Statistical Analysis of the OCEAN Personality Model

Daniel Bromberg 215776477, daniel.bromberg2005@gmail.com
Moussaffi John Meir 322923244, johnhmm2001@gmail.com

Bar Ilan University, Department of Mathematics

Abstract

This study investigates differences in the Big Five personality trait *Openness* across racial groups using a publicly available dataset from OpenPsychometrics.org. We aim to determine whether certain groups have significantly higher or lower median Openness scores and to quantify the practical magnitude of these differences. Given the ordinal and non-normal nature of Likert-scale scores, the analysis employs nonparametric techniques, including the Kruskal–Wallis test for overall group differences, nonparametric ϵ^2 as an effect size, pairwise permutation tests on median differences with Holm's correction, and Cliff's δ to assess stochastic dominance. Results show statistically significant differences in median Openness, with Europeans and Africans generally ranking highest and Middle Eastern, Indian, and Southeast Asian groups lowest. However, all observed effect sizes are small ($|\delta| \leq 0.31$, $\epsilon^2 = 0.053$), indicating substantial overlap between distributions. These findings underscore the distinction between statistical and practical significance and highlight the importance of effect size reporting in personality research.

1 Introduction

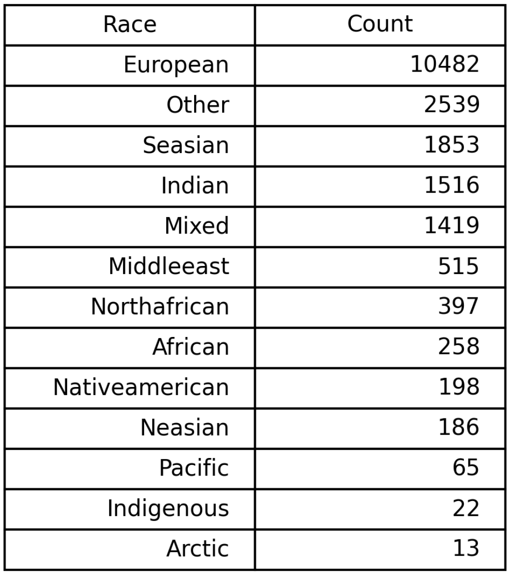
This project examines how the Big Five trait "Openness" varies across racial groups using survey data from OpenPsychometrics.org. Our aim is to identify which groups are statistically indistinguishable and which differ, and to quantify the magnitude of those differences.

Linking personality to demographics is common, but conclusions depend on the model and methods used. A brief historical note: in the 1940s William H. Sheldon proposed somatotypes (ectomorph, mesomorph, endomorph) and claimed links between physique and temperament; the approach has since been criticized for subjective measurement, biased samples, and weak evidence, and is not used today. By contrast, the Big Five (OCEAN) model is dimensional, emerging from lexical studies and later factor analysis, and is the dominant descriptive framework in contemporary personality research. Because correlational analyses do not establish causation, we treat our findings as descriptive patterns rather than causal claims.

2 Dataset

The dataset used in this project was obtained from OpenPsychometrics[1], an open-source platform that provides raw data from online personality tests. The BIG5 (5/18/2014) dataset includes anonymous responses to a 50-item Big Five personality questionnaire, rated on a 1–5 Likert scale, along with demographic information such as age, gender, race, and country of origin. Additionally, the dataset comes with a text description of the personality questionnaire, with the full list of questions that appeared in it.

Each of the five traits—Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism—is represented by the mean score of 10 corresponding items. We'll treat these trait scores as estimates of the continuous personality variables. E.g., for a person whose answers in the ten questions for Agreeableness are $[1, 2, 5, 5, 3, 2, 3, 2, 1, 5]$, his score would be $\overline{A} = 2.9$. Some items were reverse-worded (e.g., "I don't talk a lot" for Extraversion) so they were reverse-scored by replacing the value with $6 - x$ before averaging.



Race	Count
European	10482
Other	2539
Seasian	1853
Indian	1516
Mixed	1419
Middleeast	515
Northafrican	397
African	258
Nativeamerican	198
Neasian	186
Pacific	65
Indigenous	22
Arctic	13

We can observe that some races are more represented in the dataset than others. We put all the races with under 200 participants and mixed under “Other”.

