# Did machine learning reveal symbolism, emotionality, and imaginativeness as primary predictors of creativity?

**Assessing the reproducibility and replicability of Spee et al.'s findings on creativity in Western art**

**STATS 604 - Project 3**

**Due November 1st, 2023**

# Contents

# 1 Introduction

## 1.1 Background

Creativity is a fundamental but ambiguous concept in the visual arts. It has been demonstrated that there is an association between the perceived creativity of a work of art and its overall quality [Hag+12; PLT17], but the underlying factors that shape individuals' perceptions of creativity are not well-established. This discrepancy is the central focus of a recent article published in *Nature*'s *Scientific Reports*, in which Spee et al. attempt to identify the perceived attributes of a work of art — e.g., color, symbolism, emotion — that contribute most heavily to individuals' assessments of its creativity [Spe+23].

Spee et al. conducted a study in which 78 non-experts were asked to rate 54 paintings according to 17 different attributes, each on a 100-point scale. The participants were also asked to judge the creativity of each painting on the same 100-point scale. The authors trained a random forest regression model to analyze these data, emphasizing that this approach is well-suited for capturing the nonlinear relationships that are thought to exist between creativity and the selected attributes [Mar+16; VW20]. Their results support this hypothesis of nonlinearity and suggest that symbolism, emotionality, and imaginativeness are the most prominent predictors of creativity in Western paintings. In this report, we evaluate the reproducibility and replicability of Spee et al.'s work.

## 1.2 Objectives and outline

Our analysis is guided by the following three questions:

1. Can we **reproduce** Spee et al.'s results using their published data and the procedures outlined in their article?
2. What methodological **decisions, omissions, or ambiguities** may have affected the authors' findings or our ability to reproduce them?
3. How robust are the authors' conclusions? Can we **replicate** their findings if we alter a few of their decisions and assumptions?

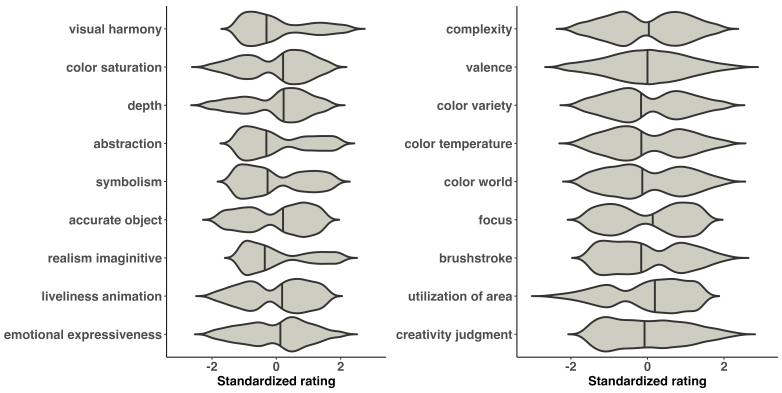
In section 2, we outline Spee et al.'s study procedure and data set in greater detail, and we successfully reproduce several of their exploratory plots. Section 3 constitutes the bulk of our reproduction efforts — we describe and implement the authors' random forest algorithm, which involves cross-validation for hyperparameter tuning and permutation testing for assessing variable importance. We obtain results that are very similar to the authors' with respect to mean absolute error and variable importance, and we reproduce their partial dependence plots. In section 4, we theorize that the authors did not adequately account for variation in creativity judgments across individuals or across painting styles. We implement two alternative approaches that account for these additional variables, and we compare the results to the authors' original findings. Finally, we conclude in section 5 by discussing the limitations and ramifications of our efforts. We reflect on what our analysis says not only about Spee et al.'s research on creativity, but also about reproducible and replicable research.

# 2 Study description and data

## 2.1 Study procedure

Spee et al. recruited a sample of 78 psychology students from the University of Vienna to participate in their study. The authors recorded the sex, age, and education of each participant, and they also had them complete a questionnaire to gauge their interest, knowledge, and experience in art [Spe+20]. These characteristics were published in a Figshare repository along with the data collected during the study [Spe+22], and we were able to verify the sample statistics reported by the authors. Approximately 70 percent of the participants were female, and their ages ranged from 19 to 35 (mean = 24.23, sd = 3.45). Their average art knowledge on a scale from 0 to 36 was 7.03 (sd = 3.75), and their average art interest on a scale from 7 to 71 was 40.00 (sd = 13.91). The authors only used this information to verify that the participants were so-called art novices, and they did not incorporate it into their analysis in any other way. This is likely defensible, as the data provide little evidence that any of the demographic traits or questionnaire answers were strongly associated with individuals' creativity judgments.

**Figure 1:** Sample distributions of standardized response (`creativity judgment`) and attributes
*(Corresponds to Figure S5 in supplement of [Spe+23]; black line indicates median)*




Each participant was shown images of the same set of 54 paintings in a random order. These paintings varied in both style (representative, impressionistic, and abstract) and genre (portrait, landscape, and still life), with an equal number of paintings in each of these categories. However, the columns corresponding to genre were empty in the authors' data set, so we were unable to consider this information in our analysis. The participants were asked to rate each painting according to 17 different attributes (see section 2.2) by dragging a slider along a bipolar 100-point scale for each attribute (e.g., for the "emotionality" attribute, the two poles of the scale were "emotionless" and "emotionally loaded"). They were also asked to rate the creativity of each painting on the same 100-point scale. A total of 4,206 sets of creativity judgments and attribute ratings were recorded — 54 for each of the 78 participants, with six observations missing due to recording errors.

