LMU
LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

INSTITUT FÜR STATISTIK

Simon Klau, Felix Schönbrodt, Chirag Patel, John Ioannidis,
Anne-Laure Boulesteix, Sabine Hoffmann

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

Technical Report Number 232, 2020
Department of Statistics
University of Munich

http://www.statistik.uni-muenchen.de

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

[1]Institute for Medical Information Processing, Biometry, and Epidemiology, Ludwig-Maximilians-Universität München, Munich, Germany

[2]Department of Psychology, Ludwig-Maximilians-Universität München, Munich, Germany

[3]LMU Open Science Center, Ludwig-Maximilians-Universität München, Munich, Germany

[4]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

[5]Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA, USA

[6]Department of Epidemiology and Population Health, Stanford University, Stanford, CA, USA

[7]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA

[8]Department of Statistics, Stanford University, Stanford, CA, USA

February 5, 2020

*Corresponding author: e-mail: simon.klau@yahoo.de, Department of Medical Information Processing, Biometry and Epidemiology, Ludwig-Maximilians-Universität München, Marchioninistr. 15, D-81377, Munich, Germany

Abstract

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 observational associations, namely data pre-processing, model and sampling uncertainty. We analyze their behavior for varying sample sizes for two associations in personality psychology. 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 the transparent reporting of the stability of research findings.

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

1 Introduction

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, Anaya, & Brown, 2017), many researchers agree that this is not the major problem causing the crisis (Gelman & Loken, 2014; Ioannidis, Munafo, Fusar-Poli, Nosek, & David, 2014). Instead, the problems seem to be more subtle and partly due to the multiplicity of possible analysis strategies (Goodman, Fanelli, & Ioannidis, 2016; Open Science Collaboration, 2015). In this vein, there is evidence that the instability of observational associations can partly be explained by the fact that researchers tend to run several analysis strategies on a given data set, but to report only one of them selected post-hoc (Simmons, Nelson, & Simonsohn, 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. 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 even before being able to fit 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 & 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 increases the false positive rate while simultaneously decreasing 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 the results of a large number of possible analysis strategies (Muñoz & Young, 2018; Simonsohn, Simmons, & Nelson, 2015; Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016; Young, 2018), including the vibration of effects, proposed by Ioannidis (2008) and further developed by Patel, Burford, and Ioannidis (2015), and Palpacuer et al. (2019). Alternatively, the flexibility in the choice of analysis strategies can be reduced before analyzing the data through preregistration and registered reports (Chambers, 2013; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). Similarly, the instability of observational findings arising from sampling uncertainty can be assessed through resampling (Meinshausen & Bühlmann, 2010; Sauerbrei, Boulesteix, & Binder, 2011) or sampling uncertainty can be reduced by increasing the sample size (Button et al., 2013; Maxwell, 2004; Schönbrodt & Perugini, 2013). While the solutions which have been 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 (2019) 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.

2 Methods

2.1 The data and research questions of interest

We use a large data set from the SAPA project personality test (Condon, Roney, & Revelle, 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 observational associations between the Big Five (agreeableness, conscientiousness, extraversion, neuroticism, openness to experience) and the 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 $n \in \{500, 5000, 15000, 50000, 84543\}$, where 84543 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 (Malouff, Thorsteinsson, Schutte, Bhullar, & Rooke, 2010) and between extraversion and physical activity (Rhodes & Smith, 2006). Additional results on the association between agreeableness and smoking (Malouff, Thorsteinsson, & Schutte, 2006), neuroticism and obesity (Gerlach, Herpertz, & Loeber, 2015), and conscientiousness and education (Sorić, Penezić, & Burić, 2017) can be found in the Supplementary Material, together with results on openness and physical activity, for which no evidence for an association could be found (Rhodes & Smith, 2006).

2.2 Quantifying the instability of observational associations due to different sources of uncertainty through the vibration of effects framework

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 2.1 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 observational associations through the vibration of effects framework proposed by Ioannidis (2008). 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, resulting from the inclusion or exclusion of a number of potential control variables. To quantify the variability in these results, they calculate 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. Furthermore, the authors propose visualizing -log10(p-values) and hazard ratios with volcano plots. These plots allow easy detection of a Janus effect, which is characterized by significant results in both positive and negative directions.

In this work, we will refer to the type of vibration investigated by Patel et al. (2015) as ‘model vibration’ and extend the framework to subsamples of the data and data pre-processing choices in order to compare model vibration to ‘sampling vibration’ and ‘data pre-processing vibration’, which we will introduce in more detail in sections 2.2.2 and 2.2.3. Following the proposal of 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.

2.2.1 Model vibration

In order to assess model vibration for a given association of interest, we will consider a logistic regression model for which we take any possible combination of control variables into account. 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.

2.2.2 Sampling vibration

To quantify sampling vibration, we use a resampling-based framework where we draw a large number of random subsets from our data set and fit the same logistic regression model on each of these subsets. In particular, we draw 1000 subsets of size $0.5n$, with n as the number of observations from the data sets defined in section 2.2, which comprise different numbers of observations themselves. Although each subset is drawn without replacement, the observations of subsets overlap between repetitions.

2.2.3 Data pre-processing vibration

The data pre-processing choices we are considering include 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 (definition 1) or a subgroup is defined by excluding participants who are younger than 18 (definition 2), which can be justified by their inability to legally provide consent (Barchard & Williams, 2008). Furthermore, studies about associations involving the Big Five personality traits are often carried out on subgroups of gender or countries, as was for instance shown by Malouff et al. (2006) and Malouff et al. (2010) for the variables smoking and physical activity. Therefore, with regards to the gender of participants, we either perform the analyses with all participants (definition 1), only with male participants (definition 2), or only with female participants (definition 3). Finally, 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 (definition 1), or we include only participants from the United States, which presents the single largest country in the data set. In this case (definition 2), we exclude the control variable specifying the continent from the analyses. In total, this results in $3 \cdot 2 \cdot 2 = 12$ 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 that are used to define outliers. Furthermore, it is either possible to remove or winsorize outlier values (Osborne & 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 (definition 1). As a second option, we delete observations with absolute z-values greater than 2.5 (definition 2). Finally, we perform winsorization to achieve absolute z-values less than or equal 2.5 (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 and covariates In the definition of the outcome and covariates, we only consider the influence of different pre-processing choices for the three variables smoking, physical activity and education. 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 & 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 (definition 1) or based on non-smokers (never smokers and study participants who did not smoke the previous year) versus all other study participants (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' (definition 1) or, alternatively, 'less than once per month' versus 'less than once per week or more' (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 (definition 1), because they will soon obtain a university degree or to the group with a low level of 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 (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 (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. Here, we either choose a minimal factor loading of 0.3 (definition 1) or of 0.4 (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 1152 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.

2.3 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. 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, if we focus on sampling vibration only, we need to decide on a favorite model as well as a favorite data pre-processing choice. As favorite data pre-processing choice, we consider data 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.

2.4 Comparing the vibration due to the choice of the analysis strategy with sampling vibration

In addition to the investigation of individual types of vibration, we aim at quantifying the joint impact of model and data pre-processing choices on the variability of results. For simplicity, we will refer to the combination of a model and all necessary data pre-processing choices as analysis strategy. Correspondingly, this combination of choices results in $1024 \times 1152 = 1179648$ 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 884736.

