ESTIMATING THE RELIABILITY & ROBUSTNESS OF RESEARCH

ERROR REVIEW

Joel, S., Eastwick, P. W., & Finkel, E. J. (2017). Is romantic desire predictable? Machine learning applied to initial romantic attraction. *Psychological Science, 28*(10), 1478-1489. https://doi.org/10.1177/0956797617714580

*reviewed by*

**Florian Pargent**, LMU Munich

Feb 17, 2025

*For the sections below, indicate whether you discovered any errors using the dropdown menu. Describe the* ***errors*** *you discovered, the* ***methods*** *that you used to find them, and the* ***amount of time*** *you invested in the search. Refer to specific files to allow verification of your review. For the assessment below, make sure to check if authors have provided* ***supplementary analyses*** *as these may clear concerns arising from the (interpretation of) the primary analyses. If you have written code yourself for the review, please attach it to the report. Please indicate the version of software/packages you used to run the original code and/or your own code.*

# 

# I. METHODS, MEASUREMENT, AND DESIGN

**1. Design** No errors found

Are there errors in the conceptual design of the study? E.g., flawed randomisation technique

*Time spent: minutes/hours*

I think this study uses a strong conceptual design and is a solid example of how machine learning methods can be used to investigate predictive psychological research questions. I carefully examined the manuscript and did not find any conceptual issues.

**2. Measurement** Didn't check

Are there any measures, techniques, or devices that were incorrectly applied or inappropriate for the specific task described in the paper?

*Time spent: minutes/hours*

I did not find any issues concerning measurement. However, I did not check whether the large number of predictor variables (182 constructs in Sample A and 112 constructs in Sample B) that were preselected by the authors from an even larger item pool described elsewhere were theoretically appropriate.

**3. Preregistration Consistency** Not applicable

Are there substantial deviations from the preregistration, particularly undisclosed ones?

*Time spent: minutes/hours*

The manuscript does not mention any preregistration.

**4. Sampling** No errors found

Is there an error in the sampling strategy? Is the power analysis reproducible? Does the model used for the power analysis match the model in the substantive analyses? Were separate power analyses conducted for all primary analyses?

*Time spent: minutes/hours*

Because the manuscript focuses on predictive modeling and does not use hypothesis testing, no power analysis is reported. However, the authors used simulations to show that, in theory, their machine learning models are capable of predicting relationship desire. I re-executed the original source code to run these simulations that are reported in the supplementary Table S16 and S17 of the manuscript and did not find any inconsistencies. General details on my reproducibility checks are reported in the review sections **Computational Reproducibility of Reported Statistics** and **Reproducibility report**.

**5. Other Aspects Related to Methods and Measures**  Didn't check

*Time spent: minutes/hours*

# II. DATA, CODE, AND STATISTICAL ANALYSES

## 

**1. Code Functionality** No errors found

Does the provided code run without the need to make any adjustments and without errors? If not, what steps were needed to get it to run (if it was eventually possible)?

*Time spent: minutes/hours*

I could not find any documented software versions in the public resources. Even if those were available, I suspect that it would have been quite difficult to exactly reproduce the numerical results of the machine learning analyses after so many years. I have instead attempted to reproduce the analysis with current software versions and checked whether my results are consistent with the published numbers, that is whether the interpretation of any results and the resulting claims would differ based on my computations.

After I downloaded the data and put it into the same folder as the syntax files, I was able to run all original R scripts with current versions of R and all required R packages without any adjustments to the original source code. Following the conceptual framework by Kohrt et al. (2024), I performed a so-called *re-execution reproduction*. I documented the process and the results of my reproduction on a website (<https://florianpargent.github.io/error-review_Joel_2017/>) and the underlying Github repository (<https://github.com/FlorianPargent/error-review_Joel_2017>). I give a more detailed report of the reproduction in the review section **Reproducibility report.**

**2. Computational Reproducibility of Reported Statistics** Errors found

Is there a clear traceability of reported stats to code? Does the code output match what’s reported in the paper? Are all reported statistics findable within the analysis code?

*Time spent: minutes/hours*

I attempted to verify the consistency of all results reported in the main text of the manuscript, in Tables 2, 3, 4 of the manuscript and in the supplemental tables S16 and S17.