Figure 1: Number of participants of each race

3 Results

The null hypothesis is that the median of ’Openness’ is the same in all races. We want to see which races are more open then the others and if that’s statistically significant. First, we want to check the normality of openness.

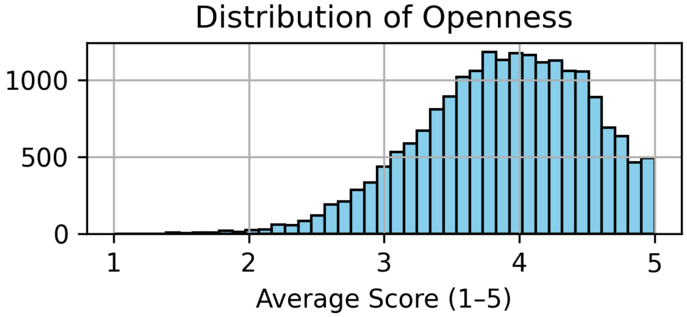


Figure 2: Openness histogram

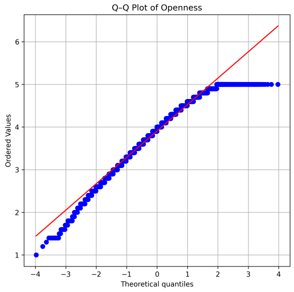


Figure 3: Opennes QQplot

There is a long left tail and the right stops with a thick end at 5, so the distribution is asymmetric. In the QQplot we can see that the middle looks fine but the tails are not on the line. For further validation, we’ll check the distribution of the Openness trait in each race separately.