In the joint investigation of model and data pre-processing choices, the calculation of ROR is straightforward and can give an estimate for the absolute amount of vibration caused by the analysis strategy.

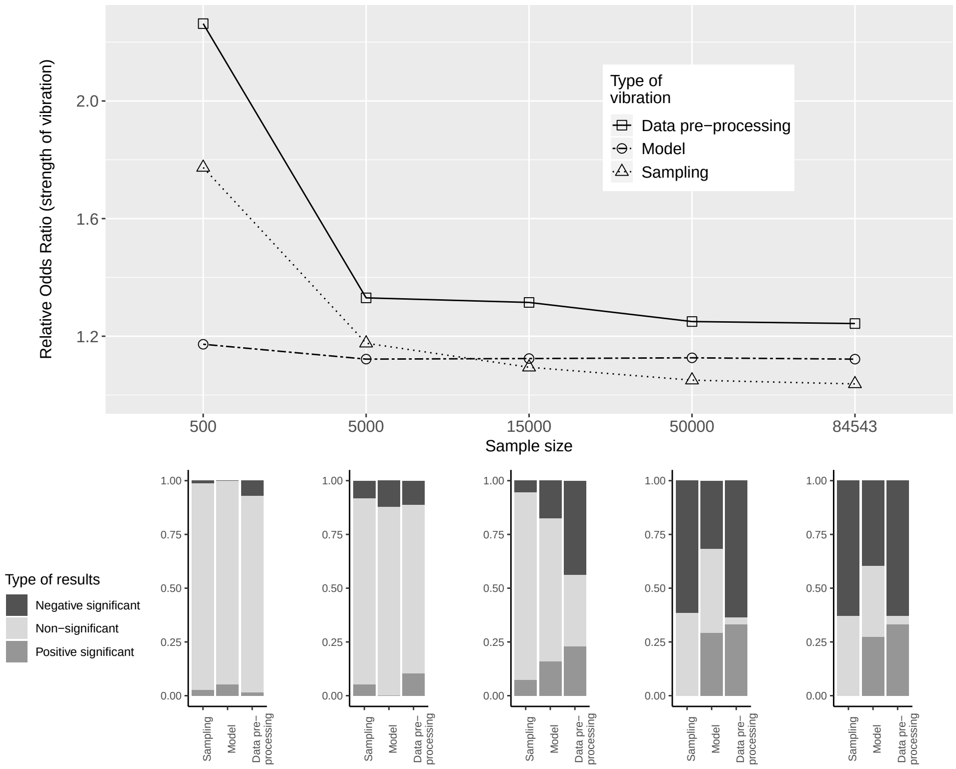


Figure 1: Data pre-processing, model, and sampling vibration for different sample sizes (top panel), and bar plots visualizing the type of results in terms of significance of estimated effects (bottom panel) for the association between neuroticism and relationship status.

Additionally, 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. This is done by modelling log(OR), corresponding to the regression coefficient of the predictor of interest, with two categorical covariates, indicating data pre-processing and model choices, in a linear model and by performing a variance decomposition through an analysis of variance (ANOVA).

3 Results

3.1 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 1 and 2, respectively.

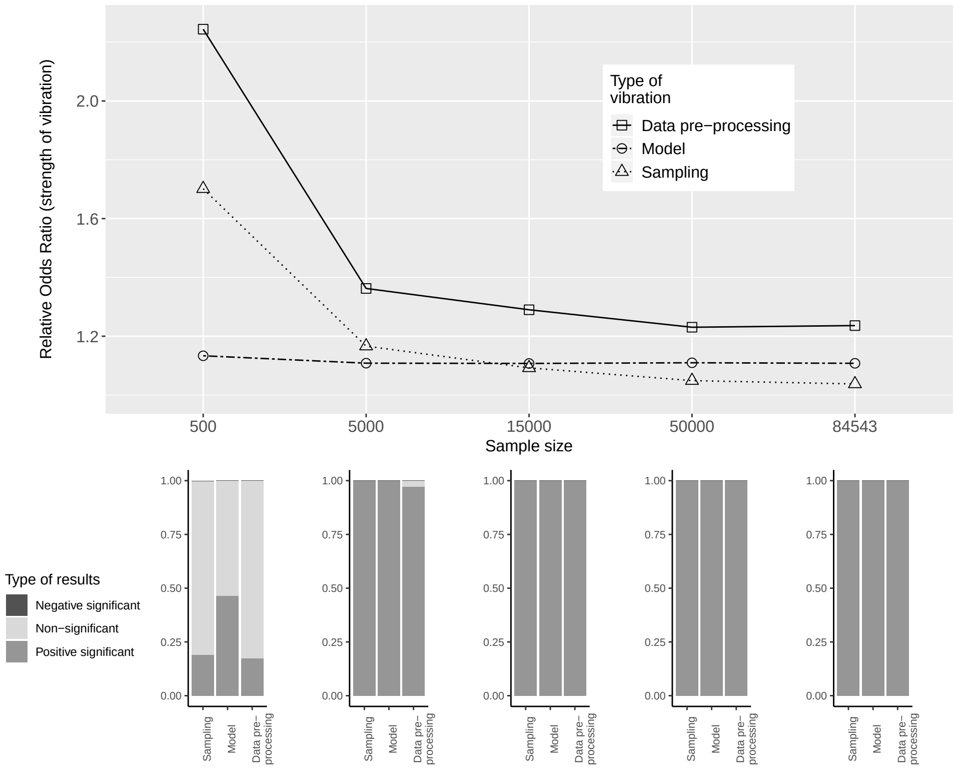


Figure 2: Data pre-processing, model, and sampling vibration for different sample sizes (top panel), and bar plots visualizing the type of results in terms of significance of estimated effects (bottom panel) for the association between extraversion and physical activity.

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-processing, model and sampling). For all investigated sample sizes, data pre-processing vibration is higher than model and sampling vibration for both associations of interest. For the lowest sample size of 500, high RORs can be observed, for instance of 2.26 for the association between neuroticism and relationship status. For larger sample sizes, data pre-processing vibration decreases and tends to a value of 1.24 for both associations of interest. A similar behavior can be observed for sampling vibration, but with consistently lower RORs. These tend to 1 for large sample sizes, in contrast to those RORs quantifying data pre-processing vibration. Therefore, the influence of a specific sample can be expected to be negligible for sufficiently large sample sizes. Compared to data pre-processing and sampling vibration, model vibration 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.

