Meta-Psychology, 2023, vol 7, MP.2020.2556
https://doi.org/10.15626/MP.2020.2556
Article type: Original Article

Open data: Not Applicable
Open materials: Yes
Open and reproducible analysis: Yes
Open reviews and editorial process: Yes
Preregistration: No

Edited by: Danielsson, H., Carlsson, R.
Reviewed by: Parsons, S., Young, C., Antonoplis, S.
Analysis reproduced by: Batinović, L., Tobias, C.
All supplementary files can be accessed at OSF:
https://doi.org/10.17605/OSF.IO/8MZV3

Comparing the vibration of effects due to model, data pre-processing, and sampling uncertainty on a large data set in personality psychology

Simon Klau[1,2], Felix D. Schönbrodt[3,4], Chirag J. Patel[5], John P.A. Ioannidis[6,7,8,9], Anne-Laure Boulesteix[1,4], and Sabine Hoffmann[1,4]

[1]Institute for Medical Information Processing, Biometry, and Epidemiology, Munich, Germany
[2]Leibniz-Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany
[3]Department of Psychology, Ludwig-Maximilians-Universität München, Munich, Germany
[4]LMU Open Science Center, Ludwig-Maximilians-Universität München, Munich, Germany
[5]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[6]Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA, USA
[7]Department of Epidemiology and Population Health, Stanford University, Stanford, CA, USA
[8]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA
[9]Department of Statistics, Stanford University, Stanford, CA, USA

Researchers have great flexibility in the analysis of observational data. If combined with selective reporting and pressure to publish, this flexibility can have devastating consequences on the validity of research findings. We extend the recently proposed vibration of effects approach to provide a framework comparing three main sources of uncertainty which lead to instability in empirical findings, namely data pre-processing, model, and sampling uncertainty. We analyze the behavior of these sources for varying sample sizes for two associations in personality psychology. Through the joint investigation of model and data pre-processing vibration, we can compare the relative impact of these two types of uncertainty and identify the most influential analytical choices. While all types of vibration show a decrease for increasing sample sizes, data pre-processing and model vibration remain non-negligible, even for a sample of over 80000 participants. The increasing availability of large data sets that are not initially recorded for research purposes can make data pre-processing and model choices very influential. We therefore recommend the framework as a tool for transparent reporting of the stability of research findings.

Keywords: metascience, researcher degrees of freedom, stability, replicability, Big Five

In recent years, a series of attempts to replicate results of published research findings on independent data have shown that these replications tend to produce much weaker evidence than the original study (Open Science Collaboration, 2015), leading to what has been referred to as a 'replication crisis'. While there have been a number of widely publicized examples of fraud and scientific misconduct (Ince, 2011; van der Zee et al., 2017), many researchers agree that this is not the major problem causing the crisis (Gelman and Loken, 2014; Ioannidis et al., 2014). Instead, the problems seem to be more subtle and partly due to the multiplicity of possible analysis strategies (Goodman et al., 2016; Open Science Collaboration, 2015). In this vein, there is evidence that the instability of empirical associations can be partly explained by the fact that researchers tend to run several analysis strategies on a given data set, but report only one of them selected post-hoc (Simmons et al., 2011).

Indeed, there are a great number of implicit and explicit choices that have to be made when analyzing observational data. It is necessary to make various decisions when specifying a probability model to study the association between possible predictor variables and an outcome of interest (Leamer, 1983). In addition to possible choices involved in the specification of a probability model, denoted as 'model uncertainty' in the following, there are numerous judgments and decisions that are required prior to fitting the model to the data. When pre-processing the data, there are many possibilities regarding not only the definition of predictor and outcome variables, but also data inclusion and exclusion criteria, and the treatment of outliers (Wicherts et al., 2016). We denote this type of uncertainty as 'data

pre-processing uncertainty'.

Apart from the problems arising through the multiplicity of possible analysis strategies, there seem to be more fundamental issues in the analysis of observational data that originate from the low statistical power which characterizes many psychological studies (Maxwell, 2004; Szucs and Ioannidis, 2017). In psychology, effect sizes tend to be small and sample sizes are typically small to moderate. This combination leads to studies with low statistical power and therefore high sampling uncertainty when the same analysis strategies are applied to different samples with the aim of answering the same research question. High sampling uncertainty decreases the chances of being able to replicate the results of studies that detect a true effect.

In recent years, a plethora of solutions to the replication crisis have been proposed in different disciplines. There are several approaches that allow the reporting of results for a large number of possible analysis strategies (Muñoz and Young, 2018; Simonsohn et al., 2015; Steegen et al., 2016; Young, 2018), including the vibration of effects which was proposed by Ioannidis (2008) and further developed by Patel, Burford, and Ioannidis (2015), Palpacuer et al. (2019), and Klau, Hoffmann, Patel, Ioannidis, and Boulesteix (2021). Alternatively, the flexibility in the choice of analysis strategies can be reduced before analyzing the data through preregistration and registered reports (Chambers, 2013; Wagenmakers et al., 2012). Similarly, the instability of empirical findings arising from sampling uncertainty can be assessed through resampling (Meinshausen and Bühlmann, 2010; Sauerbrei et al., 2011) or sampling uncertainty can be reduced by increasing the sample size (Button et al., 2013; Maxwell, 2004; Schönbrodt and Perugini, 2013). While the solutions proposed so far address important pieces of the problem by either focusing on the multiplicity of analysis strategies or on sampling uncertainty, it is important to be able to investigate sampling, model, and data pre-processing uncertainty in a common framework to understand the full picture. Klau, Martin-Magniette, Boulesteix, and Hoffmann (2020) rely on a resampling procedure to compare method and sampling uncertainty, but focus their application on the selection and ranking of molecular biomarkers.

In this work, we use the vibration of effects approach (Ioannidis, 2008) to assess model, data pre-processing, and sampling uncertainty in order to provide a tool for applied researchers to quantify and compare the instability of research findings arising from all three sources of uncertainty. We study this instability for varying sample sizes for two associations in personality psychology, namely between neuroticism and relationship status, and extraversion and physical activity, by analyzing a large and publicly available data set.

The vibration of effects framework

The vibration of effects framework to quantify the effect of model, data pre-processing, and sampling uncertainty

The vibration of effects framework (Ioannidis, 2008; Patel et al., 2015) provides researchers with a tool to assess the robustness of their research findings in terms of alternative analysis strategies. In particular, it allows the quantification of the impact of different choices on the stability of results, helping researchers identify the most influential analysis choices. In this respect, the vibration of effects framework has some conceptual similarities to specification curve analysis (Simonsohn et al., 2015), multiverse analysis (Steegen et al., 2016) and multi-analyst experiments (Aczel et al., 2021). All these approaches try to assess whether results vary according to different specifications. In contrast to the latter approaches, the vibration of effects framework presents the effect estimates and p-values resulting from a large number of analysis strategies simultaneously in a volcano plot. Moreover, vibration of effects considers a more comprehensive set of possible strategies, while the specification curve and multiverse analysis include a step where the researchers try to define what are reasonable specifications (a task that is often difficult) and multi-analyst experiments typically request many different analysts to make independently their best choice of the analysis strategy. While the vibration of effects approach was initially proposed and applied to assess model uncertainty, it has recently been extended to enable comparison of the relative impact of different analysis choices with measurement and sampling uncertainty (Klau et al., 2021).

The vibration of effects framework can be used in the context of modeling an association of interest, i.e., when estimating the effect of a predictor of interest on an outcome of interest to obtain effect estimates and corresponding p-values, while controlling for the effect of several covariates.

In the application of the framework by Patel et al. (2015), the authors consider the association between a predictor of interest and a survival outcome, and assess the vibration by defining a large number of models which result from the inclusion or exclusion of a number of potential covariates. In this work, we will refer to the type of vibration investigated by Patel et al. (2015) as 'model vibration', apply the framework to subsamples of the data (Klau et al., 2021), and extend it to data pre-processing choices in order to compare

model vibration to 'sampling vibration' and 'data pre-processing vibration'. To quantify sampling vibration, we use a resampling-based approach where we draw a large number of random subsets from our data set and fit the same model on each of these subsets. Furthermore, we fit a model for a large number of data pre-processing strategies in order to assess data pre-processing vibration. These choices could, for example, include the handling of outliers, eligibility criteria, or the definition of predictor and outcome variables. Examples for the implementation of these three types of vibration are provided in section *Applying the vibration of effects framework to the SAPA dataset*.