## 2.2 Exploratory data analysis

The authors did not make their code publicly available, but they published their data in a Figshare repository [Spe+22]. Their data set contains 4,206 observations of the response variable `creativity judgment` and the 17 predictor attributes, which are listed in Figure 1 and Figure 2. Here, we examine these attributes by reproducing several figures shared by the authors in the supplement of their article.

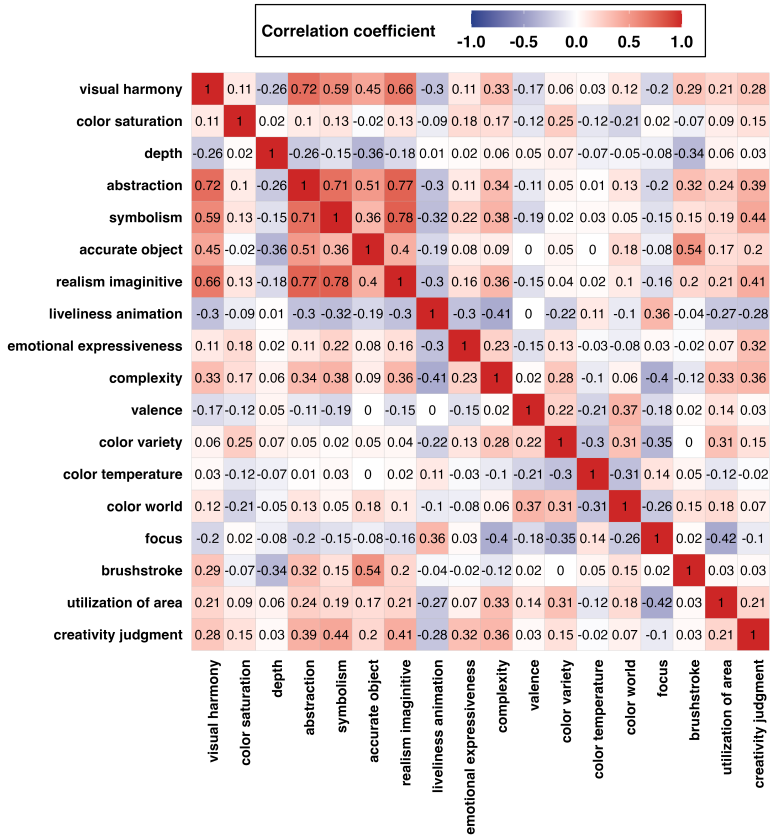
Figure 1 is our replica of Figure S5 in the authors' supplement, and we find that it matches the authors' figure almost exactly. It depicts the sample distributions of the response and the attributes, after standardization. We observe that several attributes have bimodal distributions. This suggests that many participants may have treated the attribute rating process similarly to a binary classification task — instead of calibrating their ratings across the full 100-point scale, they tended to drag the slider from the middle of the scale toward one of the poles by a consistent magnitude. This phenomenon was not mentioned by the authors, and it is not clear if or how it may have affected their results.

In Figure 2 (which is the authors' Figure S3), we examine the correlations between the attributes, as well as their correlations with the response. We observe relatively strong positive correlations between several attributes, such as `abstraction`, `symbolism`, `realism imaginative`, and `visual harmony`. We theorize that the participants may have struggled to distinguish between these more conceptual attributes.

Given this relatively large collection of attributes, one might suggest using principal component analysis (PCA) to construct a smaller set of predictors that retain much of the information in Figure 1 and Figure 2. Spee et al. implemented PCA on the correlation matrix, and they reported the proportion of variance explained by each principal component in Figure S4 of their supplement. We have exactly reproduced their PCA results in Figure 3. This figure suggests that it would be reasonable to proceed with the first four principal components — this is where the elbow falls in the left panel, and these four components explain roughly 60 percent of the total variance in creativity judgments. However, the authors instead emphasize that all but one of the components are necessary to explain more than 99 percent of the variance, and thus they decide to use the original predictors. This is a fair decision, but it does not necessarily justify their initial use of PCA.

We claim that PCA reveals additional insights about Spee et al.'s data set that the authors either failed

**Figure 2:** Correlation heatmap for `creativity judgment` and attributes
*(Corresponds to Figure S3 in supplement of [Spe+23])*




to explore or failed to report. In Figure 4, we plot the scores for the first two principal components and label the points (each corresponding to one participant's rating of one painting) by their creativity judgment and style. The left panel indicates that the first two components stratify ratings by their creativity – higher creativity ratings tend to correspond to lower scores in the first PC direction. The right panel illustrates that the first two components also separate ratings for paintings of different styles. Viewed together, these two plots suggest that `style` may explain a nonnegligible amount of the total variance of `creativity judgment`, as the patterns of separation in the two panels are very similar. This is notable because Spee et al. did not include `style` as a predictor in their random forest algorithm; they seemingly used it only to ensure that they considered a variety of artwork styles in their study. We will return to this potentially overlooked source of variability in section 4.

# 3 Reproduction of main findings

## 3.1 Algorithm

Spee et al. synthesize their procedure in the "machine learning based data analysis approach" section of their article [Spe+23]. They list the software version used for the analysis and the art attributes that serve as their predictors, and then they justify their use of random forests and provide a detailed description of the steps they took in their analysis.