*Exceptions:* I did not attempt to verify the results in Table 1 of the main text because I am not familiar with the BLOCKO program that was used to compute the reported social-relations-model analyses. I also did not attempt to verify the results in the supplemental tables (except from S16 and S17) due to time constraints, as well as the descriptive statistics of the predictor variables reported in the main text on page 1481.

A more detailed report of my *re-execution reproduction* (Kohrt et al., 2024) can be found in the review section **Reproducibility report.** In summary, the reproduction clearly strengthened the support for all claims made in the original manuscript.

However, while working through the original manuscript, I detected some minor inconsistencies, that I want to document here:

1. On page 1480, the manuscript states:  
   *“Sample B consisted of 187 undergraduate students (93 women and 94 men; mean age = 19.6 years, SD = 1.2) who attended one of eight such events in 2007.”*  
     
   However, in the two data files Data/Testing actor.csv and Data/Testing partner.csv (which both contain the predictor variable *Gender*), one gender value is missing. Either the reported descriptive gender statistics in the manuscript are incorrect, or the missing data point in the testing datasets is incorrect. I have documented this inconsistency on the website of my reproducibility check here:  
   <https://florianpargent.github.io/error-review_Joel_2017/Syntax/Document%20Inconsistencies.html>
2. The abstract of the manuscript states:  
   *“Random forests models predicted 4% to 18% of actor variance and 7% to 27% of partner variance; crucially, however, they were unable to predict relationship variance using any combination of traits and preferences reported before the dates.”*However, on page 1483, the manuscript describes the results in Table 2 as:  
   *“The resulting models predicted approximately 5% to 18% of the variance in actor desire and 18% to 27% of the variance in partner desire.”*The lowest explained variance in actor desire in Table 2 of the manuscript is reported as 4.95% and the lowest explained variance in partner desire is reported as 18.48%. I cannot say with absolute certainty where the 4% and 7% come from. However, I strongly suspect that the sentences in the abstract (and in the discussion, where the same numbers are repeated on page 1486) do not only refer to the results reported in Table 2 but also to the results of the test set predictions reported in Table 3. If the authors interpreted the squared correlation between predicted and actual scores as explained variance, this would result in 4% for actor desire (0.19^2=0.0361) and 7% for partner desire (0.26^2=0.0676) after rounding. I did not find any hint in the manuscript or the source code whether the authors might have computed this squared correlation.

**3. Data Processing Errors** Indeterminable

Are there substantive errors during the preparation or cleaning of data (e.g. duplication of rows during a merge) prior to substantive analyses and hypothesis tests?

*Time spent: minutes/hours*

The authors provide only the preprocessed data in their public resources accompanying the manuscript. For this reason, I had problems verifying how the dependent variables that were used in the machine learning analyses were actually computed.

On page 1481, the manuscript states:  
*“On the interaction-record questionnaire, participants completed a three-item measure of their romantic desire for that individual: “I really liked my interaction partner”, “I was sexually attracted to my interaction partner,” and “I am likely to say ‘yes’ to my interaction partner.” These items were rated on a 9-point scale (1 = strongly disagree, 9 = strongly agree). For Sample A, α was .88 (M = 5.04, SD = 2.11); for Sample B, α was .87 (M = 4.93, SD = 1.90).”*

The provided datasets include only the combined dependent variable but not the three individual items of the aggregated measure. The manuscript never explicitly states how the three items are combined (although I would expect that the sum or mean score across the three items was computed). Because I did not have access to the three individual items, I could not reproduce the reliability estimates (Cronbach’s alpha of the three-item measure for samples A and B) reported in the manuscript.