Figure 1 shows three possible patterns of vibration of effects generated with fictive data. Since our application of the vibration of effects will focus on binary outcomes in the context of logistic regression models, we present these figures with odds ratios (OR) as effect estimates. In the left panel, a regular pattern is visualized where all effects are positive (OR > 1) and significant ($p < 0.05$). This is recognizable by a vertical line, where OR = 1, and a horizontal line, where $p = 0.05$, illustrating the significance threshold, respectively. Furthermore, dotted lines provide information about the 1st, 50th and 99th percentile of results. Such a regular pattern demonstrates the robustness of the estimated effect to alternative model specifications, data pre-processing options or to resampling, depending on the type of vibration that is presented. The second panel demonstrates a pattern which is characterized by significant and non-significant results in both positive and negative directions – here, the median OR is close to one. We refer to this pattern as the 'Janus effect', in allusion to the two-headed ancient Roman god (Patel et al., 2015). While a Janus effect pattern indicates that there is no consistent association between the predictor of interest and the outcome, the occurrence of both positive and negative significant results can lead to researchers selectively reporting a significant finding in the desired direction if they try a number of possible analysis strategies. Finally, the right panel contains a more irregular pattern. Such a pattern can, for example, result from the inclusion or exclusion of a particular covariate, or by different choices in the definition of a covariate. By highlighting the data points referring to such a definition, the results can be visually connected to these choices.

To quantify the variability in these results, Patel et al. (2015) propose two summary measures, namely relative hazard ratios and relative p-values (RP). These summary measures are defined as the ratio of the 99th and 1st percentile of hazard ratios and the difference between the 99th and 1st percentile of -log10(p-value), respectively. Following Patel et al. (2015), we define the relative odds ratio (ROR) as the ratio of the 99th percentile and 1st percentile of the OR. The ROR provides a more robust and intuitive measure of variability than the variance. The minimal possible value of ROR is 1, indicating no vibration of effects at all, while larger ROR values indicate larger vibration.

Comparing the vibration of effects due to different types of uncertainty and identifying the most influential analytical choices

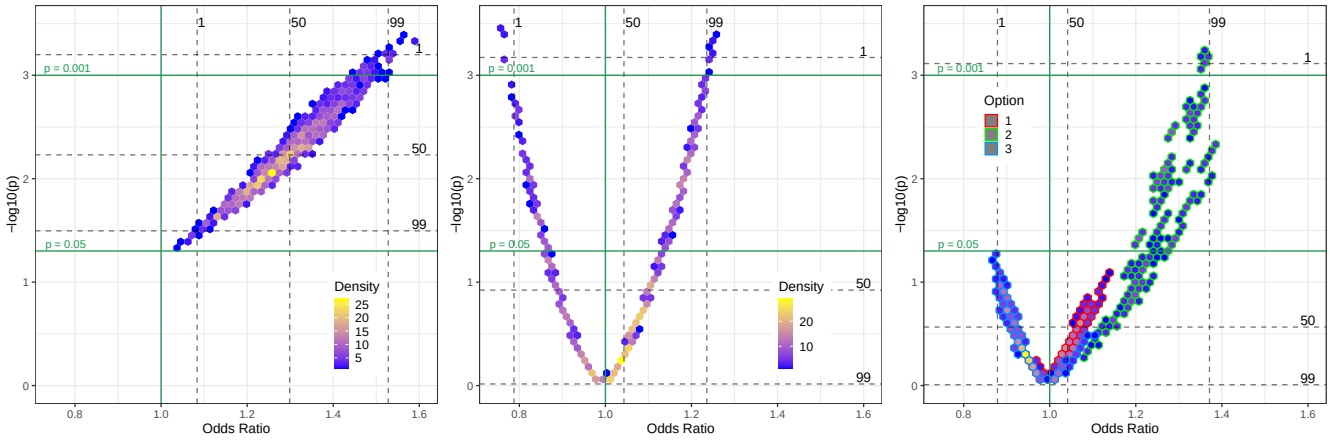
For an association of interest, model, data pre-processing, and sampling uncertainty can be compared through the vibration of effects framework. In order to assess the variability in effect estimates and p-values for one type of vibration, the other types of vibration have to be fixed to a 'favorite' specification. For instance, when focusing on sampling vibration only, decisions on a favorite model as well as a favorite data pre-processing choice must be made.

In addition to the investigation of individual types of vibration, the joint impact of model and data pre-processing choices on the variability of results can be quantified. For simplicity, we will refer to the combination of a model and all necessary data pre-processing choices as an analysis strategy. In the joint investigation of model and data pre-processing choices, the calculation of ROR is straightforward and can give an estimate for the total amount of vibration caused by the analysis strategy. Additionally, the relative impact of data pre-processing and model choices on the vibration that is caused by the choice of the analysis strategy can be quantified and it is possible to identify the model and data pre-processing choices that explain the largest variation in results. To do so, we can use a linear model in which we describe the association between the effect estimate of interest as an outcome variable (in our case the log(OR)) as a function of two categorical covariates, indicating data pre-processing and model choices. By performing a variance decomposition through an analysis of variance (ANOVA), we can determine the data pre-processing choices and model choices that most contribute to the total amount of vibration caused by the analysis strategy.

In the following section, we will give detailed examples of the application of the vibration of effects framework regarding model, sampling and data pre-processing choices.

Figure 1

Vibration of effects with fictive data.



Applying the vibration of effects framework to the SAPA dataset

The data and research questions of interest

For the application of the vibration of effects, we use a large data set from the SAPA project personality test (Condon et al., 2017) which is publicly available at the Dataverse repository (https://dataverse.harvard.edu/dataverse/SAPA-Project). The sample consists of 126884 participants who were invited to complete an online survey between 2013 and 2017 in order to evaluate the structure of personality traits. The data set comprises information about a large pool of 696 personality items which were completed by the participants on a 6-point scale ranging from 1 (*very inaccurate*) to 6 (*very accurate*) and a set of additional variables including gender, age, country, job status, educational attainment level, physical activity, smoking status, relationship status and body mass index (BMI) of participants.

In this work, we use these data to assess the extent to which associations between the Big Five (agreeableness, conscientiousness, extraversion, neuroticism and openness to experience) and the five outcome variables physical activity, educational achievement, relationship status, smoking habits and obesity are influenced by data pre-processing, model, and sampling uncertainty. In order to investigate the behavior of the three types of vibration with increasing sample size, we consider different subsets of the original data with subset sizes $n \in \{500, 5000, 15000, 50000, 84045\}$, where 84045 is the size of the complete data set after excluding participants with missing observations. Lower sample sizes than the original sample size were obtained by generating random subsamples from the original data set, without replacement. In the application of our framework, we consider six associations of interest, comprising five for which we found empirical evidence in the psychological literature. In the presentation of our results, we focus on the association between neuroticism and relationship status and between extraversion and physical activity (Rhodes and Smith, 2006).

There is a large body of evidence on the association between neuroticism and relationship satisfaction (Dyrenforth et al., 2010; Malouff et al., 2010; O'Meara and South, 2019), which might for instance be explained by cognitive biases in the interpretation of ambiguous situations (Finn et al., 2013). Concerning the association between extraversion and physical activity, Eysenck, Nias, and Cox (1982) suggested that individuals with high levels of extraversion would be more likely to start sports and to excel in them because the bodily activity would satisfy their sensation seeking behavior. According to Wilson and Dishman (2015), an association between extraversion and physical activity may also result from the fact that extraverts are more social and outgoing, making them more exposed to situations that offer the possibility to be physically active. Additional results on the association between agreeableness and smoking (Malouff et al., 2006), neuroticism and obesity (Gerlach et al., 2015), and conscientiousness and education (Sorić et al., 2017) can be found in the Supplementary Material, together with results on openness and physical activity, for which no evidence of an association could be found (Rhodes and Smith, 2006).

Quantifying and comparing the effect of model, sampling and data pre-processing uncertainty

We describe each association of interest through a logistic regression model in which we estimate the effect of the predictor of interest (e.g., neuroticism or extraversion) on the binary outcome of interest (e.g., relationship status or physical activity) to obtain odds ratios (OR) and corresponding p-values, while controlling for the effect of several covariates. As potential control variables, we consider all variables introduced in section *The data and research questions of interest* that are not part of the association of interest. For instance, the association between neuroticism and relationship status comprises the control variables age, gender, continent, job status, BMI, smoking, education, physical activity, conscientiousness, agreeableness, extraversion and openness. For the association between physical activity and extraversion, we replace these two variables in the list of potential control variables with neuroticism and relationship status. This results in a total number of 12 control variables for each associations of interest.

We quantify the instability of these associations through the vibration of effects framework introduced in section *The vibration of effects framework to quantify the effect of model, data pre-processing, and sampling uncertainty*.

Model vibration

In order to assess model vibration, we consider all possible combinations of control variables as described in the introduction of the framework. Following Patel et al. (2015), we will consider age and gender as baseline variables which are included in every model, resulting in a total number of $2^{10} = 1024$ possible models for a given association of interest.

Sampling vibration

To quantify sampling vibration, we follow the strategy of drawing a large number of random subsets from our data set and fitting the same logistic regression model on each of the subsets, as outlined in the introduction of the framework. In particular, we draw 1000 subsets of size $0.5n$, with n as the number of observations from the data sets defined in section *The data and research questions of interest*, which comprise different numbers of observations themselves. Although each subset is drawn without replacement, the observations of subsets overlap between repetitions.