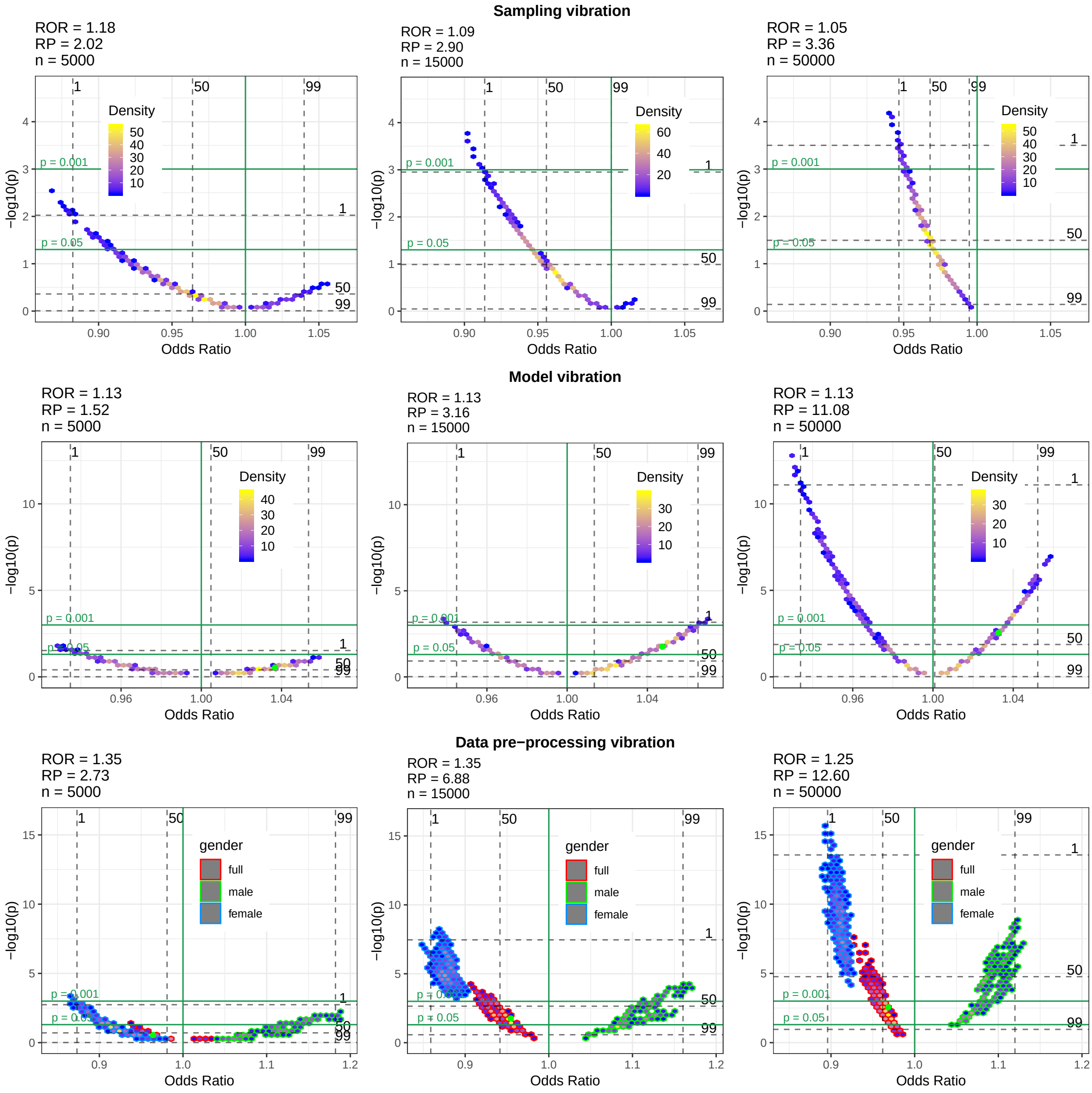
In the lower panels of Figures 1 and 2, bar plots provide information about the percentage of significant results for each sample size and each type of vibration for the three categories: "negative significant", "non-significant", and "positive significant". For all three types of vibration, most results are not significant for a sample size of 500 while for the larger sample sizes the results are mostly significant: Here, the association between neuroticism and relationship status shows a Janus effect with both negative and positive significant results for model and data pre-processing vibration. For sampling vibration, on the other hand, only negative-significant or non-significant effects can be observed for large sample sizes. 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 & Smith, 2006). Hence, a Janus effect cannot be observed for this association.

These results are underlined by volcano plots (Figures 3 and 4), which contain exact patterns of -log10(p-value) and ORs for three different sample sizes. Here, irregular patterns in data pre-processing vibration can be detected, which contrasts with the more consistent patterns of sampling and model vibration. A closer look at the data pre-processing vibration reveals that three clusters can be clearly distinguished, resulting from the pre-processing choice for the control variable gender. For male participants, neuroticism is associated with a committed relationship. On the other hand, there are two clusters with the opposite sign: Both female participants as well as the full data without subgroups for the variable gender are associated with a predominantly negative association between neuroticism and relationship status. The larger the sample size, the more clearly the clusters can be distinguished.

3.2 The relative impact of model and data pre-processing 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 5 for the association between neuroticism and relationship status, and Figure 6 for the association between extraversion and physical activity. In these figures, the top panels allow for a comparison of this joint vibration, also referred to as vibration due to the analysis strategy, and sampling vibration. The vibration caused by the analysis strategy is higher than sampling vibration for both associations. For a low sample size of $n = 500$, it is considerably higher than for larger sample sizes with RORs of 2.02 and 2.00 for the association between neuroticism and relationship status, and extraversion and physical activity, respectively. For a sample size of 5000, ROR values of 1.39 and 1.36 can be observed for these associations, which indicate lower vibration. For larger sample sizes, however, RORs do not show any further obvious decrease. For sample sizes greater than 500, the vibration remains in the range of ROR values of 1.34 and 1.40 for the association between neuroticism and relationship status. Similarly, the RORs for sample sizes greater than 500 are in the range of 1.27 and 1.36 for the association between extraversion and physical activity.

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



Sampling vibration

Model vibration

Data pre−processing vibration

Figure 3: 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).