On page 1481-1482, the manuscript describes how the dependent variable for actor, partner and relationship desire were computed:

*“We next separated each report of romantic desire (e.g., Male 1's reported desire for each of his 12 speed dates) into these three statistically independent components. First, we calculated actor desire-the extent to which the participant liked his or her speed-dating partners on average—by subtracting the romantic desire grand mean from the average of each participant's approximately 12 reports of romantic desire. Second, we calculated partner desire—the extent to which the participant was liked by his or her speed-dating partners on average-by subtracting the romantic desire grand mean from the average of the approximately 12 reports of romantic desire about that participant. Third, we calculated relationship desire-the extent to which the participant liked a particular partner above and beyond his or her actor effect and the partner's partner effect-by subtracting the grand mean, the participant's actor effect, and the partner's partner effect from the participant's report of romantic desire for that partner. In our analyses, we attempted to predict each of these three components separately.”*

Because I did not have access to the raw data, I could not verify that the dependent variables of the different prediction analyses were computed according to this description. When looking at the dependent variables in the different preprocessed datasets, the dependent variable usually has a mean close (but not identical) to 0 and a standard deviation close (but not identical) to 1. For example, the summary statistics of the dependent variable in the dataset used to predict actor desire in sample A are (<https://florianpargent.github.io/error-review_Joel_2017/Syntax/Background%20Measures%20Predicting%20Actor%20Desire.html>):  
  
> Background\_Actor\_SampleA <- read.csv(file="Level 2 predicting ActorGM, Sample A.csv" , header=1)

> mean(Background\_Actor\_SampleA$DiggingActorGM)

[1] -0.005191881

> sd(Background\_Actor\_SampleA$DiggingActorGM)

[1] 1.020518

I am wondering whether the dependent variable has not only been centered (subtract the mean) but also standardized (divide by the standard deviation) at some point in the preprocessing. If some observations had been removed afterwards, this would explain why the mean and standard deviations do not match 0 and 1 exactly. I did not find any comment on standardization in the manuscript or the syntax files.

Note that in addition to a lack of transparency, the exact computation of the dependent variable matters with respect to the interpretation of the reported MSE values in the manuscript. The MSE is an absolute performance measure that is not bounded upwards. As a consequence, the MSE is sensitive to the scale of the dependent variable, such that the computed MSE value is not the same with and without standardization of the dependent variable.

**4. Model Misspecification** Errors found

Are there any consequential issues with the assumptions or the form of a statistical model (e.g., overfitting, wrong distribution assumption) used to describe data?

*Time spent: minutes/hours*

I did *not* find any major issues with the applied machine learning models that could weaken my confidence in the claims made in the original article.

However, I have found three minor issues, where I would argue that the methodology could have been improved:

## 4.1 Undocumented handling of missing values

The manuscript does not mention whether the analyzed data contained missing values, and if so, how missing values were handled during the statistical analyses. Looking at the data shows that missing values exist both in the predictor variables and the dependent variable. Looking at the code shows that rows with missing data were automatically deleted when performing variable selection as well as training the final random forest models (na.action = na.omit). No imputation techniques or missing indicators were used. I see two potential issues with this approach:

1. The decision to remove observations with missing values instead of using some form of imputation might have affected the estimated performance of the random forest models. But more importantly, the decision how to deal with missing values in predictive research affects how the final performance estimates can be interpreted. When missing values are removed during model training instead of using imputation methods or missing indicators (see for example, Sisk et al., 2023), the final model can only be applied to observations without missing values. When missingness in the predictor or the dependent variable is related to the true value in the dependent variable (different causal mechanisms are possible here), this limits the scope of predictive performance estimates based only on complete observations. Although I do not have a strong opinion on what would have been the best solution on how to handle missing data for the current study, I feel like these issues should at least have been discussed in the manuscript.
2. Another consequence of removing observations with missing values is that the performance estimates reported in the manuscript are based on reduced and varying sample sizes. Without imputation, no prediction can be computed for each test set observation with at least one missing value in a predictor variable. As a consequence, all performance estimates reported in the manuscript are potentially based on different sample sizes. This fact should have been mentioned in the manuscript. Otherwise, the reader thinks that the sample size reported in the “Participants” section also applies to all reported random forest analyses.