Data pre-processing vibration

The data pre-processing choices we are considering comprise the handling of outliers, eligibility criteria, and the definition of predictor and outcome variables. These data pre-processing choices are based on studies found in the literature. For a given association of interest, we fit a logistic regression model for each data pre-processing strategy.

Eligibility criteria. The eligibility criteria are based on the variables age, gender and the country of participants. For age, either the full group of participants is included in the analyses (age eligibility criterion definition 1) or a subgroup is defined by excluding participants who are younger than 18 (age eligibility criterion definition 2), which can be justified by their inability to legally provide consent (Barchard and Williams, 2008). Furthermore, studies about associations involving the Big Five personality traits are often carried out on subgroups of countries, for instance as shown by Malouff et al. (2006) and Malouff et al. (2010) for the variables smoking and physical activity. Therefore, we distinguish two alternative study populations based on the participants' country. Either all participants are included in the analyses and continent is considered as a categorical control variable (country eligibility criterion definition 1), or we include only participants from the United States, which presents the single largest country in the data set. In this case (country eligibility criterion definition 2), we exclude the control variable specifying the continent from the analyses. In total, this results in $2 \times 2 = 4$ possible combinations for the definition of eligibility criteria.

Handling of outliers. A further data pre-processing choice is the handling of outliers. A variety of different outlier definitions can be found in the literature. Bakker and Wicherts (2014), for instance, provide a large range of z-values (which is the number of standard deviations that a value deviates from the mean) that are used to define outliers. Furthermore, it is either possible to remove or winsorize outlier values (Osborne and Overbay, 2004). Here, we focus on three different choices concerning all continuous covariates, comprising the five personality dimensions, as well as age and BMI: Firstly, we perform no further pre-processing with these covariates (outlier definition 1). As a second option, we delete observations with absolute z-values greater than 2.5 (outlier definition 2). Finally, we perform winsorization to achieve absolute z-values less than or equal 2.5 (outlier definition 3). Thereby we replace values with $z > 2.5$ by 2.5, and values with $z < -2.5$ by -2.5.

Dichotomization of outcome variables and covariates. In the definition of the outcome variables and covariates, we only consider the influence of different pre-processing choices for the three variables smoking (which is the outcome variable in the association between agreeableness and smoking, see results in the Supplementary Material, and a covariate in all other as-

sociations), physical activity (which is the outcome variable in the association between extraversion and physical activity and between openness and physical activity, see results in the Supplementary Material, and a covariate in all other associations) and education (which is the outcome variable in the association between conscientiousness and education, see results in the Supplementary Material, and a covariate in all other associations). All three variables are recorded with a certain number of categories (nine categories for smoking, six categories for physical activity and seven categories for education) and have to be dichotomized in order to be able to model them as a binary outcome in a logistic regression model. For all three variables, literature search revealed a lack of common definitions. For smoking and physical activity for instance, summaries of these definitions are provided by Malouff et al. (2006) and Rhodes and Smith (2006), respectively. Similarly, the term education is very ambiguous, and even the more specific phrase of academic achievement exhibits a large variety of definitions (Fan and Chen, 2001). Therefore, we aim at reasonable dichotomizations of our given categories. For smoking, we either consider a definition based on never smokers vs. all other categories of smoking (smoking definition 1) or based on non-smokers (never smokers and study participants who did not smoke the previous year) versus all other study participants (smoking definition 2). For physical activity, we either assume a definition based on the two categories 'less than once per week' versus 'once per week or more' (physical activity definition 1) or, alternatively, 'less than once per month' versus 'less than once per week or more' (physical activity definition 2). Finally, in the definition of education we distinguish between study participants with a high level of education and study participants with a low level of education. In this distinction, we either assign current university students to the group with a high level of education (education definition 1), because they will soon obtain a university degree or to the group with a low level of education (education definition 2), as they have not obtained a degree yet. All other variables (job status, relationship status, BMI) are included in the analyses without considering alternative pre-processing choices. Therefore, we should acknowledge that the vibration of effects due to pre-processing choices can be larger than what is illustrated here. For more details on the variables which were collected in the SAPA project, we refer to Condon et al. (2017).

Personality scores. The definitions of the five personality dimensions, i.e., openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism, are based on the corresponding personality items. There are a large number of different strategies to combine several items to a scale value. Indeed, the SAPA data set contains almost 700 items that were designed to assess personality, but each participant only completed a subset of these items. In order to determine a score on each of the personality dimensions, a correlation matrix, which is based on pairwise complete cases can be analyzed through factor analysis. As the Big Five personality traits were initially constructed as orthogonal factors (Saucier, 2002), we consider orthogonal rotation techniques as a first option (factor rotation definition 1) for the factor analysis. However, Saucier (2002) argues that the scales used to measure the Big Five are not orthogonal in practice. In fact, a more common option in factor analysis of the personality traits is the use of oblique rotation techniques (factor rotation definition 2). The assignment of items to the five personality dimensions can be realized by determining a minimal factor loading that has to be achieved to assign an item to a factor but there is no consensus in the literature on an optimal cut-off value for such a minimal factor loading. Here, we either choose a minimal factor loading of 0.3 (factor loading definition 1) or of 0.4 (factor loading definition 2). The score of a participant can then be calculated by taking the mean score of all items that were assigned to a given factor. This strategy might lead to missing values for some participants on the personality dimensions as it is only reasonable to calculate such a score if there is a minimum number of completed items. Here, we use a required minimum value of 5 completed items.

While there are numerous analysis strategies to determine the personality score of a participant, it is not in the scope of this study to consider all possible analysis strategies. Therefore, we limit the number of possible data pre-processing strategies by only considering the two choices, orthogonal vs. oblique rotation, and mean scores on items assigned to a factor with loadings greater than 0.3 or 0.4. While these variable definitions are based on the raw data set with all observations, the other data pre-processing choices are subsequently implemented on the data sets of different sizes.

The combination of the definition of personality scores with all other data pre-processing choices results in 384 different data pre-processing strategies in total. These represent only a subset of a larger number of choices that may be made, in theory. However, in practical terms, they represent the main choices that are likely to be considered.

Table 1

Data pre-processing choices

	Original categories	Definition 1 (favorite)	Definition 2	Definition 3
Eligibility criteria				
Age		All participants	Only ≥ 18	
Country		All participants	Only from US	
Handling of Outliers		No pre-processing	Exclusion if $\|z\| > 2.5$	Winsorization if $\|z\| > 2.5$
Dichotomization of outcome and covariates				
Smoking	Never smokers	'Never smokers'	Non-smokers ('never smokers'	
	Not in the last year	vs.	and 'not in the last year')	
	Less than once a month	all other participants	vs.	
	Less than once a week		all other participants	
	1 to 3 days a week			
	Most days			
	Everyday (5 or less times)			
	Up to 20 times a day			
	More than 20 times a day			
Physical activity	Very rarely or never	Less than once a week	Less than once a month	
	Less than once a month	vs.	vs.	
	Less than once a week	once a week or more	'less than once a week' or more	
	1 or 2 times a week			
	3 or 5 times a week			
	More than 5 times a week			
Education	Less than 12 years	High (incl. 'currently in	High	
	High school graduate	college/university')	vs.	
	Currently in college/university	vs.	low (incl. 'currently in	
	Some college/university, but did not graduate	low	college/university')	
	College/university degree			
	Currently in graduate or professional school			
	Graduate or professional school degree			
Personality scores				
Rotation technique		Oblique	Orthogonal	
Minimal factor loading		0.3	0.4	

Comparing the vibration of effects due to different types of uncertainty

For each association of interest, we quantify and compare model, data pre-processing, and sampling uncertainty through the vibration of effects framework for varying sample sizes. Our favorite data pre-processing choice is pre-processing without any subgroup analysis, without special handling of outliers, and with variable definition 1 for education, smoking and physical activity. Additionally, the favorite definition of the personality traits is performed with the oblique rotation technique and factor loadings greater than 0.3. Our favorite model choice simply consists in the model that contains all potential control variables. Furthermore, if the aim is to assess data pre-processing vibration or model vibration, we define the full data set as our favorite sample.

In addition to the investigation of individual types of vibration, we will compare the joint impact of model and data pre-processing choices on the variability of results with sampling vibration. Here, the combination of data pre-processing and model choices results in $1024 \times 384 = 393216$ analysis strategies. However, not every possible combination yields useful and valid results. For instance, when we consider the data pre-processing choice where the association of interest is only explored for participants from the US, the model including continent as a control variable is not valid. Thus, the total amount of feasible analysis strategies falls to 294912. Moreover, we quantify the relative impact of data pre-processing and model choices on the vibration that is caused by the choice of the analysis strategy as previously described.