The authors train a random forest regression model to predict `creativity judgment` based on the 17 attributes introduced in the previous section. They view this as a favorable method due to its robustness against multicollinearity, efficiency, and ability to capture interactions and nonlinear associations. However, the authors do not provide sufficient explanation for some of the design choices underlying their training procedure. We claim that there is significant undisclosed flexibility in their approach, especially regarding hyperparameter tuning. Their procedure is outlined in Algorithm 1.

The main loop consists of four steps. First, the data set is randomly split into train (80%) and test (20%)

**Figure 3:** Proportion of variance in `creativity judgment` explained by principal components
*(Corresponds to Figure S4 in supplement of [Spe+23])*

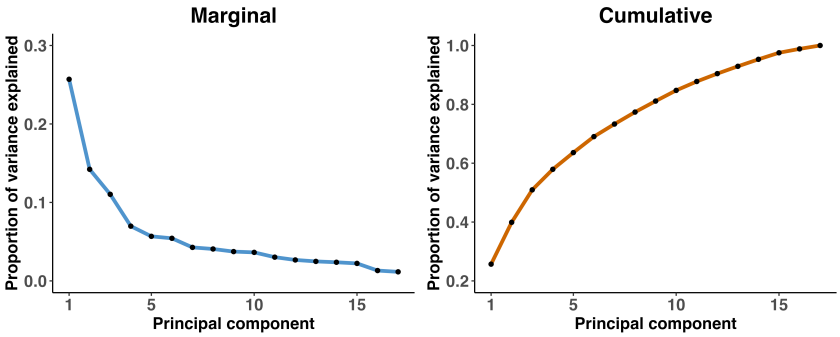
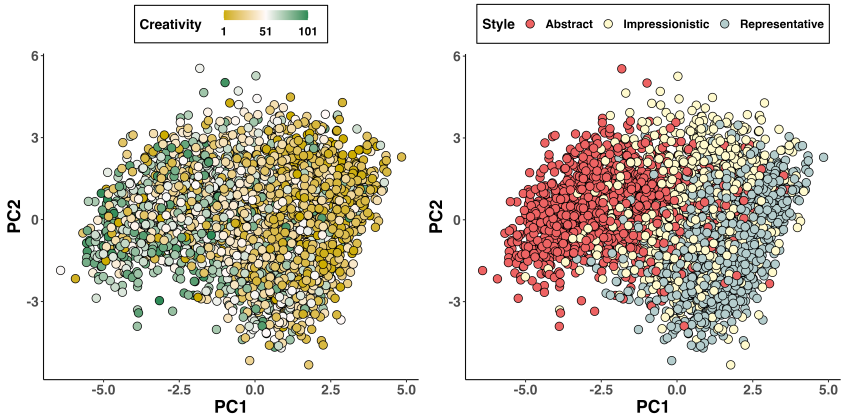



**Figure 4:** Scores for the first two principal components, labeled by `creativity judgment` and `style`




sets. Second, hyperparameter selection is conducted on the (scaled) train set using nested cross-validation and a probabilistic approach to parameter search called Bayesian model optimization (BMO). Next, with the optimal hyperparameter setting determined by the previous step, a regression forest is refitted on the whole training set. Finally, measures of goodness of fit are computed, and using this model as a baseline, permutation tests are performed on both the response and the predictors to assess the statistical significance of the goodness of fit and variable importance, respectively.

Regular cross-validation uses the same data to tune and select model parameters and evaluate model performance. This introduces bias into the procedure and may yield a model that overfits the data [CT10]. Nested cross-validation offers a more rigorous approach for hyperparameter tuning, at the cost of additional computation. It is designed to tune model parameters and evaluate model performance on different data — it uses an inner loop to fit a model to each training set and select hyperparameters over the validation set, and the estimation of generalization error occurs in the outer loop.

Regarding hyperparameter search, traditional grid search procedures can be computationally expensive and may not always converge to the best set of parameters, especially when dimensionality is high. This is where methods like Bayesian model optimization (BMO) [MM91] come into play, which offer a probabilistic approach to hyperparameter tuning. The authors use a Bayesian update procedure for an underlying Gaussian process by iteratively maximizing an objective function. This can speed up computation since it does not require searching over the whole hyperparameter space.

Finally, before the end of each outer loop, the model is assessed on the respective test set. The goodness of fit is measured with (1) the prediction coefficient of determination (prediction $R^2$), and (2) the mean absolute error (MAE). To assess statistical significance, the model is refitted using the same train set, but with the response values shuffled across data instances. This procedure is repeated multiple times, and the model's goodness of fit is deemed significant if, under the null hypothesis (that there is no association between the response and predictors, so the obtained metric appeared by chance), no more than 5% of the metrics obtained with the shuffled data are more extreme than the one obtained with the

**Algorithm 1** Random forest with hyperparameter tuning & permutation tests for variable importance [Spe+23]

**Input:** n_outer, n_inner, n_permute, proposed_points (parameter for BMO)

**for** $i$ in 1:n_outer **do**
  Train, Test ← split(data, split_ratio)
  Train_mean = mean(Train)
  Train_sd = sd(Train)
  Train = scale(Train)
  Test = scale(Test, mean=Train_mean, sd=Train_sd)

  BPP ← BMO(Train, n_inner, proposed_points, split_ratio)  ▷ BMO: Bayesian Model Optimization

  Fit a random forest with BPP on Train  ▷ BPP: Best Performing Parameters

  Measure prediction performance on Test

  Permutation test for goodness of fit with n_permute shufflings

  **for** $j$ in 1:n_outer **do**
    Permutation test for variable importance of the $j$th predictor with n_permute shufflings
  **end for**
**end for**

original data. For variable importance, a similar permutation procedure is conducted for each predictor; here, however, the shuffling is performed on the corresponding column of the test set, so there is no need to refit a model. This permutation effectively nullifies the predictor's relationship with the response. The variable importance of a predictor is hence defined as the reduction in prediction $R^2$ caused by the shuffling, and p-values can be computed using the procedure described above.