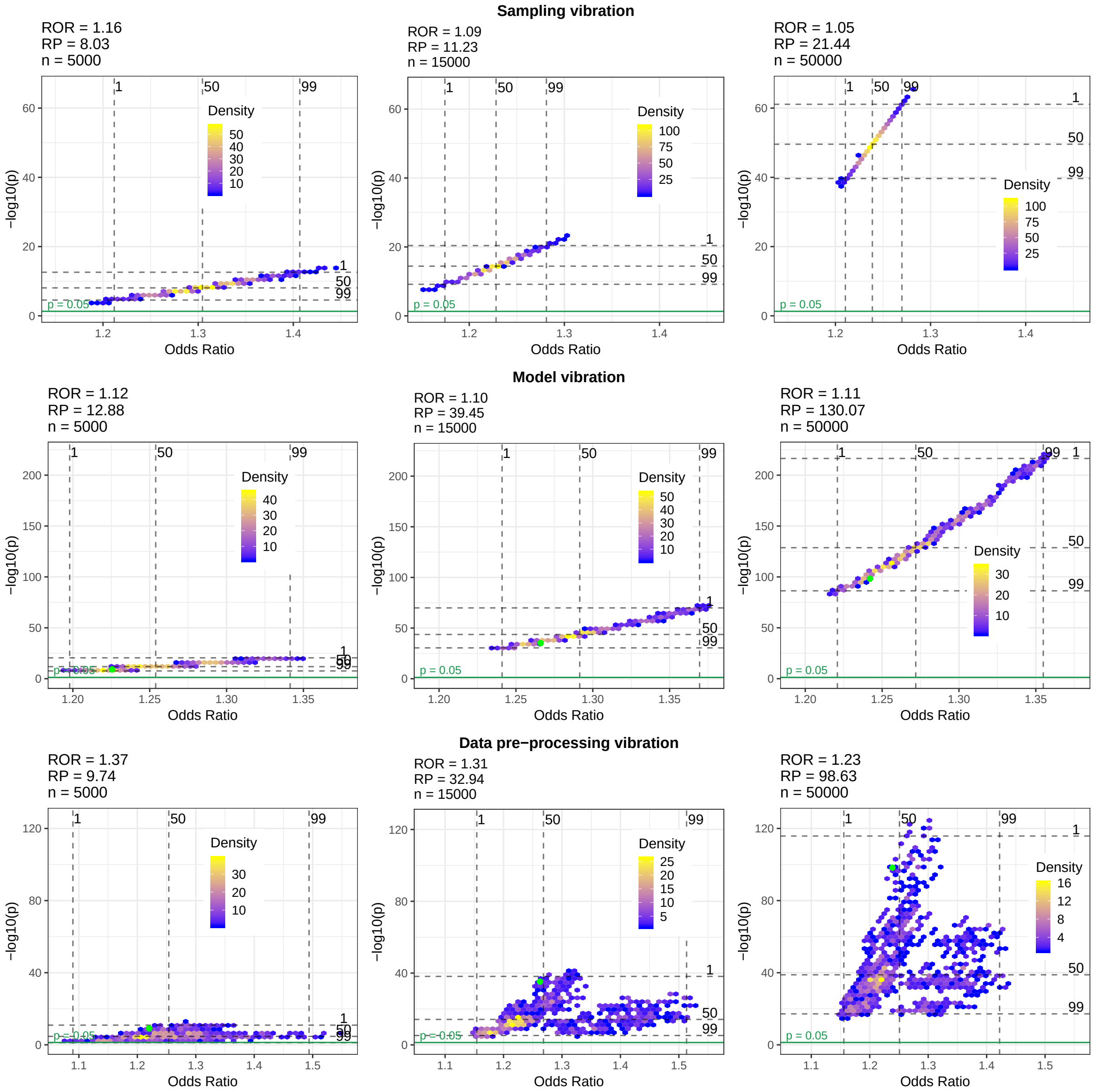


Figure 4: Volcano plots for different types of vibration and different sample sizes (n) for the association between extraversion and physical activity. The summary measures ROR and RP indicate relative odds ratios and relative p-values, respectively. Green dots indicate results obtained with favorite model choices (middle row) and favorite data pre-processing choices (bottom row).

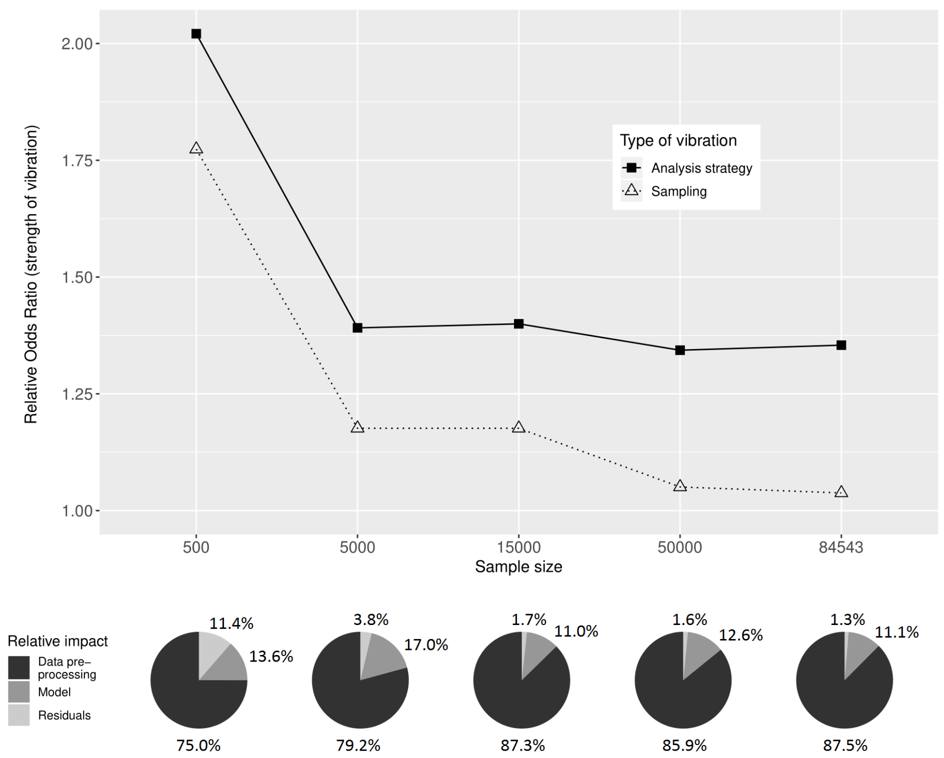


Figure 5: Cumulative model and data pre-processing vibration ('analysis strategy') compared to sampling vibration (top panel), and relative impact of model and data pre-processing vibration for different sample sizes (bottom panel) for the association between neuroticism and relationship status.

choices by far exceeds the impact of model vibration. Indeed, at most 22.2% of the total vibration due to the analysis strategy can be explained by model choices for the association between extraversion and physical activity. For the association between neuroticism and relationship status, the relative model impact is even lower, with a maximum value of 17%.

A more detailed investigation of data pre-processing vibration as part of the total vibration shows that the variable gender has the largest impact of the data pre-processing choices on the vibration of effects for the both associations of interest. Indeed, for the association between neuroticism and relationship status, 86% of data pre-processing vibration can be explained by the impact of gender for the largest sample size, which is in accordance with Figure 5. For the association between extraversion and physical activity, the relative impact of gender on data pre-processing vibration is 59.2% for the full data set.

4 Discussion

4.1 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

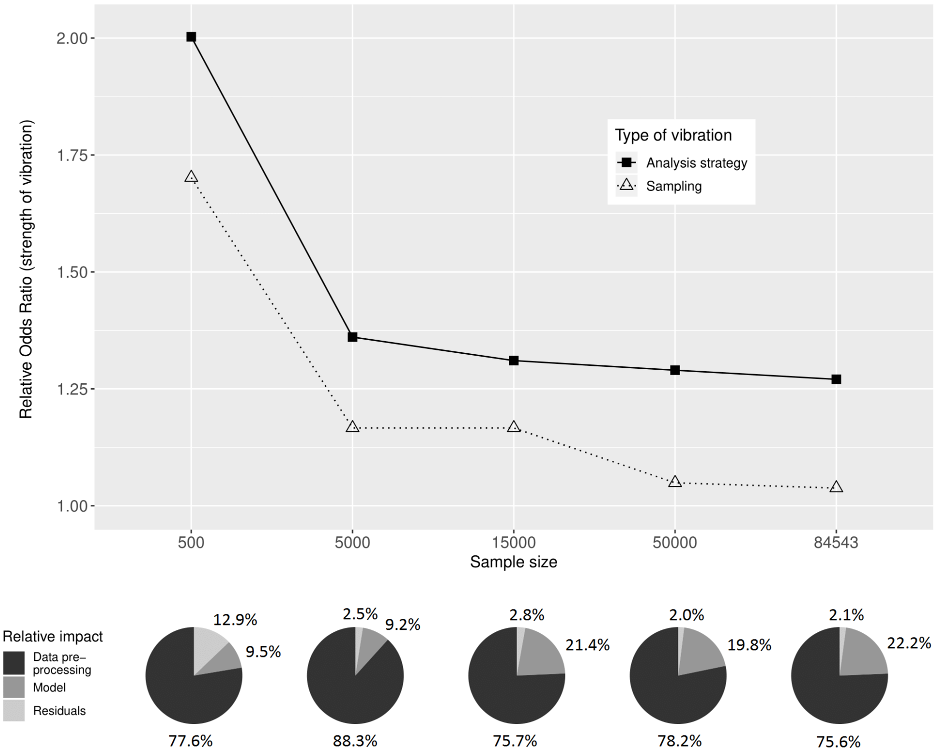


Figure 6: Cumulative model and data pre-processing vibration (‘analysis strategy’) compared to sampling vibration (top panel), and relative impact of model and data pre-processing vibration for different sample sizes (bottom panel) for the association between extraversion and physical activity.