## 4.2 Biased performance evaluation due to variable selection

For the random forest analyses reported in Table 2 of the manuscript, the authors always used the full dataset (of either sample A or B) when performing variable selection, training and evaluating the performance of the final model (for a short discussion of this topic, see for example Pargent et al., 2023). Although the VSURF package used for variable selection implements some strategies to avoid overfitting, not strictly separating the data used for variable selection and performance evaluation can be expected to have biased the performance estimates reported in Table 2 at least to some extent. This bias might explain some of the difference in predictive performance between the OOB estimates in Table 2 of the manuscript and the lower test set performance for sample B in Table 3. A strategy to obtain less biased, more conservative performance estimates in Table 2 would have been to use nested resampling and embed the variable selection into the performance evaluation loop. For example, one could use k-fold cross-validation in the outer loop that is used to compute MSE and explained-variance. In each training set of the cross-validation, first VSURF would be used for variable selection, and then the selected variables would be used to train a random forest model. The trained model would then be used to compute predictions for the observations in the corresponding test set. This nested procedure would ensure that the variable selection procedure never had access to the observations in the corresponding test set of the outer cross-validation loop that is used for performance evaluation.

I want to emphasize that the lack of separating the observations used for variable selection and the evaluation of predictive performance in the final models (Table 2) is not a big issue, because the manuscript also includes a true external validation using the training and testing analyses based on two independent samples (Table 3).

On a side note, if the manuscript were written today, common standards probably would demand that the performance estimates in Table 2 (and 4) would be computed based on a resampling strategy with an explicit separation of training and test data, such as k-fold cross-validation or subsampling. Now that computations have become faster and cheaper, and there have been discussions of bias for specific types of prediction tasks (e.g., Janitza & Hornung, 2018), the classical OOB estimate for random forests is rarely used anymore in practice (except for the context of variable important measures).

## 4.3 Biased performance evaluation due to dependent observations

The manuscript correctly mentions that in models predicting relationship desire where data sets were organized at Level 1, some rows are not independent because they belong to the same person. This dependency has not been fully taken into account when computing the OOB performance estimates for relationship desire reported in Tables 2 and 4 (for a short discussion of this topic, see for example Pargent et al., 2023). The methodological literature has increasingly emphasized blocked resampling (Roberts et al., 2017) as a strategy to produce more realistic performance estimates in cases of dependent observations. Conceptually, the performance estimates in Table 2 can be thought to simulate how well the model would predict the relationship desire for a new speed-date in which the two participants are already known (the participants were already part of the training dataset), but the exact pairing of participants is new (the two participants had not dated each other in the training dataset). This is a different applied setting compared to the scenario where the trained model is requested to predict the relationship desire for a speed-date of two completely new participants. The predictive performance of the second scenario is what is estimated with the training testing analyses using the independent samples A and B in Table 3. Because the participants were different in both samples, using the model trained on sample A to predict the relationship desire for an observation from sample B reflects the applied setting where a prediction is made for a new speed-date with new participants the model has not been trained on previously. This conceptual difference might explain some of the difference in predictive performance between the OOB estimates in Table 2 of the manuscript and the lower test set performance for sample B in Table 3.

I want to emphasize that the lack of using blocked resampling is not a big issue, because the manuscript also includes the training testing analyses based on two independent samples. These analyses reported in Table 3 provide by far the strongest evidence in the manuscript for the claim that predicting actor and partner desire could be possible in dating practice. The test set estimates of sample B in these analyses do not suffer from the dependency issue and can be considered more realistic estimates for the most relevant applied setting of predicting completely new speed-dates. I also want to note that blocked resampling was rarely discussed at the time when the manuscript was published in 2017. However, if the manuscript were written today, I would expect a more rigorous discussion of this important issue when using resampling estimates of predictive performance.

**5. Erroneous/Impossible/Inconsistent Statistical Reporting** Errors found

Are there inconsistencies between test statistics, degrees of freedom, and p-values? Are there implausible degrees of freedom between compared SEM models? Are there point estimates outside the confidence interval bounds?

*Time spent: minutes/hours*

Because the manuscript focuses on machine learning analyses, there are no p-values reported (except from the last column in Table 3 of the main text) and only a small number of confidence intervals (in the “Training and testing analyses” section of the main text). I did not find any major inconsistencies.

