.215%
Exposure:Rater	0.221%
Photo:Exposure:Rater:Residual	10.149%

From the table above, we can see that the exposure does not account for much of the total variance in scores. In other words, the exposure of the photos does not have much effect on the rating of the photo. We also noticed that the photos accounted for thirty percent of the total variance so we decided to run G-Theory again after setting exposure level 3 as the baseline by subtracting its score from the other exposure levels for each rate and photo.

Sources of Variation	Percent Total Variance	Updated Percent Total Variance
Photo	29.139%	0.142%
Exposure	0.009%	0.050%
Rater	25.172%	0.000%
Photo:Exposure	0.095%	0.558%
Photo:Rater	35.215%	45.923%
Exposure:Rater	0.221%	1.412%
Photo:Exposure:Rater:Residual	10.149%	51.915%

Above is the G-theory results with the updated percent total variance on the right column with exposure level 3 as the baseline. And we can see that exposure still does not account for much of the total variance in scores.

B. Randomized Block Design

A randomized block design allows for subjects to be placed into subgroups called blocks, such that the variability within the blocks is less than the variability between the blocks. For our analysis, the blocks are each individual rater that scored the photos, and we also removed the effect from the photos by averaging all of the photos' scores since the spread of all ratings within all photos was immense. This design reduces variability within treatment conditions and potential confounding, producing a better estimate of treatment effects.

The following is the data structure we used to run our analysis:

Rater	E1	E2	E3	E4	E5
1	Mean(R1, E1)	...	...	...	Mean(R1, E5)
2	Mean(R2,E1)				...
3	Mean(R3,E1)				...
...	...				...
199	Mean(R199,E1)				Mean(R199,E5)
200	Mean(R200,E1)	...	...	...	Mean(R200,E5)

- **Assumptions**: This design assumes zero interaction between blocks and treatments. We also assume that Rater is random.
- **Strengths**: With an effective blocking variable (a blocking variable that is strongly related to the dependent variable but not related to the independent variable), the design can provide more precision than other independent groups designs of comparable size.
- **Shortcomings**: The design assumes zero interaction between blocks and treatments. If an interaction exists, tests of treatment effects may be biased.

After reviewing the results above, we can see that exposure is statistically significant after removing the effect of the photos since its F-value is large and p-value is less than 0.05. Despite the large F-value, we also noticed that the Sum of Squares for exposure is extremely small when compared to the Sum of Squares for Raters. Meaning, exposure is not practically significant in the perception of beauty even though it's statistically significant when compared to the effect of raters and photos.

$$\eta_p^2 = \frac{SS_{Exposure}}{SS_{Exposure} + SS_{Residual}} = \frac{0.80}{0.80 + 39.76} \approx 0.01972387$$

We calculated practical significance by calculating Partial eta-squared, which is the sum of squares of our treatment divided by the sum of squares of our treatment plus the sum of squares of the residual. Its effect size thresholds are as follows: $small \approx 0.01$, $medium \approx 0.06$, and $large \geq 0.14$.

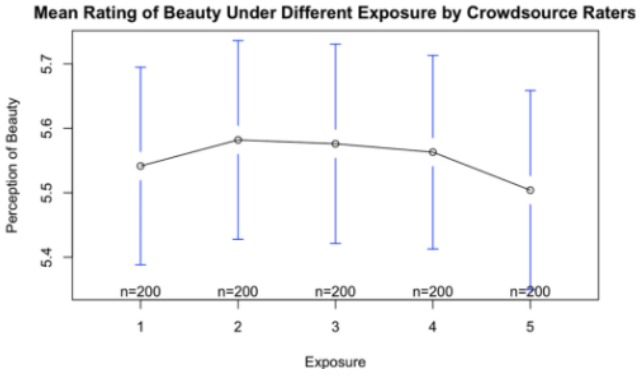
Our resulting partial eta-squared value was about 0.02, meaning, it has a relatively small effect on the total variation of the perception of beauty even after removing the effect of photos and blocking the raters.

Even though exposure was not practically significant, we still conducted post-hoc analysis using pairwise t-tests using the Bonferroni correction since it was statistically significant.

Table of P-values (<0.05 = significant)

	Exposure 1	Exposure 2	Exposure 3	Exposure 4
Exposure 2	1	-	-	-
Exposure 3	1	1	-	-
Exposure 4	1	1	1	-
Exposure 5	1	1	1	1

The above table concludes that there is no significant difference across exposure levels in the perception of beauty since a p-value of less than or equal to 0.05 signifies significance, but all of the p-values are 1, which is the largest value it can be.



Above is a graphical representation of our post-hoc analysis results which agrees with our conclusion from the table of p-values. We cans see that the means are relatively the same (within a 0.1 difference of each other) and the blue bars for each exposure level, which represents two standard deviations from the mean, is also roughly the same. In other words, the ratings across all of the exposure levels essentially have the same mean and standard deviation and no overall significant difference.