## 3.2 Implementation details and results

In Spee et al.'s analysis, they repeated the outer loop 128 times (`n_outer = 128`). In each repetition, the hyperparameters being searched over are the minimum sample size of split nodes (2 to 128), the minimum sample size of leaf nodes (1 to 128), and the maximum number of features considered at each split (1 to total number of features). To find the best-performing hyperparameters, they employed BMO with 128 inner iterations (`n_inner = 128`) and 96 initial points. Starting from each of the 96 randomly-chosen initial points, BMO sequentially attempts 128 combinations of parameters. Hence, a total of $96 \times 128 = 12{,}288$ hyperparameter settings are tested, which is substantially smaller than an exhaustive grid search. Within each outer loop, the shuffling was repeated 64 times (`n_permute = 64`) for the permutation test of the response and each predictor. The authors reported that all hyperparameters other than those specified above were set at their default values.

We attempted to replicate their analysis under the exact same settings, but we found it computationally costly to fit 12,288 regression forests in each outer loop. Given the limited time and computational resources available to us, we decided to slightly reduce the scale of the computation without sacrificing the validity of our replication. We found that qualitatively, the authors' results are quite resistant to different choices of hyperparameters. As such, we reduced the number of inner iterations and initial points to 64 and 24, respectively. All of the remaining settings are the same as those of the authors.

We summarize the prediction performance of the model in Table 1. Our results are very close to those of Spee et al. ($17.5 \pm 0.94$ and $0.30 \pm 0.05$), although our standard deviations are slightly smaller. Statistical significance was perfectly reproduced. These results indicate that the model's predictions differed from the observed creativity judgments by 17.3 points, on average. The average prediction $R^2$ was 0.33, implying that the model explains approximately 30% of the total variance in `creativity judgment`.

In terms of variable importance, in each outer iteration, we compute the average reduction of $R^2$ across the 64 permutation procedures for each predictor. We create a box plot (Figure 5) that illustrates the 128 average reductions for each predictor. This plot matches Figure 1 of the original article very closely, including the locations and heights of the boxes, the lengths of the error bars, and the exact ordering of the ten predictors with the highest variable importance. The only difference is that our figure depicts a greater degree of statistical significance for several predictors; this actually provides stronger support for their claim that symbolism, emotionality and imaginativeness are the most important attributes.

**Figure 5:** Importance of attributes for predicting `creativity judgment`
*(Corresponds to Figure 1 in [Spe+23])*

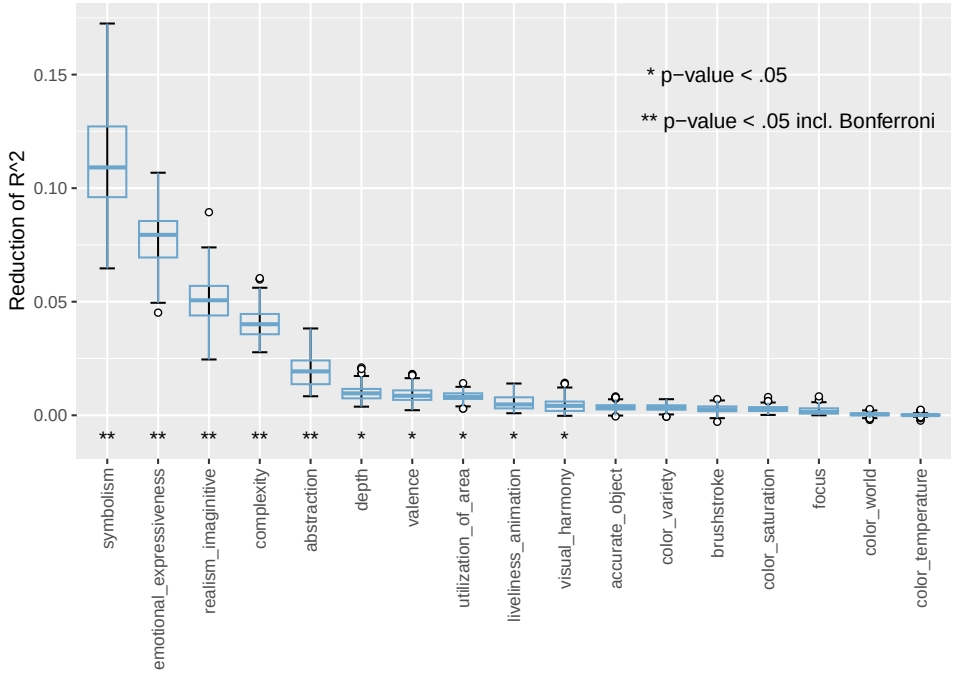



**Table 1:** Metrics for prediction performance and goodness of fit
*(Corresponds to Table 2 in [Spe+23])*

| Response | Average MAE ± sd | p-value of MAE | Average $R^2$ ± sd | p-value of $R^2$ |
|---|---|---|---|---|
| Creativity | 17.31 ± 0.44 | $p < 0.001$ | 0.33 ± 0.02 | $p < 0.001$ |