Results

The variability in effect estimates for one type of vibration

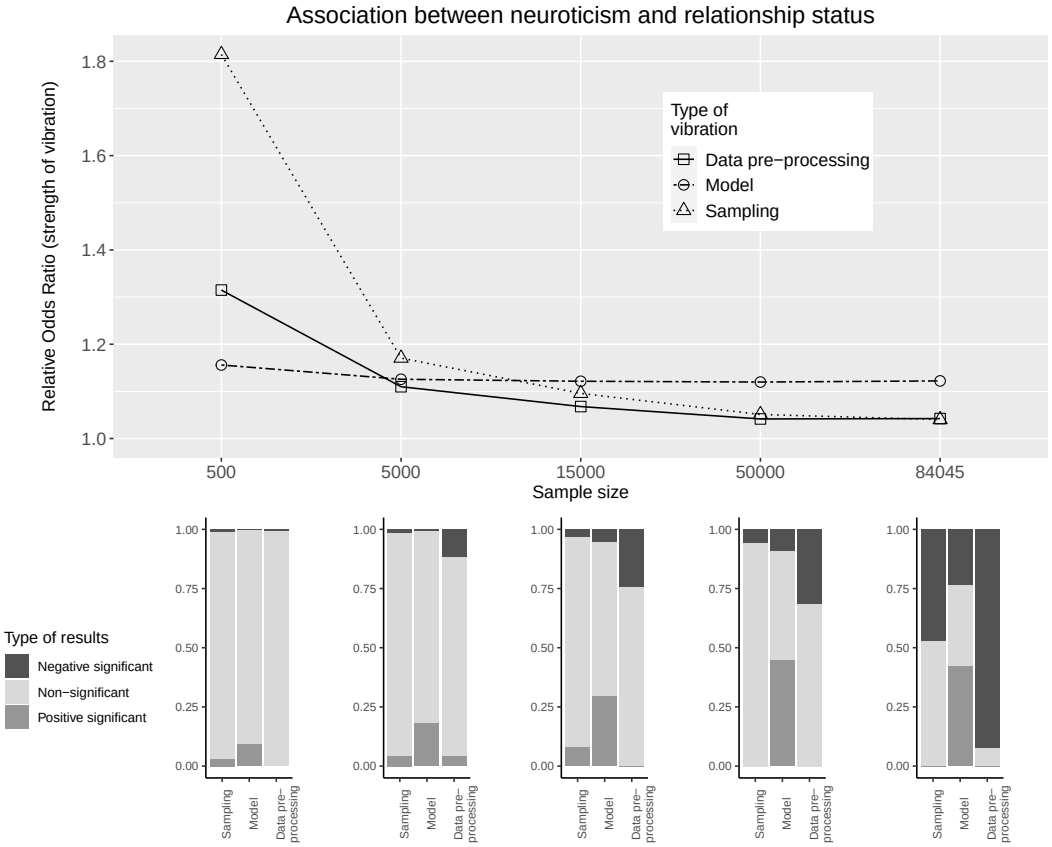
For more stable results, we repeat the analyses of all types of vibration for sample sizes of 500, 5000 and 15000 ten times and average the results across the obtained RORs. For the visualization of vibration patterns, however, we choose one representative plot out of the total number of ten. For a sample size of 50000, we consider the variability between RORs as negligible and run the analyses on only one sampled data set.

For the association between neuroticism and relationship status and the association between extraversion and physical activity, results of measures quantifying the variability in effect estimates for one type of vibration are visualized in Figures 2 and 3, respectively. Corresponding figures for the other associations are provided in the Supplementary Material.

In the upper panels, RORs are displayed against the sample size n for the three types of vibration (data pre-

Figure 2

Data pre-processing, model, and sampling vibration for different sample sizes (top panel), and bar plots visualizing the type of results in terms of significance of estimated effects (bottom panel) for the association between neuroticism and relationship status.



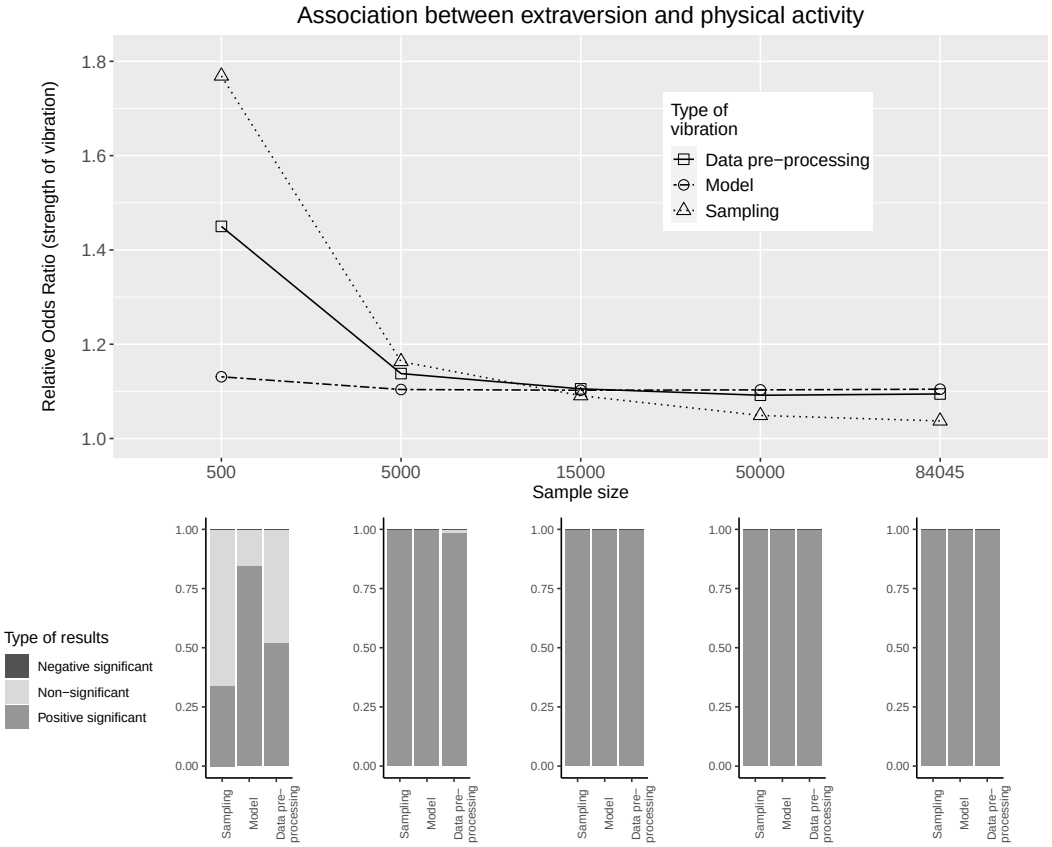
processing, model, and sampling). For both associations, sampling vibration is higher than model and data pre-processing vibration for low sample sizes ($n = 500$ and $n = 5000$). For the lowest sample size of $n = 500$, the ROR quantifying sampling vibration is close to 1.8 (1.81 for the association between relationship status and neuroticism and 1.77 for the association between physical activity and extraversion). For larger sample sizes, sampling vibration decreases and tends to an ROR of 1. Therefore, the influence of a specific sample can be expected to be negligible for sufficiently large sample sizes. Focusing on the two other types of vibration, data pre-processing vibration is larger for low sample sizes than model vibration, and decreases for increasing sample size, however, without approximating an ROR of 1. Model vibration, in contrast, is less influenced by the sample size. Although we observe a slight decrease for RORs quantifying model vibration for increasing sample sizes, it is lower than sampling and data pre-processing vibration for small sample sizes and does not tend to a value of 1 for larger sample sizes.

In the lower panels of Figures 2 and 3, bar plots provide information about the percentage of significant results for each sample size and each type of vibration for the three categories: „negative significant“, „non-significant“, and „positive significant“. For all three types of vibration, most results are not significant for a sample size of $n = 500$ while for the larger sample sizes, the results are mostly significant. For the largest sample size, the association between neuroticism and relationship status shows a Janus effect with both negative and positive significant results for model vibration. On the other hand, for sampling and data pre-processing vibration, only negative-significant or non-significant effects can be observed.

For the association between extraversion and physical activity, all types of vibration yield positive significant effects for sample sizes larger than 5000, which is in accordance with the results from the literature (Rhodes and Smith, 2006). Hence, a Janus effect can-

Figure 3

Data pre-processing, model, and sampling vibration for different sample sizes (top panel), and bar plots visualizing the type of results in terms of significance of estimated effects (bottom panel) for the association between extraversion and physical activity.



not be observed for this association.

The volcano plots in Figures 4 and 5 allow investigating the behavior of the three types of vibration in more detail by providing the exact patterns of -log10(p-value) and ORs for the three sample sizes n = 5000, n = 15000 and n = 50000. For the association between neuroticism and relationship status, we can distinguish a clear Janus effect for sampling vibration with both positive and negative results for all three sample sizes. For model vibration, we initially only observe positive results for a sample size of n = 5000, but with increasing sample size, there are also results indicating a negative association between neuroticism and relationship status. In contrast, for data pre-processing vibration, we observe both positive and negative non-significant results for n = 5000 and n = 15000 whereas for a large sample size of n = 50000 the volcano plot clearly indicates a negative association, even though only about one third of the results are significant. In summary, for the association between neuroticism and relationship status, the results and conclusions critically depend on the chosen analysis strategy and there is a high potential for researchers to find contradictory findings on the same data set if they make different analytical choices.