C. Random Intercept Model

We decided to include the region of the rater into our analysis by creating a random intercept model which lets the intercept (Rater) vary across all values of the random variable, but keeps the slopes fixed (Exposure and Region). A mixed design similar to our Randomized Block Design, but helps us analyze the different Exposure levels in each Region.

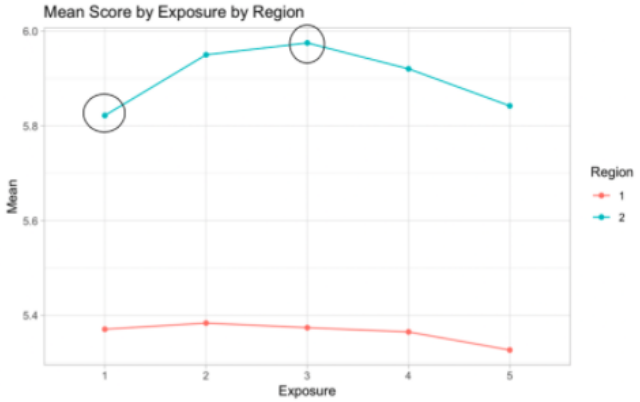
- **Assumptions**:
  1) Level 2 residuals for different groups are uncorrelated.
  2) Level 1 residuals for different observations are uncorrelated.
  3) Level 2 and Level 1 residuals are uncorrelated.
  4) Residuals and covariates are uncorrelated

  Note: Level 2 variables represent grouping variables for the observations and Level 1 variables represent individual observations.
- **Strengths**: Can be extended to higher-level models with repeated observations within individuals within clusters
- **Shortcomings**: Unable to estimate the effect of any variable that does not vary within clusters, which holds for all grouping variables.

	Estimate	P-value
Intercept	5.369767	<2e-16
Region2	0.580937	0.000317
Exposure1	0.011628	0.675042
Exposure2	0.014729	0.595406
Exposure4	0.006202	0.823066
Exposure5	-0.04031	0.14637
Region2:Exposure1	-0.129938	0.005359
Region2:Exposure2	-0.024588	0.597376
Region2:Exposure4	-0.054089	0.245437
Region2:Exposure5	-0.090267	0.055426

In the above table, the intercept uses region 1 and exposure level 3 as the baseline, so the estimate of 5.37 for the intercept is the mean score of photos under exposure level 3 rated by raters from region 1. As for the other intercept estimates, they are the differences between the score for that level and the baseline intercept.

Overall, we can see that the difference in scores between the regions is significant and specifically under region 2, the difference between exposure level 1 and 3 are also significant.



The plot above is a graphical representation of the results from the table. We can see that the difference in men scores between the two regions is significant since region 2 (green line) is above region 1 (red line), and when we look specifically at region 2, we can see that the difference between exposure levels 1 and 3 (the circled points) is statistically different.

$$R^2_{\text{GLMM}(m)} = \frac{\sigma_f^2}{\sigma_f^2 + \sum_{l=1}^{u} \sigma_l^2 + \sigma_e^2 + \sigma_d^2}$$

Similar to the Randomized Block Design, we decided to calculate practical significance using the above equation for Marginal R-squared, which measures the proportion of variance that is explained by the fixed factor(s) alone.

The fixed factors for this analysis were exposure and region and resulted in a very small Marginal R-squared value of 0.0.52, meaning only 5.2 percent of the variance in the perception of beauty is explained by photo exposure and region of the rater.

Overall, from our random intercept model, we can conclude that 1) exposure is not a significant factor in the perception of beauty and 2) region has an effect on perception of beauty, but the effect is very small since the marginal r-squared value was low.

VII. Overall Conclusions Drawn from the Study

From the study, we were able to conclude that exposure is not a significant factor in perception of beauty. However, after looking into different regions of the rater, we saw that region was significant, but the interaction between region and exposure was not significant.

VIII. Suggestions for Further Research

Since one of the applications of this experiment was to understand whether or not the exposure filters on Instagram actually affect the perception of beauty, another study could be conducted replicating this “Instagram Method” by taking only one picture of the participant and applying different exposures afterward. This would help eliminate potential variance due to

different alignment and facial expressions across exposures of the same participant. Another suggestion would be because we had a limited dataset with only a few variables, another study could be conducted to include more information, such as the gender and race of the raters. Additionally, data should be provided on the order in which the photos were scored per rater. Because for this study the order was randomly assigned, we have no definite way of identifying consistent scoring (i.e. if someone got bored halfway through and decided to put the same answer over and over, we would have no way of knowing).

IX. Shortcomings of the Study

One shortcoming was that we had averaged scores through all photos for a particular exposure per rater, but we did include a final statistical model that kept the photo variable and averaged out the scores of all Raters. Furthermore, this method of using averages reduces the number of observations by a factor of 10, potentially masking away important information and also does not include other important information such as the spread of values. We also did not include a final statistical model including the Gender and Race of the individuals in the photos that could test additional hypotheses.

X. Appendix

For appendix and presentation, click here.