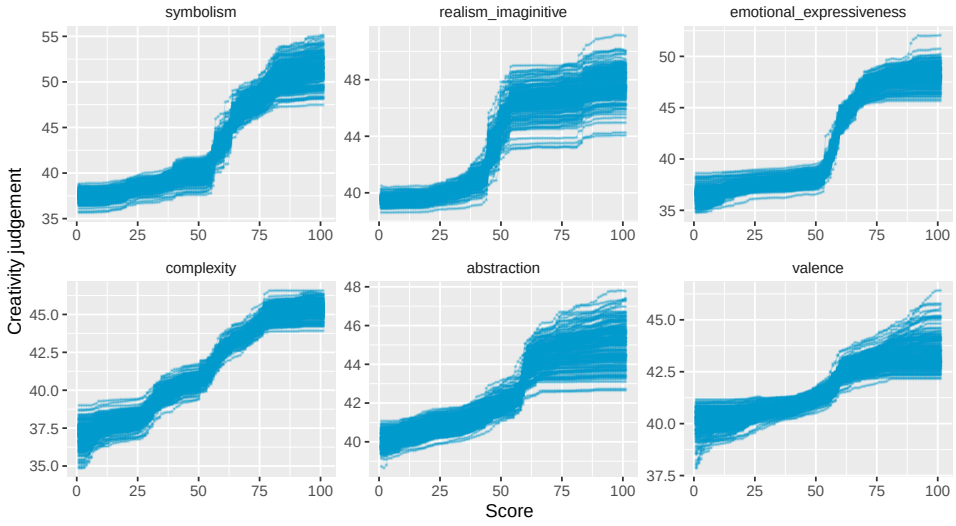
In their article, the authors constructed partial dependence plots for the six most important attributes for all 128 fitted random forests, with the goal of characterizing the relationship between `creativity judgment` and each attribute. We reproduce these plots in Figure 6. It turns out that all of the important attributes are positively associated with the response. More importantly, we confirm the authors' claim that these associations cannot be described as linear — we observe sudden nonlinear changes in Figure 6, especially for symbolism, emotionality and imaginativeness.

Setting aside any issues with model specification and feature selection, we believe that the authors' analysis is reasonable and convincing. Since their goal was to analyze and interpret the relationship between creativity and the attributes rather than to build a predictive model, the sampling randomness was properly addressed by repetition whenever it was introduced. For example, the effect of randomness introduced by data splitting was eliminated by averaging over multiple splittings, which helped improve the stability of their results. Also, since permutation tests are totally distribution-free and especially suitable for black-box methods, their justification of statistical significance was persuasive. Nonetheless, as mentioned in subsection 3.3, there are several issues in the authors' analysis. For example, two of the hyperparameters they were searching over — the minimum sample sizes of split nodes and leaf nodes — are highly related. It may have been more reasonable to tune only one of them, and additionally consider tuning the number of trees in the forest.

## 3.3 Unexplained choices

While Spee et al. exhibit a principled approach to design choices like nested cross-validation and hyperparameter tuning, there are several notable omissions and ambiguities in their exposition. For instance, their rationale for including and excluding certain attributes is somewhat subjective, and their omission of variables such as painting style raises questions about unwanted sources of variation (see section 4). Their decision to use nested cross-validation and BMO is also mostly unexplained and unjustified. They fail to discuss the trade-offs between these algorithms and their alternatives, which are

**Figure 6:** Partial dependence plots for the six most important attributes
*(Corresponds to Figures 2 and S1 in [Spe+23])*




plentiful since there exist many grid search algorithms and splitting schemes for model selection.

The authors provide no rationale behind choices like the objective function, the use of Friedman MSE as a cross-validation criterion, and the number of trees in the forest. More generally, the decision of which hyperparameters to optimize over was largely unaddressed. Furthermore, it is unclear how the authors chose the number of loops to use in their algorithm, and whether this quantity was determined a priori or if there was a stoppage criterion. This point is of special relevance when performing significance tests for variable importance [SNS11]. Lastly, the authors did not report any methodologies that they tested and found to be unsuccessful. This lack of transparency obscured the process by which they arrived at their final model, which made it more challenging for us to emulate their procedure. It also makes it harder for others to build and improve upon their work.