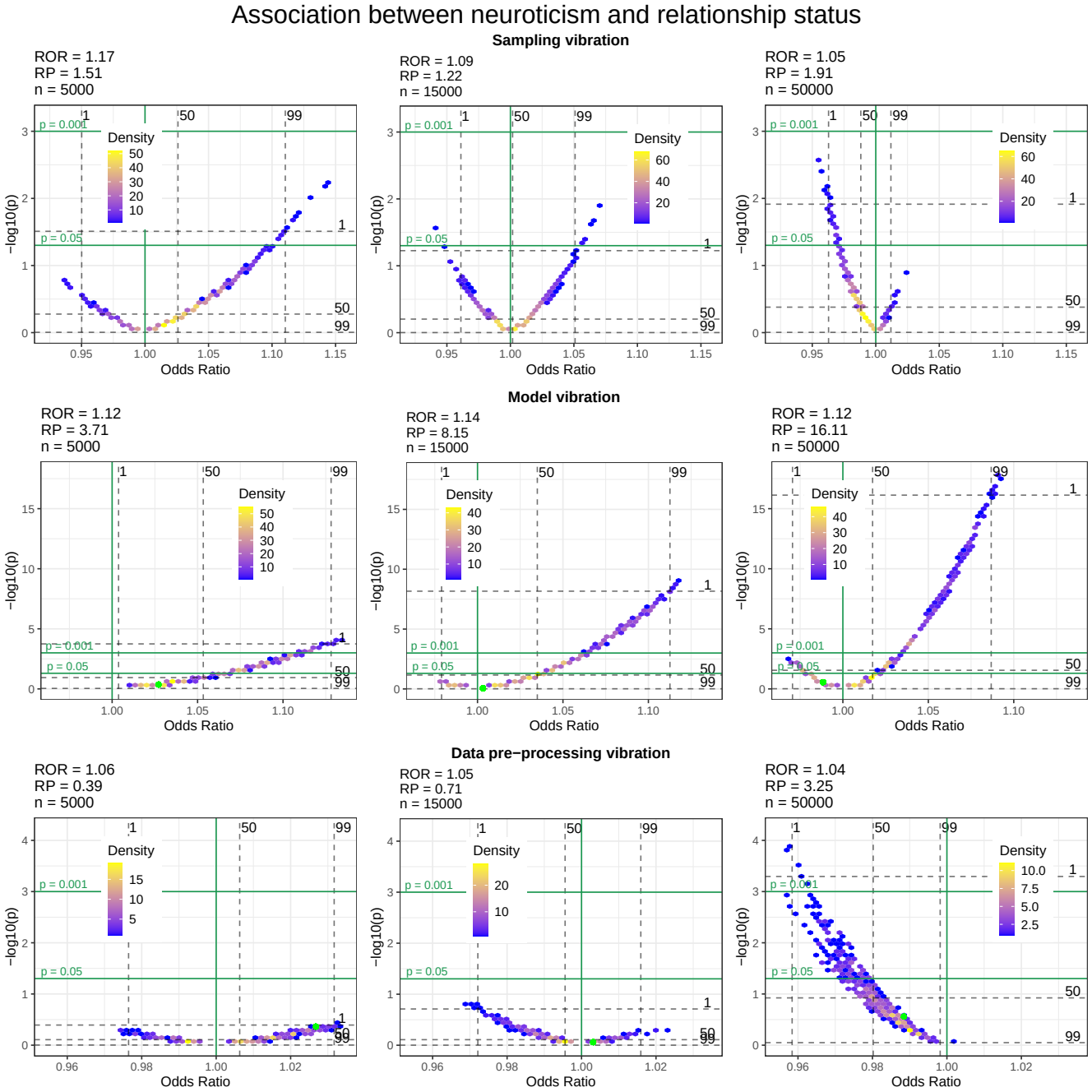
The volcano plots in Figure 5 for the association between extraversion and physical activity show a more regular pattern. We observe only positive associations for all three sample sizes with only positive significant results for n = 15000 and n = 50000, indicating that the results for this association are much more robust to the choice of the analysis strategy.

The relative impact of model and data preprocessing choices and the cumulative impact of both

Results for the total amount of vibration caused by model- and data pre-processing choices are visualized in Figure 6 for the association between neuroticism and relationship status, and Figure 7 for the association between extraversion and physical activity. In these fig-

Figure 4

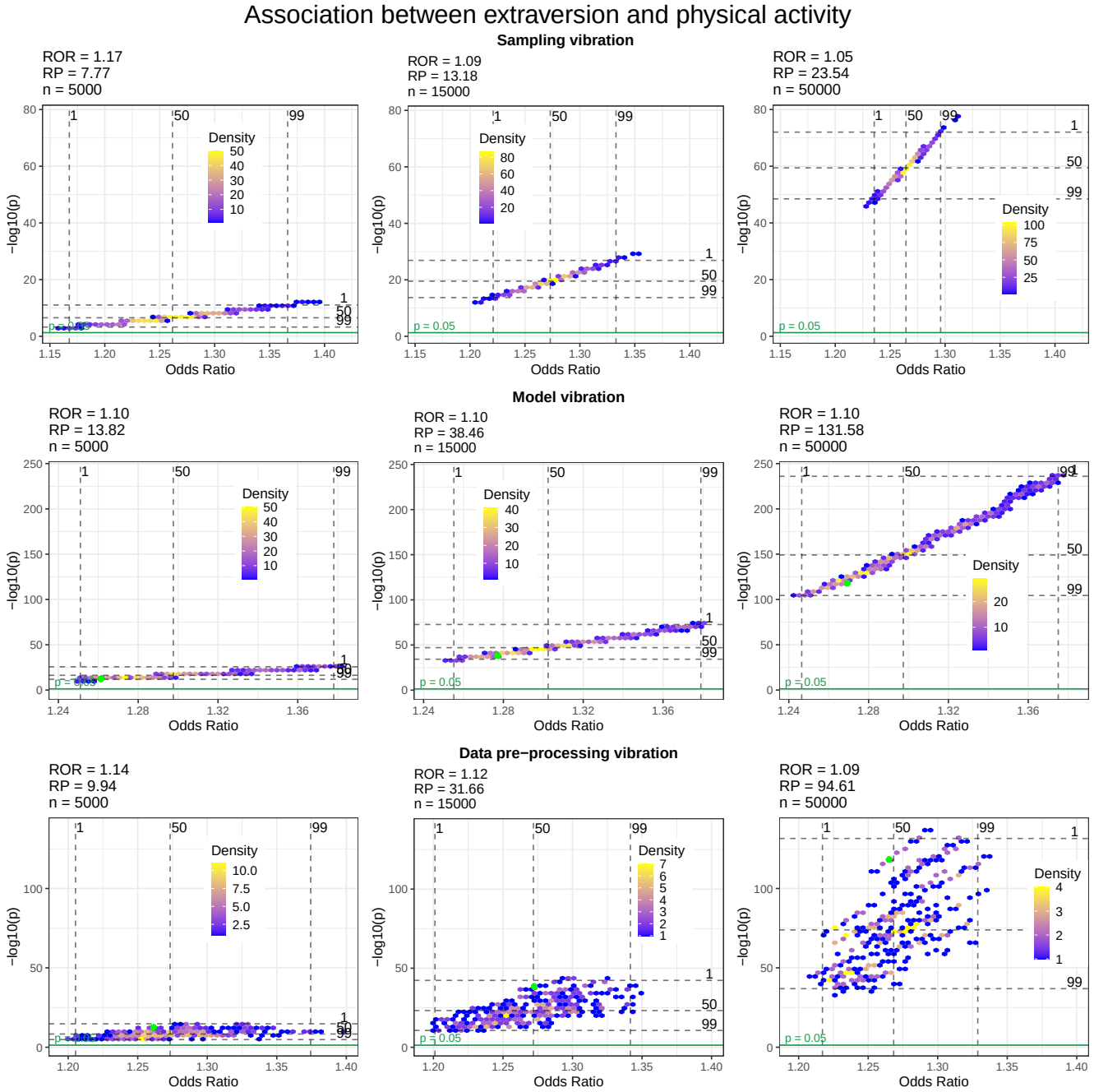
Volcano plots for different types of vibration and different sample sizes (n) for the association between neuroticism and relationship status. The summary measures ROR and RP indicate relative odds ratios and relative p-values, respectively. Green dots indicate results obtained with favorite model choices (middle row) and favorite data pre-processing choices (bottom row).



ures, the top panels allow for a comparison of this joint vibration, also referred to as vibration due to the analysis strategy, and sampling vibration. For a low sample size of $n = 500$, sampling vibration is higher than vibration caused by the analysis strategy for both associations. For a medium sample size of $n = 5000$, RORs

Figure 5

Volcano plots for different types of vibration and different sample sizes (n) for the association between extraversion and physical activity. The summary measures ROR and RP indicate relative odds ratios and relative p-values, respectively. Green dots indicate results obtained with favorite model choices (middle row) and favorite data pre-processing choices (bottom row).



corresponding to these two types of vibration are very similar (e.g. 1.18 (vibration due to the analysis strategy) and 1.17 (sampling vibration) for the association between relationship status and neuroticism, and 1.19 (vibration due to the analysis strategy) and 1.16 (sampling vibration) for physical activity and extraversion).