by Ioannidis (2008), to quantify and compare the impact of model and data pre-processing choices on the stability of observational associations. Through this extension, the vibration of effects framework allows assessment of the extent to which the instability of research findings in observational studies can be explained by the choice of analysis strategy and enables comparison of the relative 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. In addition, we quantified sampling vibration by considering the results obtained from random subsets of the data set in use. We found that data pre-processing vibration was higher than model and sampling vibration for all sample sizes considered in our analyses. For high sample sizes, sampling vibration decreased and became negligible, while model and data pre-processing vibration showed an initial decrease with increasing sample size and then remained constantly non-negligible. When considering all possible combinations of model and data pre-processing choices to compare the relative impact of each source of uncertainty, we found that data pre-processing choices explained by far more variability in results than model choices.

4.2 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 reasonable analysis strategies, i.e., those that could have been selected in an actual research project. 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 did not examine interaction terms, which may be essential in some settings.

Finally, we considered a number of possible data pre-processing strategies that is comparable to the number of models in order to allow a fair comparison of data pre-processing uncertainty and model uncertainty. As the combination of model and data pre-processing choices was in the order of magnitude of one million, it would not have been feasible from a computational point of view to consider a larger combination of models and data pre-processing strategies. As a consequence, we have to be careful when generalizing the findings of our study to other data sets and applications. 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. 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.

While the number of analysis strategies we considered in this work was limited by the computational feasibility of our analyses, it has to be noted that this number might in principle be reduced by only selecting those models that show a reasonable fit to the data. In the vibration of effects framework, the results of all possible models are reported, regardless of the fit of these models. In this respect, the vibration of effects framework differs from other approaches like Bayesian Model Averaging (Hoeting, Madigan, Raftery, & Volinsky, 1999), where a single summary measure is obtained that accounts for model uncertainty by weighting every model under consideration by its probability of being the true model. On the other hand, the vibration of effects framework shows greater flexibility than Bayesian Model Averaging as it can report the results of not only different models, but also different data pre-processing choices. Contrary to the choice of a model, data pre-processing choices may often be based on untestable assumptions, concerning for instance the nature of outlying observations, or they may arise because scientific theories are generally not precise enough to allow for a one-to-one mapping to statistical hypotheses (Steegen et al., 2016). Contrary to model uncertainty, data pre-processing uncertainty therefore cannot be reduced by comparing the fit of different data pre-processing strategies to the data, but only through conceptual rigor (Schaller, 2016) and the standardization of experimental conditions (Elson, Mohseni, Breuer, Scharkow, & Quandt, 2014).

4.3 Conclusion and Outlook

When analyzing observational data, it is necessary to make model and data pre-processing choices which rely on many explicit and implicit assumptions. The vibration of effects framework provides investigators with a tool to quantify the impact of these choices on the stability of observational associations, helping them focus their attention on the choices that have the most influence and are therefore worth further investigation or discussion. To establish it 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. The corresponding analysis took 21.6 minutes on a 64-bit Debian GNU/Linux 10 system with Intel Xeon CPU E5-2640. 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 observational 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 more and more 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, Jost, Nagler, Tucker, & Bonneau, 2015) or transaction data (Gladstone, Matz, & Lemaire, 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.

Acknowledgements

This work was funded by the DFG (individual grant BO3139/4-3). The authors of this work take full responsibilities for its content. The authors thank Alethea Charlton for language corrections.

Bibliography

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. doi: 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. doi: 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. doi: 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. doi: 10.1038/nrn3475

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

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

Elson, M., Mohseni, M. R., Breuer, J., Scharkow, M., & Quandt, T. (2014). Press CRTT to measure aggressive behavior: The unstandardized use of the competitive reaction time task in aggression research. *Psychological Assessment*, *26*(2), 419–432. doi: 10.1037/a0035569

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

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. doi: 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. doi: 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. doi: 10.1126/scitranslmed.aaf5027

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*(4), 382–401.

Ince, D. (2011). The duke university scandal – what can be done? *Significance*, *8*(3), 113–115. doi: 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. doi: 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. doi: 10.1016/j.tics.2014.02.010

Klau, S., Martin-Magniette, M.-L., Boulesteix, A.-L., & Hoffmann, S. (2019). Sampling uncertainty versus method uncertainty: A general framework with applications to omics biomarker selection. *Biometrical Journal*, 1–18. doi: 10.1002/bimj.201800309

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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 10.1177/0081175018777988

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. doi: 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. doi: 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. doi: 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. doi: 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. doi: 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. doi: 10.1080/10543406.2011.629890

Schaller, M. (2016). The empirical benefits of conceptual rigor: Systematic articulation of conceptual hypotheses can reduce the risk of non-replicable results (and facilitate novel discoveries too). *Journal of Experimental Social Psychology*, *66*, 107–115. doi: 10.1016/j.jesp.2015.09.006

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

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. doi: 10.1177/0956797611417632

Simonsohn, U., Simmons, J., & Nelson, L. D. (2015). Specification curve: Descriptive and inferential statistics on all reasonable specifications.
doi: 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. doi: 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. doi: 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. doi: 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. doi: 10.1186/s40795-017-0167-x