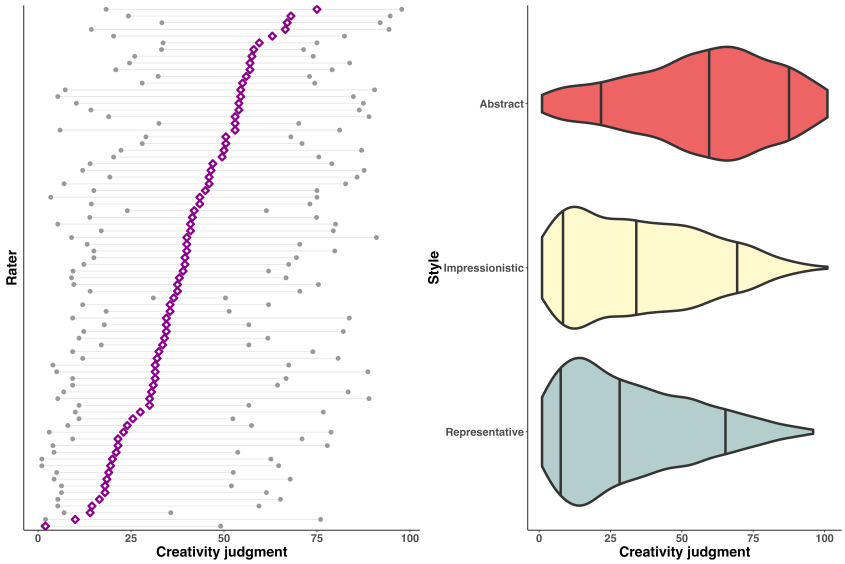
# 4 Replication of main findings

We now consider the extent to which the authors' findings resolve their main objective, which was to identify the perceived attributes of a painting that contribute to individuals' judgments of its creativity. The authors avoid making strong causal claims when discussing their results, but they frequently refer to specific attributes as "contributing to" or "having an impact" on individuals' creativity judgments. An implicit assumption underlying these conclusions is that the authors accounted for all sources of variation in creativity that they could have feasibly measured in their study — i.e., that they approximately isolated the relationship between the 17 predictor attributes and `creativity judgment`.

However, we observed a potential violation of this assumption in Figure 4, as participants' attribute ratings and creativity judgments seem to vary systematically across the three painting styles considered in the study. Specifically, ratings of abstract paintings tend to have lower PC1 scores, and lower PC1 scores correspond to higher creativity ratings. An analogous observation can be made for representative paintings, high PC1 scores, and lower creativity ratings. These patterns are corroborated in the right panel of Figure 7.

Another way in which Spee et al.'s data are not independent and identically distributed is that each participant made many ratings. We theorize that different individuals may have exhibited different tendencies when assigning attribute ratings and creativity judgments — perhaps each rater interpreted the 100-point scale in a different way. We illustrate the presence of style-to-style and rater-to-rater variability in Figure 7, where we plot the 0.1, 0.5, and 0.9 quantiles of `creativity judgment` for each rater and each style. The heterogeneity in this figure validates our concerns about these two additional sources of variability, neither of which was explicitly modeled by the authors.

**Figure 7:** 0.1, 0.5, and 0.9 quantiles of `creativity judgment` for each rater (left) and style (right)




## 4.1 Mixed-effects model

We aim to determine whether the attributes identified as important by Spee et al. will change if we reanalyze the data while accounting for variation due to `rater` and `style`. The first approach we propose for this task is a linear mixed-effects model with `creativity judgment` as the response, the 17 original attributes as fixed effects, and `rater`, `style`, and `rater:style` as random effects. One could argue that the authors' objective is better framed as an inference task than a prediction task; this is why we begin by considering a parametric model instead of a machine learning algorithm. Also, while Spee et al. detected nonlinearity in the relationships between the response and the predictors, we assume linearity in this section since it is not clear (i) how to model the form of this nonlinearity or (ii) whether this nonlinearity will persist once we account for `rater` and `style`. Finally, we claim that it is more appropriate to model `rater` and `style` as random effects than fixed effects, as their respective levels can be viewed as random samples from broader populations of raters and styles [Tay05].

We fit two candidate mixed-effects models, one that includes `rater` and `style` as random effects and another that also includes the interaction `rater:style`. We compare these nested models using a $\chi^2$-test and select the latter. Table 2 presents a summary of the selected model. We find that the random effects and fixed effects combine to explain approximately half of the total variance in `creativity judgment`, with the random effects accounting for 30 percent and the fixed effects accounting for 20 percent. The proportion of variance explained by the random effects is called the unadjusted intra-class correlation, the proportion explained by the fixed effects is called the marginal $R^2$, and the proportion explained by both is called the conditional $R^2$ [NJS17]. These results indicate that `rater` and `style` explain more of the variance in individuals' creativity judgments than the 17 original attributes, which suggests that Spee et al. failed to account for two nontrivial sources of variation even though they collected the requisite data for these variables. That being said, the fixed effects results in Table 2 are mostly consistent with those obtained in section 3.2 — for instance, `emotional expressiveness` and `symbolism` have two of the largest standardized effect sizes.

## 4.2 Incorporating additional variables into random forest

In section 4.1, we investigated the existence of individual effects and the necessity of accounting for `style`. However, note that (i) the linear mixed-effect model cannot model the potential nonlinear relationship between predictors and the response, and (ii) the metric we used to measure goodness of fit is actually the $R^2$ based on the data on which we fit the model, rather than the prediction $R^2$. These limitations make it harder to compare the results from section 4.1 with those obtained in section 3.2.