For larger sample sizes, vibration caused by the analysis strategy is larger than sampling vibration, which, as seen above, tends to an ROR of 1 for the largest sample size. Vibration caused by the analysis strategy, in contrast, does not show obvious decrease for sample sizes larger than 5000 and remains in a range between 1.13 and 1.15 (for the association between relationship status and neuroticism) and between 1.16 and 1.17 (for the association between physical activity and extraversion).

Pie charts in the bottom panels illustrate the relative impact of model and data pre-processing choices on the total vibration caused by the choice of the analysis strategy. Due to the high computational burden of the variance decomposition, we randomly select three of the ten data sets for low sample sizes of 500, 5000 and 15000 to estimate the relative impact of data pre-processing and model choices and average the results over the three selected data sets.

For both associations, the relative impact of data pre-processing choices exceeds the impact of model vibration for a sample size of $n = 500$. For sample sizes larger than 500, however, the relative model impact is larger than the relative impact due to data pre-processing. This is particularly pronounced in the association between relationship status and neuroticism, where between 79.1% ($n = 5000$) and 89.5% ($n = 50000$) of the total vibration can be explained by model choices. For the association between physical activity and extraversion, between 53.0% ($n = 5000$) and 61.7% ($n = 50000$) of the total vibration can be explained by model choices. The relative impact of data pre-processing choices is quantified by values between 35.9% ($n = 50000$) and 55.0% ($n = 500$) for this association.

A more detailed investigation of data pre-processing vibration as part of the total vibration shows that the variable age has the largest impact of the data pre-processing choices on the vibration of effects for this association between physical activity and extraversion (17.5% of the total vibration for the largest sample size). However, associations in the Supplementary Material reveal that the relative impact of data pre-processing and model choices, and the variables with the largest impact, strongly depend on the research question of interest. For the association between education and conscientiousness, for example, for the largest sample size, 97.8% of the vibration caused by the analysis strategy can be explained by data pre-processing choices with education itself as the variable with highest impact (96.1% of the total vibration explained by education).

Discussion

Summary

Researchers have great flexibility in the analysis of observational data. If this flexibility is combined with selective reporting and pressure to publish significant results, it can have devastating consequences on the replicability of research findings. In this work, we extended the vibration of effects approach, proposed by Ioannidis (2008), to quantify and compare the impact of model and data pre-processing choices on the stability of empirical associations. Through this extension, the vibration of effects framework allows identification of the choices in the analysis strategy that explain the most variation in results and comparisons of the impact of different choices with sampling uncertainty.

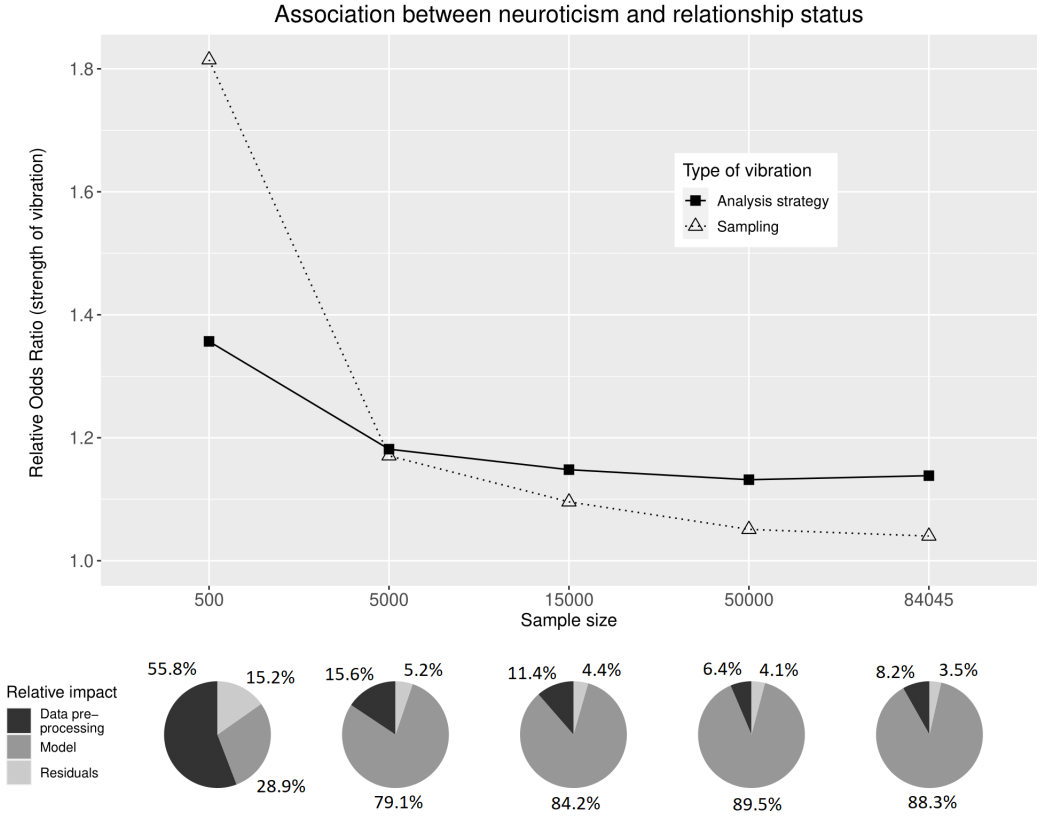
We illustrated three different types of vibration on the SAPA data set, considering reasonable data pre-processing choices and modeling strategies based on a logistic regression model, focusing on two associations of interest in personality psychology. We quantified sampling vibration by considering the results obtained from random subsets of the data set in use. We found that sampling vibration decreased with increasing sample size and became negligible, while model and data pre-processing vibration showed an initial decrease with increasing sample size and then remained constantly non-negligible. Considering all possible combinations of model and data pre-processing choices allowed us to identify the decisions which had the most influence on the variability in results. In addition to the two associations presented in the main text, we show the results of four other associations in the Supplement. These results demonstrate that our findings are not specific to the two examples discussed in our paper, but relevant to a broad variety of associations, including one where no evidence for an association could be found in the literature.

Limitations

When interpreting our results, it is important to keep in mind that both model vibration and data pre-processing vibration are in reality rather elusive concepts as they critically depend on the number and the type of analysis strategies under consideration. In theory, there are an infinite number of models and an infinite number of possible data pre-processing strategies, so any attempt to quantify the variability in an effect estimate resulting from every possible analysis strategy is doomed to fail. As it is futile to quantify the vibration in results arising from every possible strategy, we decided to focus on analysis strategies that seemed reasonable to us, i.e., those that could have been selected

Figure 6

Cumulative model and data pre-processing vibration ('analysis strategy') compared to sampling vibration (top panel), and relative impact of model and data pre-processing vibration for different sample sizes (bottom panel) for the association between neuroticism and relationship status.