However, while working through the original manuscript, I detected some minor reporting mistakes, that I want to document here:

1. On page 1485, the manuscript states:  
   *“In total, Sample A included 36 predictors (18 male and 18 female predictors) and 958 rows, and Sample B included 38 predictors (20 male and 20 female predictors) and 1,092 rows.”*I am confident that this sentence should instead read:  *“In total, Sample A included 36 predictors (18 male and 18 female predictors) and 958 rows, and Sample B included 40 predictors (20 male and 20 female predictors) and 1,092 rows.”*

I checked the dimensions of the corresponding datasets (<https://florianpargent.github.io/error-review_Joel_2017/Syntax/Post-Interaction%20Measures%20Predicting%20Actor,%20Partner,%20and%20Dyadic%20Desire.html>) e.g.:

> PostI\_malerel\_SampleB <- read.csv(file="Level 1 Post-Interaction predicting Male Dyadic Desire, Sample B.csv", header=T)

> dim(PostI\_malerel\_SampleB)

[1] 1092 41

The dataset includes 41 columns: 40 predictor variables and 1 dependent variable.

1. On page 1483, the manuscript states:  
   *“In contrast, models predicted between −4.55% and −0.18% of variance in men’s desire for women, and between −2.68% and 1.30% of variance in women’s desire for men.”*

However, Table 2 of the manuscript reports the amount of explained variance in women’s desire for men in the condition with stringent variable selection as 1.34%.

Because I cannot identically reproduce the authors computations with the precision required here (see the section **Reproducibility report**), I cannot determine which of the two numbers was actually obtained by the authors.  
  
In my own computations with current package versions, I obtained 1.32% of explained variance (<https://florianpargent.github.io/error-review_Joel_2017/Syntax/Background%20Measures%20Predicting%20Dyadic%20Desire.html>), which is very close to both numbers reported in the manuscript. Thus, I am confident that this inconsistency in the manuscript is only a copy mistake.

**6. Other Aspects Related to Data or Code**  Not applicable

*Time spent: minutes/hours*

# III. CLAIMS, PRESENTATION, AND INTERPRETATION

## 

**1. Interpretation Issues** No errors found

Throughout the entire paper, is there an incorrect substantive interpretation of data or statistical tests, causal inference issues, etc.?

*Time spent: minutes/hours*

I carefully examined the manuscript and did not find any interpretation issues.

**2. Overclaiming Generalisability** No errors found

Does the paper overclaim the generalisability of the findings with regards to stimuli, situations, populations, etc.? Is there hyping or overselling of the importance or relevance of findings?

*Time spent: minutes/hours*

I carefully examined the manuscript and did not find any problems with respect to overclaiming generalisability.

**3. Citation Accuracy** Didn't check

Are there misrepresentations of substantive claims by cited sources? Inaccurate direct quotes? Incorrectly cited or interpreted estimates? Citations of retracted papers?

*Time spent: minutes/hours*

I did not check all cited sources in the paper because I am not an expert on psychological theories of romantic attraction. For the methodological references I am more familiar with, I did not find any misrepresentation or incorrect interpretations.

**4. Other Aspects Related to Interpretation**  Didn't check

*Time spent: minutes/hours*

## 

## Reproducibility report: a re-execution reproduction of Joel et al. (2017)

## 

To check the computations reported in Joel et al. (2017), I performed a *re-execution reproduction* following the conceptual framework byKohrt et al. (2024). In this section, I provide a brief reproduction report loosely inspired by their reporting guidelines. The details of how the reproduction was executed is documented at <https://florianpargent.github.io/error-review_Joel_2017/> and the accompanying Github repository (<https://github.com/FlorianPargent/error-review_Joel_2017>):

*Reproducers' expertise:* Expert (I am familiar with the type of research, research methods and machine learning algorithms).

*Investigated work:* Joel, S., Eastwick, P. W., & Finkel, E. J. (2017). Is romantic desire predictable? Machine learning applied to initial romantic attraction. *Psychological Science*, 28(10), 1478-1489. https://doi.org/10.1177/0956797617714580.

*Investigated claims*: I investigated all major claims from the original article.