To the best of our knowledge, there is no ready-to-use package that incorporates random effects into random forest models. To address the two issues mentioned above, we simply include the factors `rater` and `style` as additional random forest predictors and redo the analysis from section 3. We now have

**Table 2:** Summary of linear mixed-effects model for `creativity judgment`

| Random effects | Variance | |
|---|---|---|
| rater | 80.06 | } Unadjusted intra-class correlation = 0.30 |
| style | 69.77 | |
| rater : style | 44.32 | |
| residuals | 315.39 | |

| Fixed effects | Estimate | Standard error | t-value | |
|---|---|---|---|---|
| (Intercept) | 41.43 | 4.95 | 8.36 | } Marginal $R^2$ = 0.20 |
| visual harmony | -2.59 | 0.47 | -5.49 | |
| color saturation | 1.29 | 0.33 | 3.86 | |
| depth | 1.38 | 0.37 | 3.75 | |
| abstraction | 1.10 | 0.57 | 1.92 | |
| symbolism | 4.38 | 0.52 | 8.50 | |
| accurate object | 0.56 | 0.45 | 1.23 | |
| realism imaginative | 1.54 | 0.56 | 2.73 | |
| liveliness animation | -1.35 | 0.36 | -3.76 | |
| emotional expressiveness | 4.92 | 0.33 | 14.91 | |
| complexity | 3.93 | 0.40 | 9.92 | |
| valence | 2.04 | 0.33 | 6.13 | |
| color variety | 0.74 | 0.36 | 2.06 | |
| color temperature | 0.20 | 0.33 | 0.61 | |
| color world | -0.49 | 0.37 | -1.32 | |
| focus | 0.24 | 0.38 | 0.65 | |
| brushstroke | -1.20 | 0.42 | -2.89 | |
| utilization of area | 0.39 | 0.36 | 1.08 | |

19 predictors — the 17 original attributes plus `rater` and `style`. We aim to determine whether this change produces any differences in prediction performance or in the importance of the art attributes. Table 3 summarizes the prediction performance of this re-implementation. We find that the difference in prediction $R^2$ between these results and subsection 3.2 is negligible. Hence, incorporating `rater` and `style` does not help explain any extra variation in the test set — i.e., it does not improve the prediction performance of the random forest algorithm.

However, if we examine the updated ordering of attribute importance (Figure 8), we find that the inclusion of `rating` and `style` *does* appear to have an impact on our results. First, almost all of the reductions of $R^2$ decrease, and they sum up to about 0.2. Since the overall prediction $R^2$ is unchanged, we infer that the proportion of the total variance explained by `rater` and `style` is approximately 0.1. This is a nontrivial proportion, and it implies that accounting for these two additional sources of variability is important. Second, looking at the ordering of the variable importance, we find that complexity surpasses imaginativeness and becomes the third most important attribute. A similar result was observed in the mixed-effects model. This suggests that the authors' original model may not have included all relevant predictors, and that their results are potentially sensitive to the inclusion of new predictors.

# 5 Discussion

While our efforts to reproduce and replicate Spee et al.'s findings on creativity in Western art were largely successful, there are several shortcomings of both the authors' analysis and our reanalysis that warrant further discussion. First, as discussed in section 3.3, the authors provided minimal justification for several design choices in their random forest algorithm, including the choice of the method itself. Other machine learning algorithms such as boosting or neural networks could have been inserted into the authors' framework without much additional effort. We did not investigate these alternative methods due to the time constraints of this project and the space constraints of this report, but it would be instructive to do so to ensure that the authors' takeaways are not a byproduct of their choice of algorithm.

Perhaps the authors' most notable oversight was their failure to control for variation in `creativity judgment` due to `rater` and `style`. This was the basis of our replication efforts in section 4. We demonstrated that accounting for these variables — either as random effects in a linear mixed-effects model or as additional random forest predictors — had a small but noticeable effect on the relative importance of the 17 attributes. For example, we found that one of the attributes mentioned in the title of Spee et al.'s paper (imaginativeness) was less important than at least one of the unmentioned attributes after we accounted for `rater` and `style`.

**Figure 8:** Importance of attributes for predicting `creativity judgment` (given `rater` and `style`)
*(Corresponds to Figure 1 in [Spe+23])*

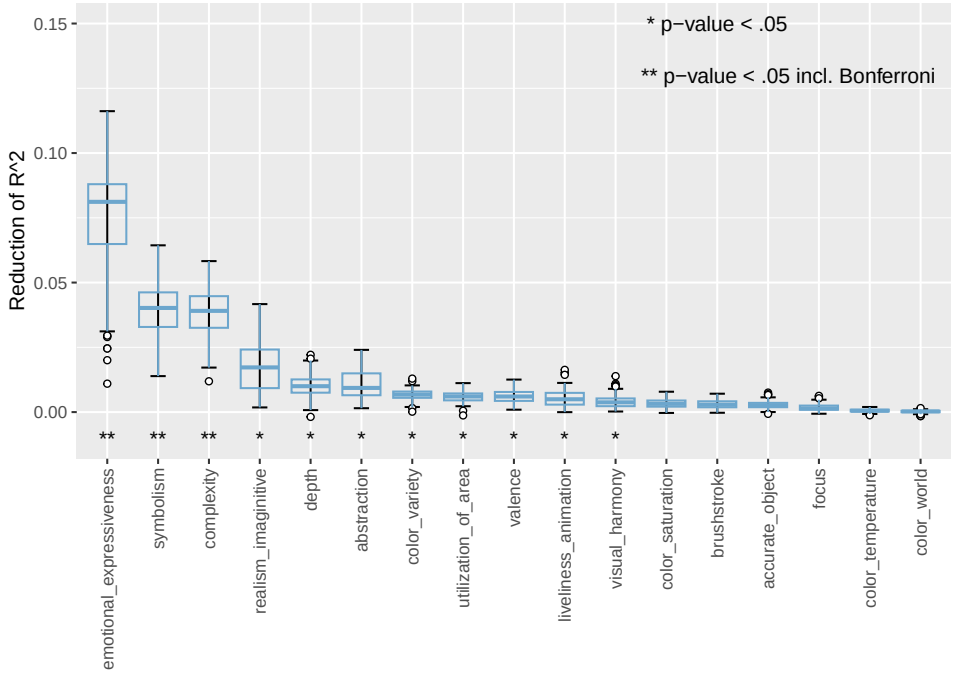



**Table 3:** Metrics for prediction performance and goodness of fit (given `rater` and `style`)
*(Corresponds to Table 2 in [Spe+23])*