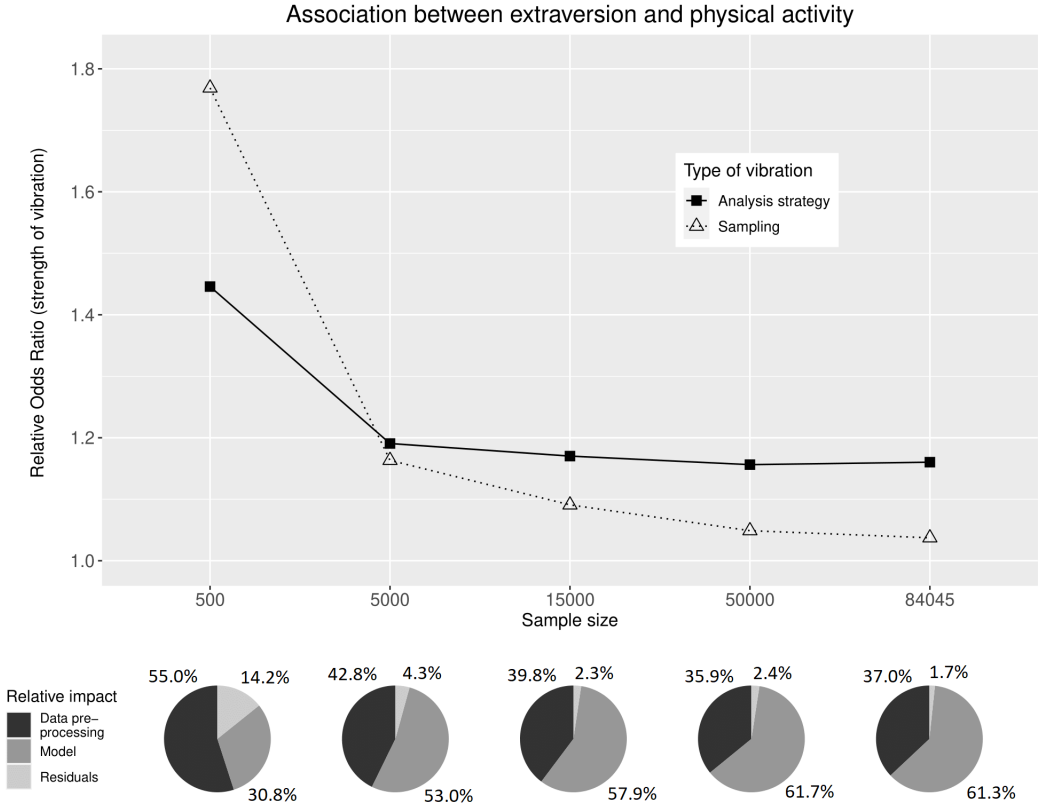
in an actual research project. While there is a firm theoretical basis to predict sampling vibration, the behavior of model and data pre-processing vibration critically depends on the particular data set and the number of possible choices under consideration. As pointed out by Del Giudice and Gangestad (2021), it is not straightforward to identify a set of reasonable analysis strategies and the inclusion of poorly justified analysis strategies in multiverse-style analyses may entail the risk of hiding meaningful effects in a "mass of poorly justified alternatives". The authors also caution against the inclusion of analysis decisions that are not truly arbitrary, because they might, for instance, modify the research question or reduce the reliability of validity with which key variables are measured. Note that the set of considered analysis strategies may sometimes also be critically limited by the available computing capacity as the computing power needed to determine which model and data pre-processing choices lead to the most variation in results also depends critically on the total number of possible analysis strategies.

Following Patel et al. (2015), we merely focused on a special type of model vibration, namely the vibration of effects that is due to the inclusion or exclusion of all potential control variables. Vibration of effects may be larger in situations where very complex models are involved, encompassing a very large number of control variables. Conversely, it may have less of an impact in data-poor studies with few variables measured and considered. Furthermore, we only considered linear effects and neither examined interaction terms nor mediator variables, which may be essential in some settings.

The definition of possible data pre-processing choices is challenging since these choices are sometimes "hidden", i.e., they are typically not discussed in great detail in a publication and some choices are completely omitted. Two recent multi-analyst experiments (Huntington-Klein et al., 2021; Schweinsberg et al., 2021), in which multiple teams of researchers were asked to answer the same research question on the

Figure 7

Cumulative model and data pre-processing vibration ('analysis strategy') compared to sampling vibration (top panel), and relative impact of model and data pre-processing vibration for different sample sizes (bottom panel) for the association between extraversion and physical activity.



same data set, found large variations in data pre-processing options among the different teams, including choices concerning the operationalization of key theoretical variables, and inclusion and exclusion criteria. In Huntington-Klein et al. (2021), no two teams of researchers reported the same sample size when analyzing the same research question on the same data set but "nearly all of the decisions driving data construction would be likely omitted from a paper, or skimmed over by a reader" (Huntington-Klein et al., 2021). Since many data pre-processing choices are not transparently reported in the literature, it is very difficult to determine a set of reasonable data pre-processing steps and multi-analyst experiments seem like the only naturalistic and convincing option to assess the full analytical variability on a given data set, assuming that the multiple analysts are reliable experts. When assessing the vibration of effects for a certain research question, both the set of considered analysis strategies and the selection of "favorite" model and data pre-processing options are to some degree arbitrary but may substantially impact the results. As the main focus of our work was to illustrate how the vibration of effects framework can be used to quantify and compare the impact of different sources of uncertainty and to identify analytical choices that have the most influence on the results, it was not in the scope of our work to quantify analytical variability, for instance by organizing a multi-analyst experiment to identify a set of reasonable data pre-processing choices.

While the vibration of effects framework is an important tool to assess the robustness of empirical findings for model, data pre-processing, measurement, and sampling uncertainty, it is not the only way to address these sources of uncertainty. As pointed out by Hoffmann et al. (2021), a variety of approaches have been proposed across different disciplines to reduce, report, integrate or accept model, data pre-processing, measurement, sampling, method and parameter uncertainty. Efforts to standardize analytical options are underway in some scientific fields building consensus among investigators

and these efforts may result in diminishing the space for potential vibration of effects. Finally, we illustrated the vibration of effects framework in logistic regression models, which is not standard in personality psychology. However, the framework can be adapted with slight modifications to more commonly used methods including, for instance, Gaussian regression and correlation analyses.

Conclusion and Outlook

When analyzing observational data, it is necessary to make model and data pre-processing choices which rely on many implicit and explicit assumptions. The vibration of effects framework provides investigators with a tool to quantify the impact of these choices on the stability of results, helping them focus their attention on the choices that have the greatest influence and are therefore worth further investigation or discussion. Alternatively, other frameworks could be raised and extended for this purpose, such as the specification curve analysis (Simonsohn et al., 2015) or multiverse analysis (Steegen et al., 2016). Compared to these frameworks, the vibration of effects allows presenting a large number of effect estimates and p-values simultaneously. Furthermore, it provides a quantitative intuitive measure of the uncertainty in form of a ratio, and is more appropriate to report sampling uncertainty.

To establish our framework as a tool, we recommend visualizing data pre-processing, model and sampling vibration with volcano plots as we have demonstrated in the Supplementary Material for the association between neuroticism and relationship status. Moreover, the systematic reporting of RORs and p-value characteristics for these types of vibration is a simple but informative guideline for quantifying the stability of published results. The framework can also be useful for readers in the interpretation of these results: When used as a tool to report the robustness of empirical associations, it helps readers (including reviewers) to interpret these results in the context of all the possible results that could have been obtained with alternative, equally justified analysis strategies. When the research data of a publication are made publicly available, which is increasingly common to enhance transparency, a reader can use the vibration of effects framework to assess the extent to which the originally reported results are fragile or incredible because they depend on very specific analytical decisions. In this vein, it is possible to specify a number of model and data pre-processing choices and to apply the framework to assess the variability in effect estimates arising from these possible analysis strategies. In our application of the framework in personality psychology, we observed many cases in which both significant and non-significant results could be obtained, depending on the choice of the analysis strategy. In extreme cases, it was even possible to obtain both positive and negative significant associations and this phenomenon persisted for a very large sample size of over 80000 participants.

The number of decisions which have to be made in the analysis of observational data becomes even more important when analyzing data that are not initially recorded for research purposes. While the increasing availability of large data sets, for instance in the form of Twitter accounts (Barberá et al., 2015) or transaction data (Gladstone et al., 2019), offer unprecedented opportunities to study complex phenomena of interest, they also increase the number of untestable assumptions which must be made in the data pre-processing and choice of model used to describe the data. In light of our results, we suggest using the vibration of effects framework as a tool to assess the robustness of conclusions from observational data.

Author Contact

Correspondence concerning this article should be addressed to Simon Klau, https://orcid.org/0000-0002-7857-1263.

Acknowledgements

We thank Alethea Charlton and Anna Jacob for valuable language corrections and David Condon for providing the data.

Conflict of Interest and Funding

The authors declare that there are no conflicts of interest with respect to the authorship or the publication of this article. This work was funded by the Deutsche Forschungsgemeinschaft (individual grant BO3139/4-3 and BO3139/7-1).

Author Contributions

S. Klau, S. Hoffmann and A.-L. Boulesteix developed the study concept. S. Hoffmann and S. Klau conducted the study and wrote the manuscript. S. Klau performed the statistical analysis. C. J. Patel, J. P. A. Ioannidis, F. D. Schönbrodt and A.-L. Boulesteix substantially contributed to the manuscript. All authors approved the final version.

Open Science Practices

This article earned the Open Materials badge for making the materials openly available. It has been verified that the analysis reproduced the results presented in the article, with minor issues due to complexity of analyses and computational time requirements. The entire editorial process, including the open reviews, are published in the online supplement.