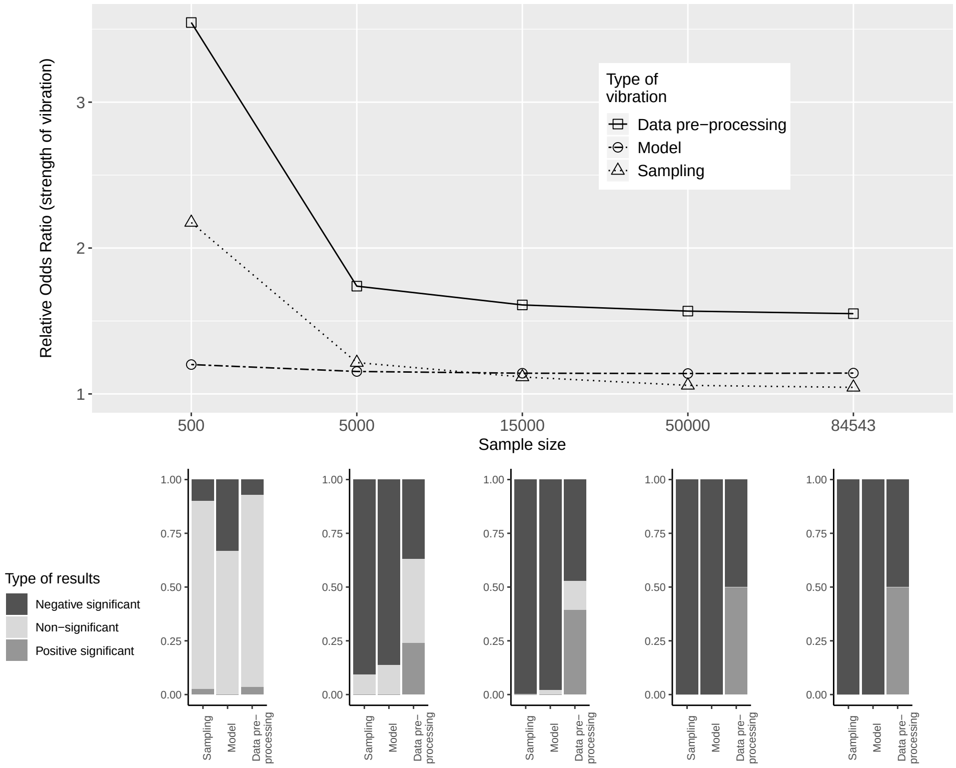
Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632–638. doi: 10.1177/1745691612463078

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., &

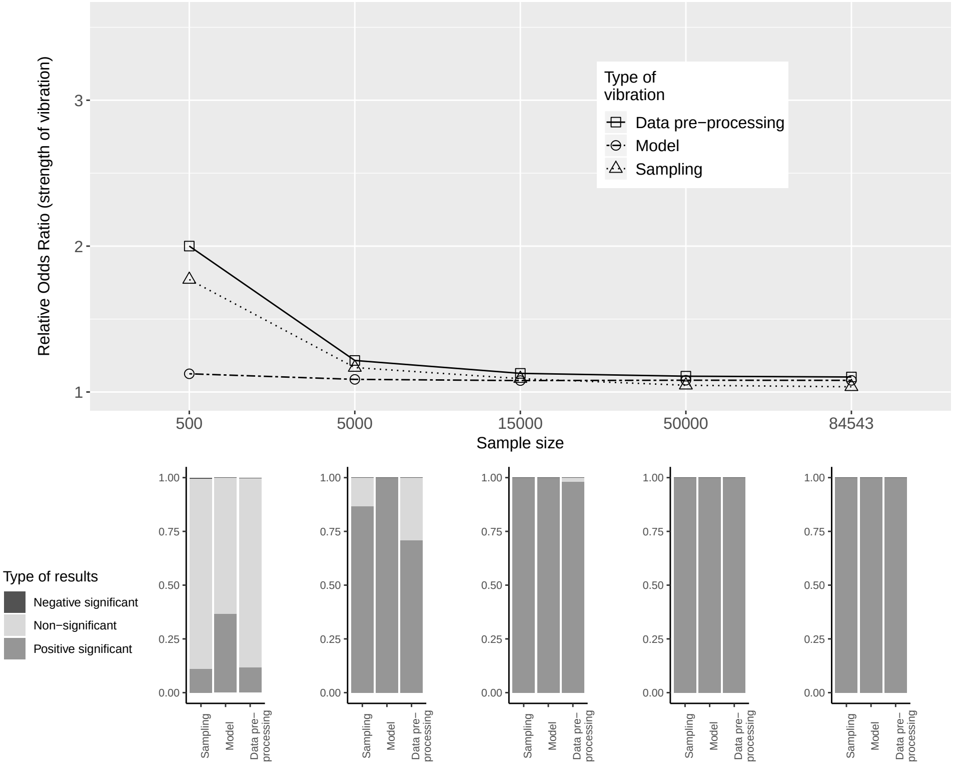
van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*(1832), 1–12. doi: 10.3389/fpsyg.2016.01832