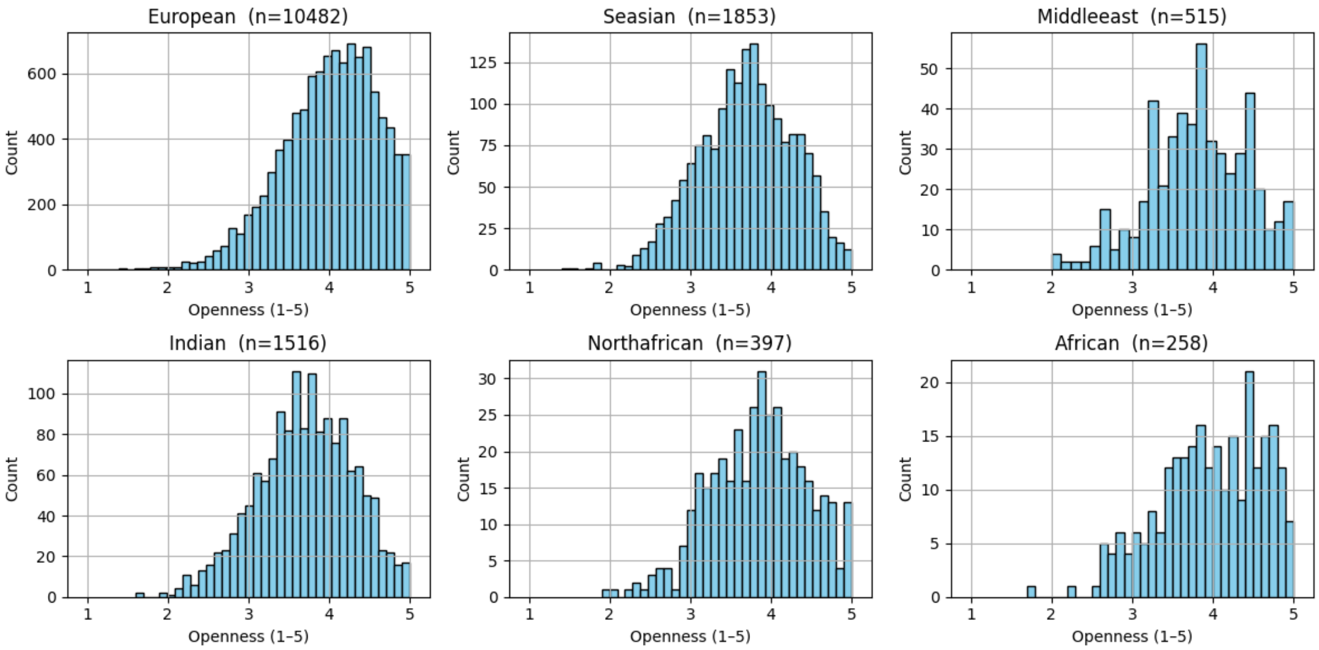


Figure 4: Histograms of openness by race

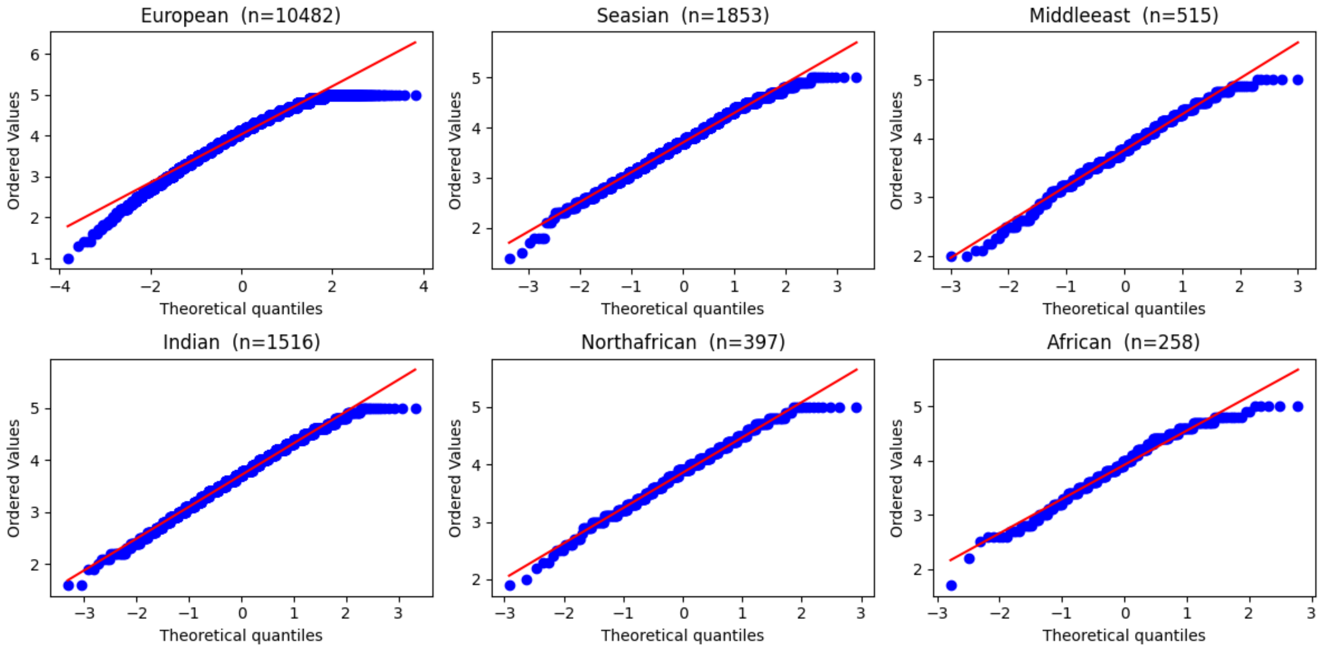
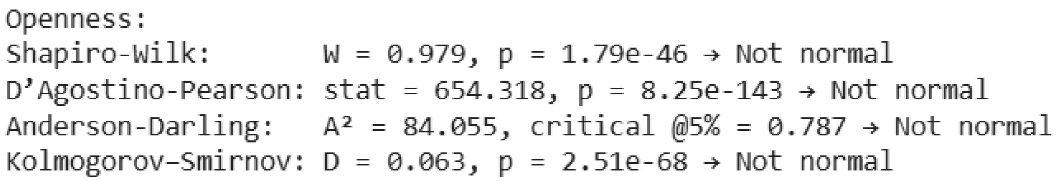


Figure 5: QQplots of openness by race

We can see similar results when we look at openness by race - the middle looks normal while there is a very thick tail at 5 and a long tail to the left. let's look at normality tests for openness:

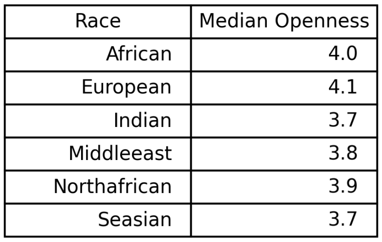


```
Openness:
Shapiro-Wilk:       W = 0.979, p = 1.79e-46 → Not normal
D’Agostino-Pearson: stat = 654.318, p = 8.25e-143 → Not normal
Anderson-Darling:   A² = 84.055, critical @5% = 0.787 → Not normal
Kolmogorov-Smirnov: D = 0.063, p = 2.51e-68 → Not normal
```

Figure 6: Normality tests results

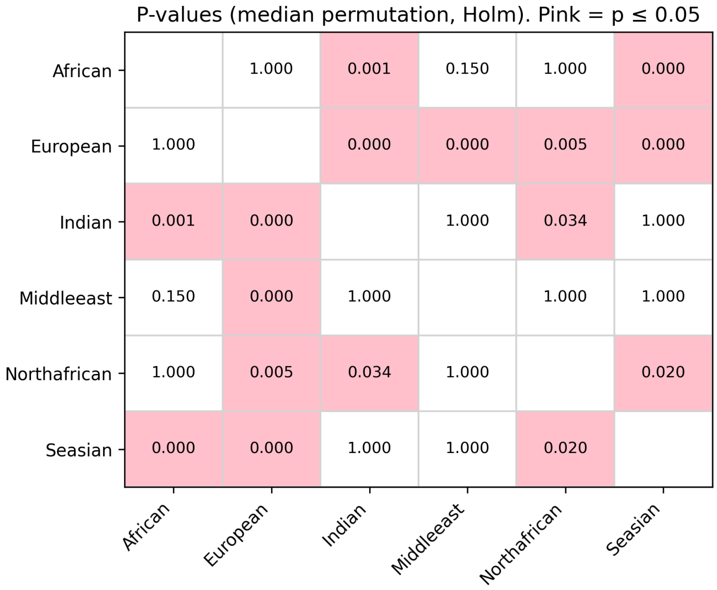
As we can see, according to all the 4 tests openness isn’t normally distributed. In conclusion, all the methods showed openness isn’t normally distributed, therefore we’ll use a parametric tests. We calculated non parametric epsilon squared for openness by race and got $\epsilon^2 = 0.053$ with p value of 0, which means only about 5% of the variability in the ranked scores is associated with race. evem thought the p value is 0 it doesn’t have much practical impact.

We used Kruskal–Wallis test if there is a significant difference between the groups. We got a p-value of 0 so there is a difference. In order to address the gaps between the group sizes we decided to do a permutation test when the null hypothesis is that the medians between each 2 races are the same instead of U tests because the races with the big sample sizes might dominate the U tests. We used Holm p value adjustment.



Race	Median Openness
African	4.0
European	4.1
Indian	3.7
Middleeast	3.8
Northafrican	3.9
Seasian	3.7

(a) Races medians of openness



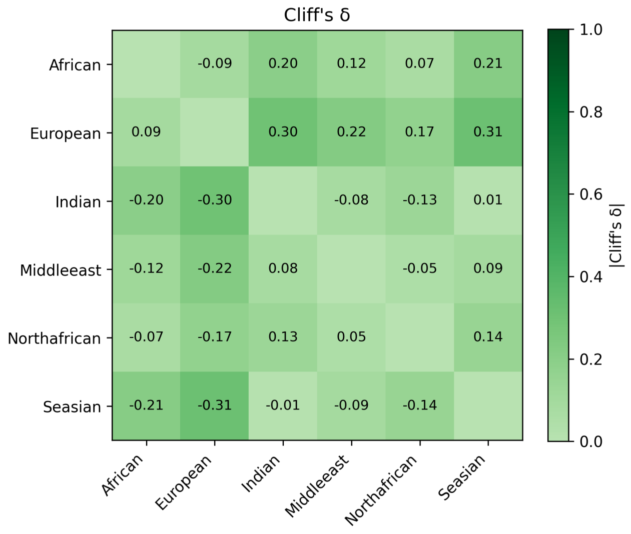
P-values (median permutation, Holm). Pink = p ≤ 0.05