References

Aczel, B., Szaszi, B., Nilsonne, G., van den Akker, O. R., Albers, C. J., van Assen, M. A. L. M., Bastiaansen, J. A., Benjamin, D. J., Boehm, U., Botvinik-Nezer, R., & Wagenmakers, E.-J. (2021). *Consensus-based guidance for conducting and reporting multi-analyst studies* [MetaArXiv]. https://doi.org/10.31222/osf.io/5ecnh

Bakker, M., & Wicherts, J. M. (2014). Outlier removal, sum scores, and the inflation of the type I error rate in independent samples t tests: The power of alternatives and recommendations. *Psychological Methods*, *19*(3), 409–427. https://doi.org/10.1037/met0000014

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, *26*(10), 1531–1542. https://doi.org/10.1177/0956797615594620

Barchard, K. A., & Williams, J. (2008). Practical advice for conducting ethical online experiments and questionnaires for United States psychologists. *Behavior Research Methods*, *40*(4), 1111–1128. https://doi.org/10.3758/BRM.40.4.1111

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Chambers, C. D. (2013). Registered reports: A new publishing initiative at Cortex. *Cortex*, *49*(3), 609–610. https://doi.org/10.1016/j.cortex.2012.12.016

Condon, D., Roney, E., & Revelle, W. (2017). A SAPA project update: On the structure of phrased self-report personality items. *Journal of Open Psychology Data*, *5*(1), 3. https://doi.org/10.5334/jopd.32

Del Giudice, M., & Gangestad, S. W. (2021). A traveler's guide to the Multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. *Advances in Methods and Practices in Psychological Science*, *4*(1), 1–15. https://doi.org/10.1177/2515245920954925

Dyrenforth, P. S., Kashy, D. A., Donnellan, M. B., & Lucas, R. E. (2010). Predicting relationship and life satisfaction from personality in nationally representative samples from three countries: The relative importance of actor, partner, and similarity effects. *Journal of Personality and Social Psychology*, *99*(4), 690–702. https://doi.org/https://doi.org/10.1037/a0020385

Eysenck, H. J., Nias, D. K. B., & Cox, D. N. (1982). Sport and personality. *Advances in Behaviour Research and Therapy*, *4*(1), 1–56. https://doi.org/10.1016/0146-6402(82)90004-2

Fan, X., & Chen, M. (2001). Parental involvement and students' academic achievement: A meta-analysis. *Educational Psychology Review*, *13*(1), 1–22. https://doi.org/10.1023/A:1009048817385

Finn, C., Mitte, K., & Neyer, F. J. (2013). The relationship–specific interpretation bias mediates the link between neuroticism and satisfaction in couples. *European Journal of Personality*, *27*(2), 200–212. https://doi.org/10.1002/per.1862

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*(6), 460–465.

Gerlach, G., Herpertz, S., & Loeber, S. (2015). Personality traits and obesity: A systematic review. *Obesity Reviews*, *16*(1), 32–63. https://doi.org/10.1111/obr.12235

Gladstone, J. J., Matz, S. C., & Lemaire, A. (2019). Can psychological traits be inferred from spending? Evidence from transaction data. *Psychological Science*, *30*(7), 1087–1096. https://doi.org/10.1177/0956797619849435

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, *8*(341), 341ps12–341ps12. https://doi.org/10.1126/scitranslmed.aaf5027

Hoffmann, S., Schönbrodt, F., Elsas, R., Wilson, R., Strasser, U., & Boulesteix, A.-L. (2021). The multiplicity of analysis strategies jeopardizes replicability: Lessons learned across disciplines. *Royal Society Open Science*, *8*(4), 1–13. https://doi.org/10.1098/rsos.201925

Huntington-Klein, N., Arenas, A., Beam, E., Bertoni, M., Bloem, J. R., Burli, P., Chen, N., Grieco, P., Ekpe, G., Pugatch, T., Saavedra, M., & Stopnitzky, Y. (2021). The influence of hidden researcher decisions in applied microeconomics.

Economic Inquiry, *59*(3), 944–960. https://doi.org/10.1111/ecin.12992

Ince, D. (2011). The duke university scandal – what can be done? *Significance*, *8*(3), 113–115. https://doi.org/10.1111/j.1740-9713.2011.00505.x

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*(5), 640–648. https://doi.org/10.1097/EDE.0b013e31818131e7

Ioannidis, J. P. A., Munafo, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention. *Trends in Cognitive Sciences*, *18*(5), 235–241. https://doi.org/10.1016/j.tics.2014.02.010

Klau, S., Hoffmann, S., Patel, C. J., Ioannidis, J. P. A., & Boulesteix, A.-L. (2021). Examining the robustness of observational associations to model, measurement and sampling uncertainty with the vibration of effects framework. *International Journal of Epidemiology*, *50*(1), 266–278. https://doi.org/10.1093/ije/dyaa164

Klau, S., Martin-Magniette, M.-L., Boulesteix, A.-L., & Hoffmann, S. (2020). Sampling uncertainty versus method uncertainty: A general framework with applications to omics biomarker selection. *Biometrical Journal*, *62*(3), 670–687. https://doi.org/10.1002/bimj.201800309

Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review*, *73*(1), 31–43.

Malouff, J. M., Thorsteinsson, E. B., & Schutte, N. S. (2006). The five-factor model of personality and smoking: A meta-analysis. *Journal of Drug Education*, *36*(1), 47–58. https://doi.org/10.2190/9EP8-17P8-EKG7-66AD

Malouff, J. M., Thorsteinsson, E. B., Schutte, N. S., Bhullar, N., & Rooke, S. E. (2010). The five-factor model of personality and relationship satisfaction of intimate partners: A meta-analysis. *Journal of Research in Personality*, *44*(1), 124–127. https://doi.org/https://doi.org/10.1016/j.jrp.2009.09.004

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*(2), 147–163. https://doi.org/10.1037/1082-989X.9.2.147

Meinshausen, N., & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(4), 417–473. https://doi.org/10.1111/j.1467-9868.2010.00740.x

Muñoz, J., & Young, C. (2018). We ran 9 billion regressions: Eliminating false positives through computational model robustness. *Sociological Methodology*, *48*(1), 1–33. https://doi.org/10.1177/0081175018777988

O'Meara, M. S., & South, S. C. (2019). Big five personality domains and relationship satisfaction: Direct effects and correlated change over time. *Journal of Personality*, *87*(6), 1206–1220. https://doi.org/10.1111/jopy.12468

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research & Evaluation*, *9*(6), 1–8.

Palpacuer, C., Hammas, K., Duprez, R., Laviolle, B., Ioannidis, J. P. A., & Naudet, F. (2019). Vibration of effects from diverse inclusion/exclusion criteria and analytical choices: 9216 different ways to perform an indirect comparison meta-analysis. *BMC Medicine*, *17*(174), 1–13. https://doi.org/10.1186/s12916-019-1409-3

Patel, C. J., Burford, B., & Ioannidis, J. P. A. (2015). Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology*, *68*(9), 1046–1058. https://doi.org/10.1016/j.jclinepi.2015.05.029

Rhodes, R. E., & Smith, N. E. I. (2006). Personality correlates of physical activity: A review and meta-analysis. *British Journal of Sports Medicine*, *40*(12), 958–965. https://doi.org/10.1136/bjsm.2006.028860

Saucier, G. (2002). Orthogonal markers for orthogonal factors: The case of the Big Five. *Journal of Research in Personality*, *36*(1), 1–31. https://doi.org/10.1006/jrpe.2001.2335

Sauerbrei, W., Boulesteix, A.-L., & Binder, H. (2011). Stability investigations of multivariable regression models derived from low-and high-dimensional data. *Journal of Biopharmaceutical Statistics*, *21*(6), 1206–1231. https://doi.org/10.1080/10543406.2011.629890

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*(5), 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

Schweinsberg, M., Feldman, M., Staub, N., van den Akker, O. R., van Aert, R. C. M., Van Assen, M. A. L. M., Liu, Y., Althoff, T., Heer, J., Kale,

A., & Uhlmann, E. L. (2021). Same data, different conclusions: Radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis. *Organizational Behavior and Human Decision Processes*, *165*, 228–249. https://doi.org/10.1016/j.obhdp.2021.02.003

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simonsohn, U., Simmons, J., & Nelson, L. D. (2015). Specification curve: Descriptive and inferential statistics on all reasonable specifications. https://doi.org/10.2139/ssrn.2694998

Sorić, I., Penezić, Z., & Burić, I. (2017). The Big Five personality traits, goal orientations, and academic achievement. *Learning and Individual Differences*, *54*, 126–134. https://doi.org/10.1016/j.lindif.2017.01.024

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, *15*(3), 1–18. https://doi.org/10.1371/journal.pbio.2000797

van der Zee, T., Anaya, J., & Brown, N. J. (2017). Statistical heartburn: An attempt to digest four pizza publications from the cornell food and brand lab. *BMC Nutrition*, *3*(54), 1–15. https://doi.org/10.1186/s40795-017-0167-x

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632–638. https://doi.org/10.1177/1745691612463078

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*(1832), 1–12. https://doi.org/10.3389/fpsyg.2016.01832

Wilson, K. E., & Dishman, R. K. (2015). Personality and physical activity: A systematic review and meta-analysis. *Personality and Individual Differences*, *72*, 230–242. https://doi.org/10.1016/j.paid.2014.08.023

Young, C. (2018). Model uncertainty and the crisis in science. *Socius*, *4*, 1–7. https://doi.org/10.1177/2378023117737206