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

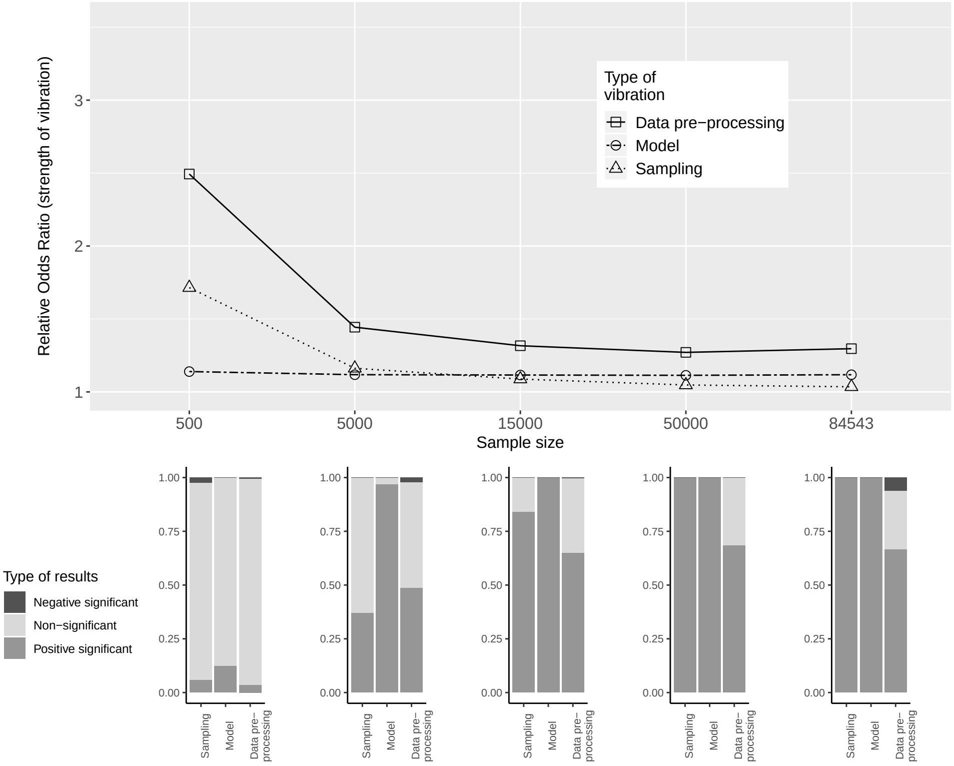
Supplementary Material



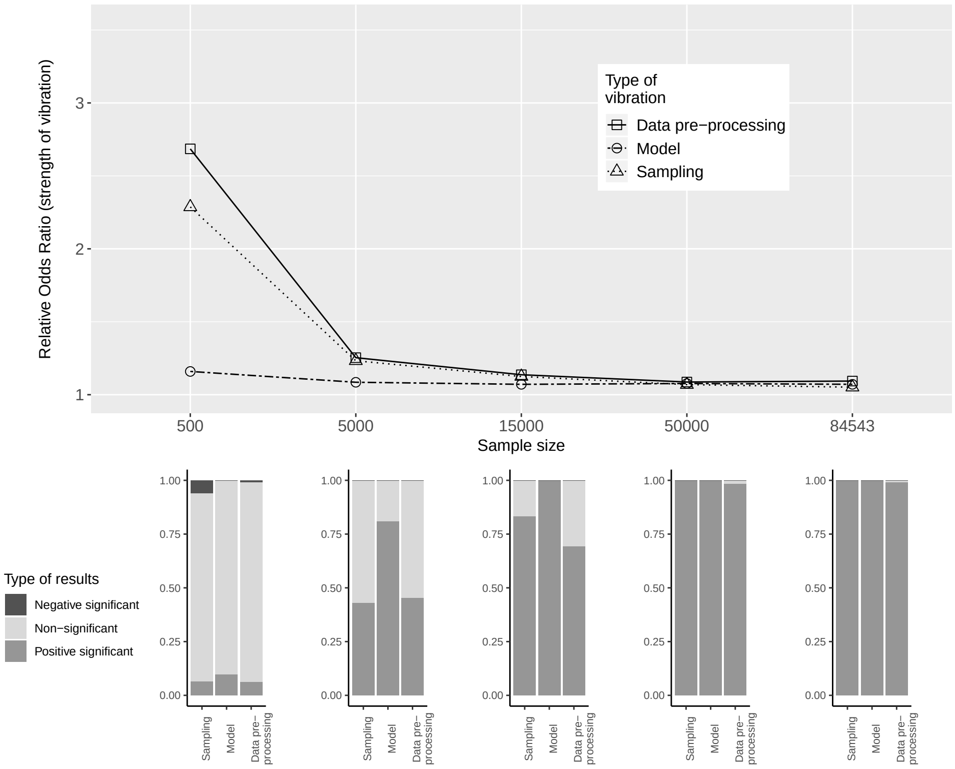
Supplementary Figure 1: Data pre-processing, model, and sampling vibration for different sample sizes (top panel), and bar plots visualizing the type of results in terms of significance of estimated effects (bottom panel) for the association between conscientiousness and education.



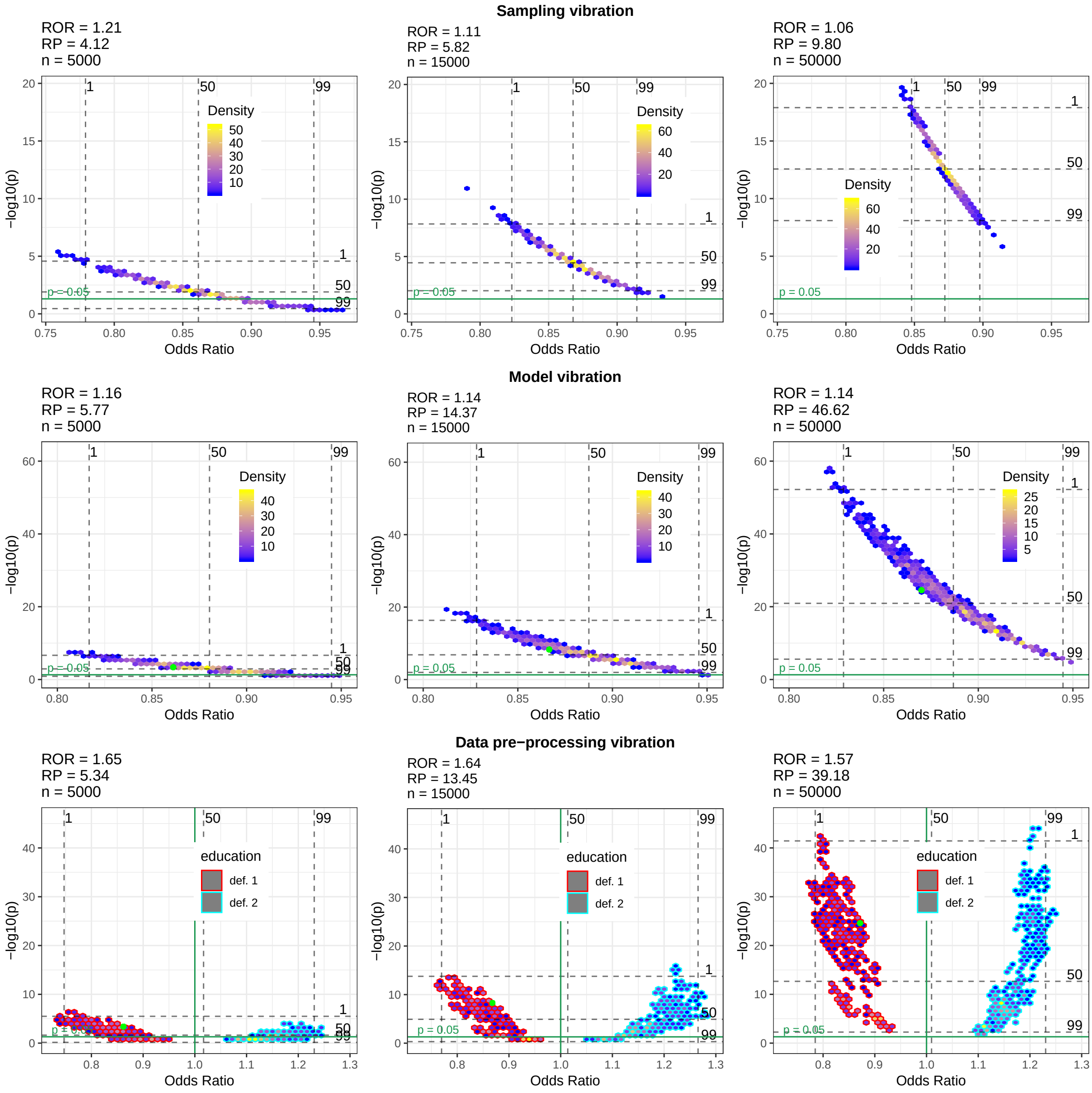
Supplementary Figure 2: Data pre-processing, model, and sampling vibration for different sample sizes (top panel), and bar plots visualizing the type of results in terms of significance of estimated effects (bottom panel) for the association between agreeableness and smoking.