	African	European	Indian	Middleeast	Northafrican	Seasian
African		1.000	0.001	0.150	1.000	0.000
European	1.000		0.000	0.000	0.005	0.000
Indian	0.001	0.000		1.000	0.034	1.000
Middleeast	0.150	0.000	1.000		1.000	1.000
Northafrican	1.000	0.005	0.034	1.000		0.020
Seasian	0.000	0.000	1.000	1.000	0.020	

(b) Holm-adjusted p-values heatmap

Figure 7: Median values and pairwise significance results.

As we can see the difference between the medians is small: the lowest one is 3.7 (Indians and South east Asians) and the highest is 4.1 for Europeans.The difference is statistically significant between all races and European except African which has p value of 1. In addition middleeasterns, Indians and South east Asians have a p value of 1 with each other. Thus Europeans and Africans competing for the title of the most open race while Middleasterns Indians and south east Asians competing for the title of the least open race. let’s look at Cliff’s delta:



What this shows
Effects are small across the board ($|\delta| \leq 0.31$), indicating substantial overlap between distributions. For example, European vs. Southeast Asian has $\delta \approx 0.31$—roughly a 63% chance that a random European score exceeds a Southeast Asian one—still a small-to-borderline-medium effect. Overall, even where p-values are significant (helped by large N), the practical differences are modest.

Figure 8: Cliff's delta

4 Methods

4.1 Cronbach's Alpha

Internal Consistency We computed Cronbach's alpha for each trait using its 10 items. The formulas of chronbach's alpha:

$$\alpha = \frac{N}{N-1}\left(1 - \frac{\sum_{i=1}^{N} \sigma_i^2}{\sigma_T^2}\right)$$

where:

- α is Cronbach's alpha (a measure of internal consistency or reliability),
- N is the number of items (e.g., questions on a test or survey),
- σ_i^2 is the variance of item i,
- σ_T^2 is the variance of the total test score (i.e., the sum of all item scores for each respondent).

All scales showed good reliability: Extraversion (0.892), Neuroticism (0.869), Agreeableness (0.832), Conscientiousness (0.813), and Openness (0.794).

4.2 Non-Parametric Epsilon-Squared

When the data are not normally distributed, a non-parametric version of Epsilon-squared (ε^2) can be used as an effect size measure after performing the Kruskal–Wallis test. It estimates the proportion of variance in the ranked dependent variable that can be explained by group membership.

$$\varepsilon^2 = \frac{H - k + 1}{n - k}$$

where:

- H is the Kruskal–Wallis test statistic,
- k is the number of groups,
- n is the total number of observations.

This statistic reflects the strength of the association between the grouping variable and the ranked outcome variable. It is suitable when ANOVA assumptions such as normality and homogeneity of variance are violated.

4.3 Permutation test (pairwise medians)

We compared groups using a two–sided *permutation test* on the *median* difference for the Openness score. For each unordered pair of races, let

$$T_{\text{obs}} = \text{median}(X) - \text{median}(Y),$$

where X and Y are the two races (with their original, possibly unequal, sizes). Under the null of *exchangeability* (no group effect), we pool the data, randomly permute labels while preserving the original sample sizes, and recompute T^* on each permuted split to form the empirical null. The two–sided p–value is

$$p = \frac{1 + \#\{\,|T^*| \geq |T_{\text{obs}}|\,\}}{B+1},$$

with $B = 30{,}000$ permutations per pair (Monte Carlo approximation).

4.4 Holm's step–down procedure

To control the family–wise error rate across multiple tests, we use Holm's method, which is uniformly more powerful than Bonferroni. Let the m raw p–values be ordered $p_{(1)} \leq \cdots \leq p_{(m)}$. Starting from the smallest, compare $p_{(i)}$ to $\alpha/(m-i+1)$. Reject $H_{(1)}$ if $p_{(1)} \leq \alpha/m$; then test $p_{(2)} \leq \alpha/(m-1)$, and so on, stopping at the first non-rejection. All hypotheses at and after that index are retained. We report the corresponding Holm–adjusted p–values for each pairwise comparison.