*Investigated result:* I attempted to verify the consistency of all results reported in the main text of the manuscript, in Tables 2, 3, 4 of the manuscript and in the supplemental tables S16 and S17. *Exceptions:* I did not attempt to verify the results in Table 1 of the main text because I am not familiar with the BLOCKO program that was used to compute the reported social-relations-model analyses. I also did not attempt to verify the results in the supplemental tables (except from S16 and S17) due to time constraints, as well as the descriptive statistics of the predictor variables reported in the main text on page 1481.

*Obtained data:* I downloaded the Data/ folder from <http://reshare.ukdataservice.ac.uk/852716/> but did not include the data files in my Github repository. To run the computations without changing the code in the syntax files, I had to place all data files from Data/ into the Syntax/ folder.

*Involved descriptions:* During the reproduction, I consulted the information reported in the manuscript (Joel et al., 2017) and the source code provided by the authors at <http://reshare.ukdataservice.ac.uk/852716/>.

*Inherited source code:* I performed a re-execution reproduction by running the original R scripts from the Syntax/ folder downloaded from <http://reshare.ukdataservice.ac.uk/852716/>.

*Applied methods standards:* I checked for consistency with common methods and reporting standards in the fields of psychology, applied statistics, and machine learning.

*Obstacles:*

* *Underspecification:*

Software versions:

Because the original software versions were not documented by the authors in their public resources, I used current versions of R and R packages and documented all versions in a [renv.lock](https://rstudio.github.io/renv/index.html) file (<https://github.com/FlorianPargent/error-review_Joel_2017/blob/main/renv.lock>). Because of different software versions than the original, my results cannot be expected to be numerically identical to the published results. However, the results of the reproduction can still be checked for consistency ([Kohrt et al., 2024](https://florianpargent.github.io/error-review_Joel_2017/#ref-kohrt_2024)) because the results of the computations should not qualitatively differ with new random numbers or minor changes to the parameter settings of statistical algorithms.

The authors have used the VSURF R package (Genuer et al., 2015) to perform variable selection. In version 1.2.0 (2022-12-14, <https://cran.r-project.org/web/packages/VSURF/NEWS>), the default settings for the number of trees performed during the different variable selection steps were restructured and reduced to speed up computations. In the documentation, the package authors recommend increasing the default settings when computational resources are not limited, which would suggest that running the original source code with the current software versions might produce less optimal results. Because the authors did not document the exact version they used (but we know that it must have been a version with the old defaults), reproducing their variable selection results (by trying different package versions and or adjusting the defaults settings when using the current package version) would have been very time-consuming so *I did not attempt this*.

Hard-coded predictor variables:

Throughout the original source code, the predictor variables used in the final random forest analyses (e.g., those reported in Table 2 and 3) were hard-coded instead of automatically extracted from the result objects of the VSURF variable selection procedure. This fact made my reproduction both more difficult and easier at the same time:

1. The hard-coded predictor variables mean that even if the original software versions could be used, running the original source code would not faithfully re-execute the complete analysis pipeline. The final random forest analyses would always include the predictor variables specified by the authors after performing their specific variable selection run. For this reason, it is nearly impossible to verify (even with the original software versions) that the authors did not make any copy mistakes when hard coding the predictor variables, because we do not know the actual results of their variable selection run.
2. The hard-coded predictor variables mean that although I had problems reproducing the number of selected variables (due to the described problems concerning the lack of documented software versions and the change in default settings in the VSURF package), I could more easily reproduce the results of the final random forest analyses because I knew which predictor variables were actually included in the models (because those are hard-coded in the source code and I did not have to make any changes).

Note that even if the used R packages had not introduced any changes to their methodology since their previous versions, the fact that the seed algorithm for random number generation was changed in R version 3.6.0 (published in 2019), means that no numerical computations that involve random number generation are guaranteed to be exactly reproducable by my results of the final random forest models. I also experimented with setting the random number generator to the old setting (sample.kind = "Rounding", in set.seed()) but could not exactly reproduce the original results.

* *Conflicting choices:* None
* *Other obstacles:* None

*Faults:* I did not find any faults responsible for the original computation not corresponding to its descriptions, data, and result.