| Response | Average MAE ± sd | p-value of MAE | Average $R^2$ ± sd | p-value of $R^2$ |
|---|---|---|---|---|
| Creativity | 17.29 ± 0.73 | $p < 0.001$ | 0.34 ± 0.04 | $p < 0.001$ |

Our efforts revealed Spee et al.'s conclusions to be relatively robust, but the generalizability of these conclusions is questionable due to the limited scope of the sample and study materials. The authors' description of their recruiting process implies that their sample of University of Vienna psychology students is a convenience sample, or at best a random sample drawn from a very specific population. It is reasonable to assume that older individuals, those with a different education level, or those from another city or country might have different perceptions of creativity in Western art, and it is not possible to capture the perceptions of these individuals using the authors' data. Also, Spee et al.'s study only considers creativity in Western art paintings. As such, their conclusions should not be extrapolated to the many other cultures and art forms in which creativity is a salient concept.

Finally, we express some cause for concern regarding the effectiveness of the 100-point Likert scale on which the participants recorded their creativity and attribute ratings. As we alluded to briefly in section 2.2, the data suggest that the raters may not have had the capacity or energy to provide calibrated ratings on such a detailed scale. Even if they did, they may have interpreted the scale differently from other raters — see Figure 7 for example, where the rater in the bottom row assigned a creativity rating close to zero for more than half of the paintings, while the rater in the top row assigned a median rating of nearly 75. One can account this variation via modeling as we did in section 4, but the authors may have been able to reduce or avoid it by using a simpler rating scheme in their study design.

Overall, Spee et al.'s work was very amenable to reproduction and replication, although it would have been ideal if they had provided more commentary regarding a few of the choices underlying their analysis. The authors published clean data and relatively detailed documentation, both of which greatly facilitated our efforts in this project. In the future, we recommend that they also publish their source code, as doing so would further enhance the transparency and accessibility of their findings.

# References


[CT10] Gavin C. Cawley and Nicola L.C. Talbot. "On over-fitting in model selection and subsequent selection bias in performance evaluation". In: *J. Mach. Learn. Res.* 11 (2010), pp. 2079–2107. ISSN: 1532-4435.

[Hag+12] Marieke Hager et al. "Assessing aesthetic appreciation of visual artworks—The construction of the Art Reception Survey (ARS)." In: *Psychology of Aesthetics, Creativity, and the Arts* 6.4 (2012), p. 320.

[Mar+16] Manuela M Marin et al. "Berlyne revisited: Evidence for the multifaceted nature of hedonic tone in the appreciation of paintings and music". In: *Frontiers in Human Neuroscience* 10 (2016), p. 536.

[MM91] J. B. Mockus and L. J. Mockus. "Bayesian approach to global optimization and application to multiobjective and constrained problems". In: *Journal of Optimization Theory and Applications* 70.1 (1991), pp. 157–172. DOI: 10.1007/bf00940509. URL: https://doi.org/10.1007/bf00940509.

[NJS17] Shinichi Nakagawa, Paul CD Johnson, and Holger Schielzeth. "The coefficient of determination $R^2$ and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded". In: *Journal of the Royal Society Interface* 14.134 (2017), p. 20170213.

[PLT17] Matthew Pelowski, Helmut Leder, and Pablo PL Tinio. "Creativity in the visual arts". In: *The Cambridge Handbook of Creativity Across Domains* (2017), pp. 80–109.

[SNS11] Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant". In: *Psychological Science* 22.11 (2011). PMID: 22006061, pp. 1359–1366. DOI: 10.1177/0956797611417632. URL: https://doi.org/10.1177/0956797611417632.

[Spe+20] Eva Specker et al. "The Vienna Art Interest and Art Knowledge Questionnaire (VAIAK): A unified and validated measure of art interest and art knowledge." In: *Psychology of Aesthetics, Creativity, and the Arts* 14.2 (2020), p. 172.

[Spe+23] Blanca T. M. Spee et al. "Machine learning revealed symbolism, emotionality, and imaginativeness as primary predictors of creativity evaluations of western art paintings". In: *Scientific Reports* 13.1 (2023). DOI: 10.1038/s41598-023-39865-1. URL: https://doi.org/10.1038/s41598-023-39865-1.

[Spe+22] Blanca T.M. Spee et al. "Dataset - How do we identify creative art?" In: (2022). DOI: 10.6084/m9.figshare.19097099.v1. URL: https://figshare.com/articles/dataset/Dataset_How_Do_We_Identify_Creative_Art_/19097099.

[Tay05] Jonathan Taylor. *Fixed vs. random effects*. Lecture slides, Stanford University. 2005.

[VW20] Eline Van Geert and Johan Wagemans. "Order, complexity, and aesthetic appreciation." In: *Psychology of Aesthetics, Creativity, and the Arts* 14.2 (2020), p. 135.