4.5 Cliff's delta

Cliff's δ is a nonparametric effect size for *stochastic dominance*: it quantifies how often values from group A exceed those from group B. If Cliff's δ is close to 0, then there is little to no stochastic dominance: a value drawn from A is about as likely to exceed a value from B as it is to be smaller, indicating negligible differences between the groups. If Cliff's δ is close to +1, then almost every value from A exceeds every value from B (strong dominance of A over B). If it is close to −1, then almost every value from B exceeds every value from A (strong dominance of B over A).

$$\delta \;=\; P(A > B) \;-\; P(A < B), \qquad \delta \in [-1, 1].$$

We have

$$P(A > B) + P(A < B) + P(A = B) = 1.$$

So by defenition it follows that

$$P(A > B) = \frac{\delta + 1 - P(A = B)}{2}.$$

Rule of thumb for magnitude: $|\delta| < 0.147$ negligible; 0.147–0.33 small; 0.33–0.474 medium; ≥ 0.474 large.

5 Discussion

Our analysis confirms that median Openness scores differ significantly across racial groups in the OpenPsychometrics dataset, but the magnitude of these differences is small. The non-parametric ε^2 value ($\varepsilon^2 = 0.053$) indicates that only about 5% of the variability in ranked Openness scores is attributable to race. While permutation tests with Holm adjustment identified several statistically significant pairwise differences, Cliff's δ values were consistently small ($|\delta| \leq 0.31$), reflecting substantial overlap between distributions.

These results highlight the distinction between statistical and practical significance: large sample sizes can yield small p-values even when group differences are modest in magnitude. The observed ranking of group medians—Europeans and Africans highest, Middle Eastern, Indian, and Southeast Asian lowest—should be interpreted descriptively rather than causally. The cross-sectional, self-selected nature of the dataset, combined with potential cultural, linguistic, and sampling biases, limits generalizability.

Future research could benefit from incorporating longitudinal designs, enabling the tracking of personality trait dynamics over time, and from controlling for potential demographic confounders such as age, gender, and socioeconomic status. Expanding the scope to include alternative or culturally adapted personality measures could also help address biases inherent in self-report questionnaires.

References

[1] O.-S. P. P. BIG5, "Raw data from online personality tests, answers to the big five personality test, constructed with items from the international personality item pool." 2014. [Online]. Available: https://openpsychometrics.org/_rawdata

[2] R. Vallat, "Pingouin: statistics in python," *Journal of Open Source Software*, vol. 3, no. 31, p. 1026, Nov. 2018.

[3] T. pandas development team, "pandas-dev/pandas: Pandas," Feb. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3509134

[4] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[5] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020. [Online]. Available: https://doi.org/10.1038/s41586-020-2649-2

[6] S. Seabold and J. Perktold, "statsmodels: Econometric and statistical modeling with python," in *9th Python in Science Conference*, 2010.

[7] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[8] M. Waskom, O. Botvinnik, D. O'Kane, P. Hobson, S. Lukauskas, D. C. Gemperline, T. Augspurger, Y. Halchenko, J. B. Cole, J. Warmenhoven, J. de Ruiter, C. Pye, S. Hoyer, J. Vanderplas, S. Villalba, G. Kunter, E. Quintero, P. Bachant, M. Martin, K. Meyer, A. Miles, Y. Ram, T. Yarkoni, M. L. Williams, C. Evans, C. Fitzgerald, Brian, C. Fonnesbeck, A. Lee, and A. Qalieh, "mwaskom/seaborn: v0.8.1 (september 2017)," Sep. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.883859

[9] G. Van Rossum, *The Python Library Reference, release 3.8.2*. Python Software Foundation, 2020.

[10] "Google. (2024). google colaboratory." [Online]. Available: https://colab.research.google.com

[11] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.

[12] M. Tomczak and E. Tomczak, "The need to report effect size estimates revisited: An overview of some recommended measures of effect size," *Trends in Sport Sciences*, vol. 21, no. 1, pp. 19–25, 2014.