*Consistency of result:* In general, all reproduced results seem consistent with the original results. All results of my re-execution reproduction are documented on the website (<https://florianpargent.github.io/error-review_Joel_2017/#results-from-rerunning-the-original-r-scripts>) and the accompanying Github repository (<https://github.com/FlorianPargent/error-review_Joel_2017>).

As outlined in the *Obstacles* section, I could not expect to obtain the same number of predictor variables when re-executing the original source code that performed the variable selection with the VSURF package with different package versions. Indeed, when running the variable selection with the current package versions, the number of variables selected differed from the numbers reported in the manuscript (e.g., Table 2 and 3). The deviation seemed to be greatest for the liberal step and smallest for the stringent step. *I did not look into this in more detail due to time-constraints.*

As outlined in the *Obstacles* section, reproducing the performance estimates of the final random forest analyses (e.g., Table 2 and 3) was independent of the reproduced results of the variable selection because the predictor variables used in the final analyses were hard-coded in the source code (and I did not make any changes). All performance estimates could be reproduced *almost identically*, despite using current package versions (deviations were mostly on the second decimal place).

*Support for claim from result:*

* *Appropriate data collection:* I consider the study design and data collection meaningful and appropriate.
* *Appropriate computation:* I consider the computations described in the descriptions meaningful and appropriate.
* *Claims follow from results:* I consider all major claims in the original article following from the results.

*Evaluation:* After conducting the reproduction, the support for the major claims in the original article are strengthened.  
I would like to make a positive mention of the fact that the authors were careful not to put a strong focus on the content of the individual predictor variables that their variable selection procedure identified as important. My experience when running the variable selection with current package versions would suggest that the exact list of selected variables might not be completely stable, and could change with different settings of the variable selection procedure.

*Communication:* The original authors were not contacted before or during the re-execution reproduction.

*Suggested improvements:* See the section **Model Misspecification** of the ERROR report**.**

*Time effort:* I found it hard to estimate the time I spent on the re-execution reproduction as well as the different categories listed in the ERROR review form. Unfortunately, I did not work on this project in a timely fashion. Several times, the project lay dormant for many weeks or even months. For this reason, I decided against making specific time estimates. I can roughly estimate that I spent at least 25 hours in total.

## References

Genuer, R., Poggi, J., & Tuleau-Malot, C. (2015). VSURF: An R Package for Variable Selection Using Random Forests, *The R Journal*, 7(2), 19-33. [https://journal.r-project.org/articles/RJ-2015-018](https://journal.r-project.org/articles/RJ-2015-018/)

Janitza, S. & Hornung, R. (2018). On the overestimation of random forest’s out-of-bag error. *PLoS ONE,* 13(8), e0201904. <https://doi.org/10.1371/journal.pone.0201904>

Joel, S., Eastwick, P. W., & Finkel, E. J. (2017). Is romantic desire predictable? Machine learning applied to initial romantic attraction. *Psychological Science*, *28*(10), 1478–1489.<https://doi.org/10.1177/0956797617714580>

Kohrt, F., Melinscak, F., McElreath, R., & Schönbrodt, F. D. (2024). *A conceptual framework for computational reproductions: Formal definitions and epistemic functions*.<https://doi.org/10.5281/ZENODO.14282874>

Pargent, F., Schoedel, R., & Stachl, C. (2023). Best Practices in Supervised Machine Learning: A Tutorial for Psychologists. *Advances in Methods and Practices in Psychological Science*, 6(3). [https://doi.org/](https://doi.org/10.1177/0956797617714580)[10.1177/25152459231162559](https://doi.org/10.1177/25152459231162559)

Roberts D. R., Bahn V., Ciuti S., Boyce M. S., Elith J., Guillera-Arroita G., Hauenstein S., Lahoz-Monfort J. J., Schröder B., Thuiller W., Warton D. I., Wintle B. A., Hartig F., Dormann C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8), 913–929. <https://doi.org/10.1111/ecog.02881>

Sisk, R., Sperrin, M., Peek, N., van Smeden, M., & Martin, G.P. (2023). Imputation and missing indicators for handling missing data in the development and deployment of clinical prediction models: A simulation study. *Statistical Methods in Medical Research*, 32(8). 1461-1477. <https://doi.org/10.1177/09622802231165001>