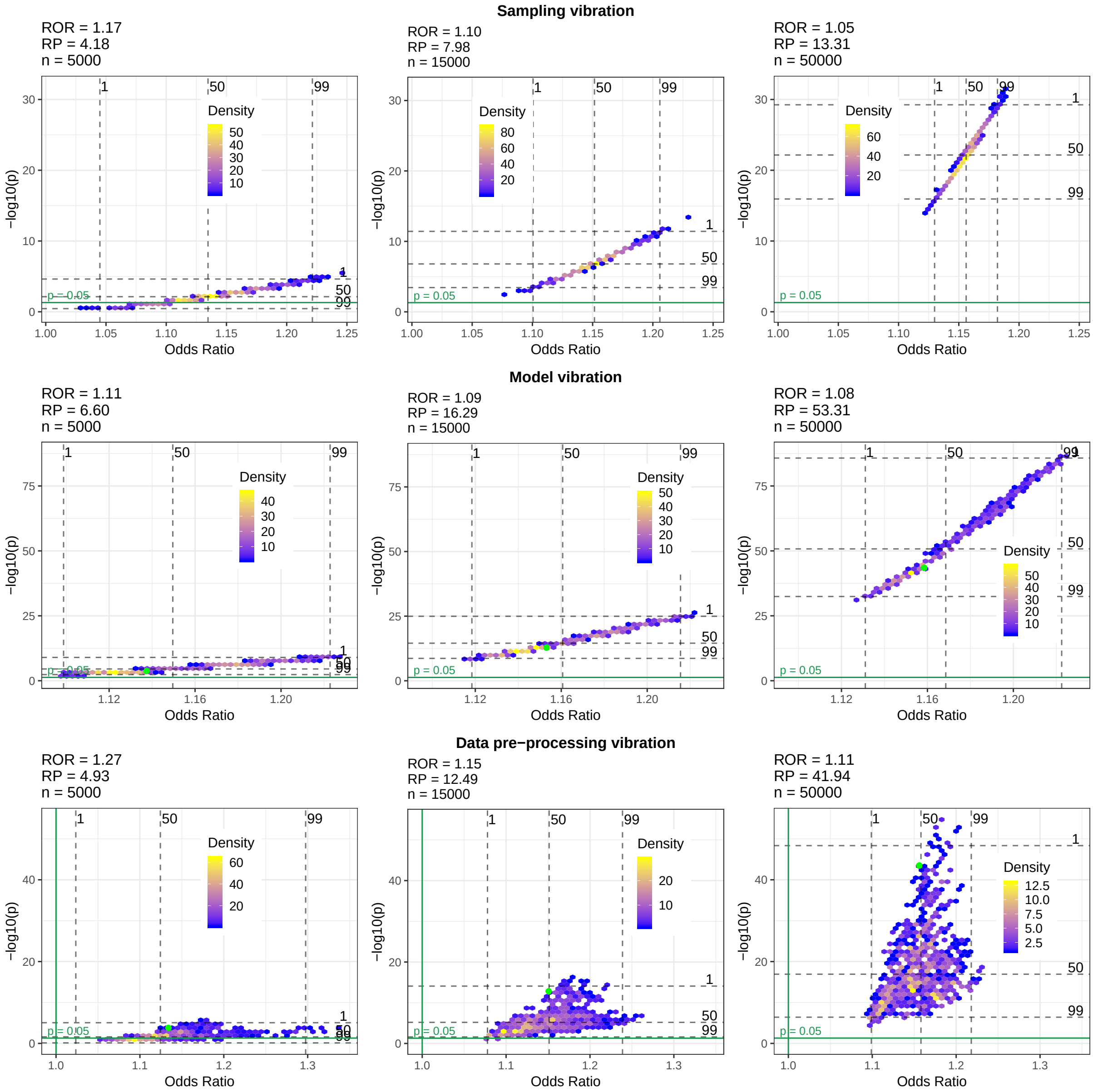
Supplementary Figure 3: Data pre-processing, model, and sampling vibration for different sample sizes (top panel), and bar plots visualizing the type of results in terms of significance of estimated effects (bottom panel) for the association between openness and physical activity.



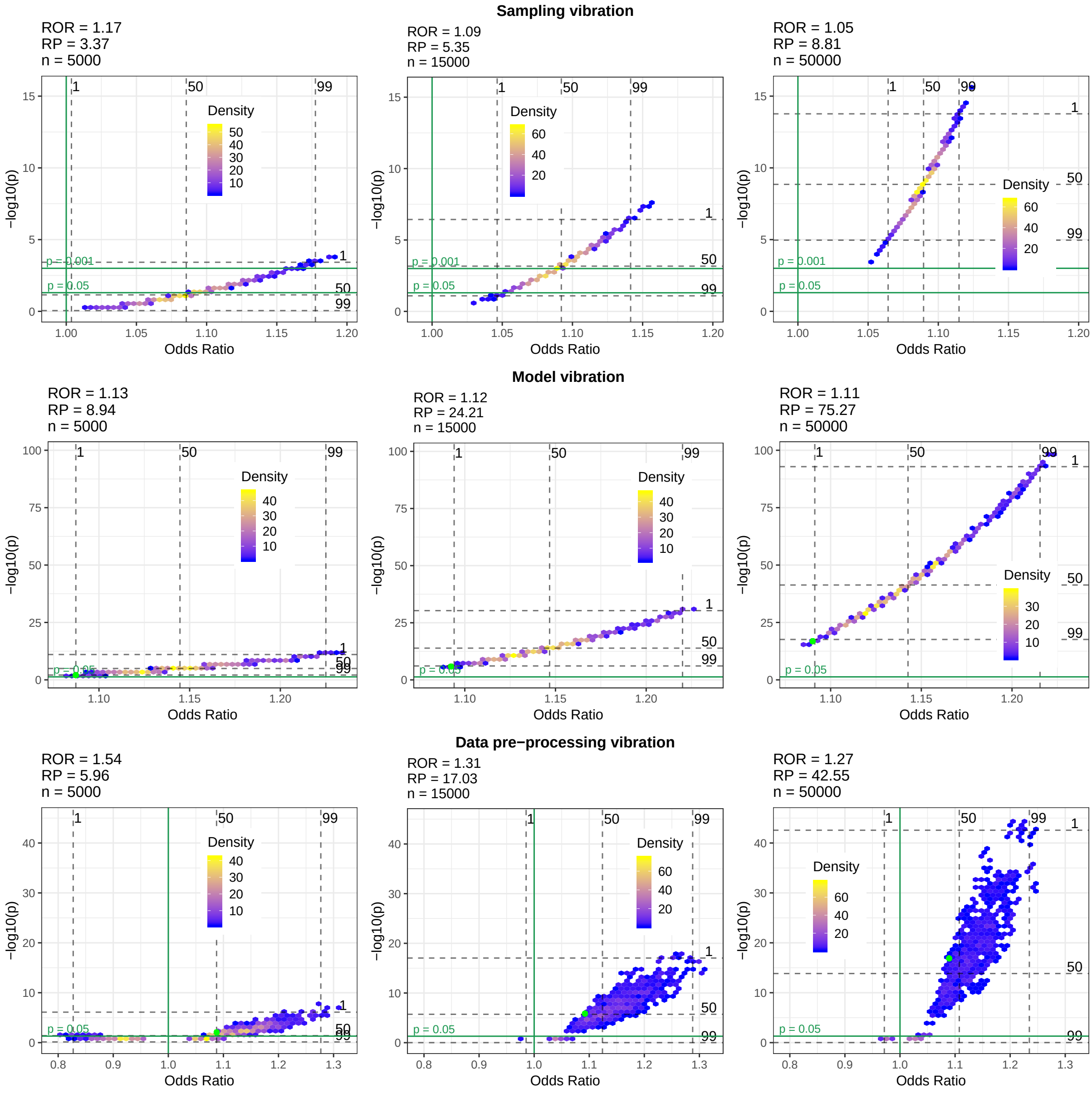
Supplementary Figure 4: Data pre-processing, model, and sampling vibration for different sample sizes (top panel), and bar plots visualizing the type of results in terms of significance of estimated effects (bottom panel) for the association between neuroticism and obesity.



Supplementary Figure 5: Volcano plots for different types of vibration and different sample sizes (n) for the association between conscientiousness and education. 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).



Supplementary Figure 6: Volcano plots for different types of vibration and different sample sizes (n) for the association between agreeableness and smoking. 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).

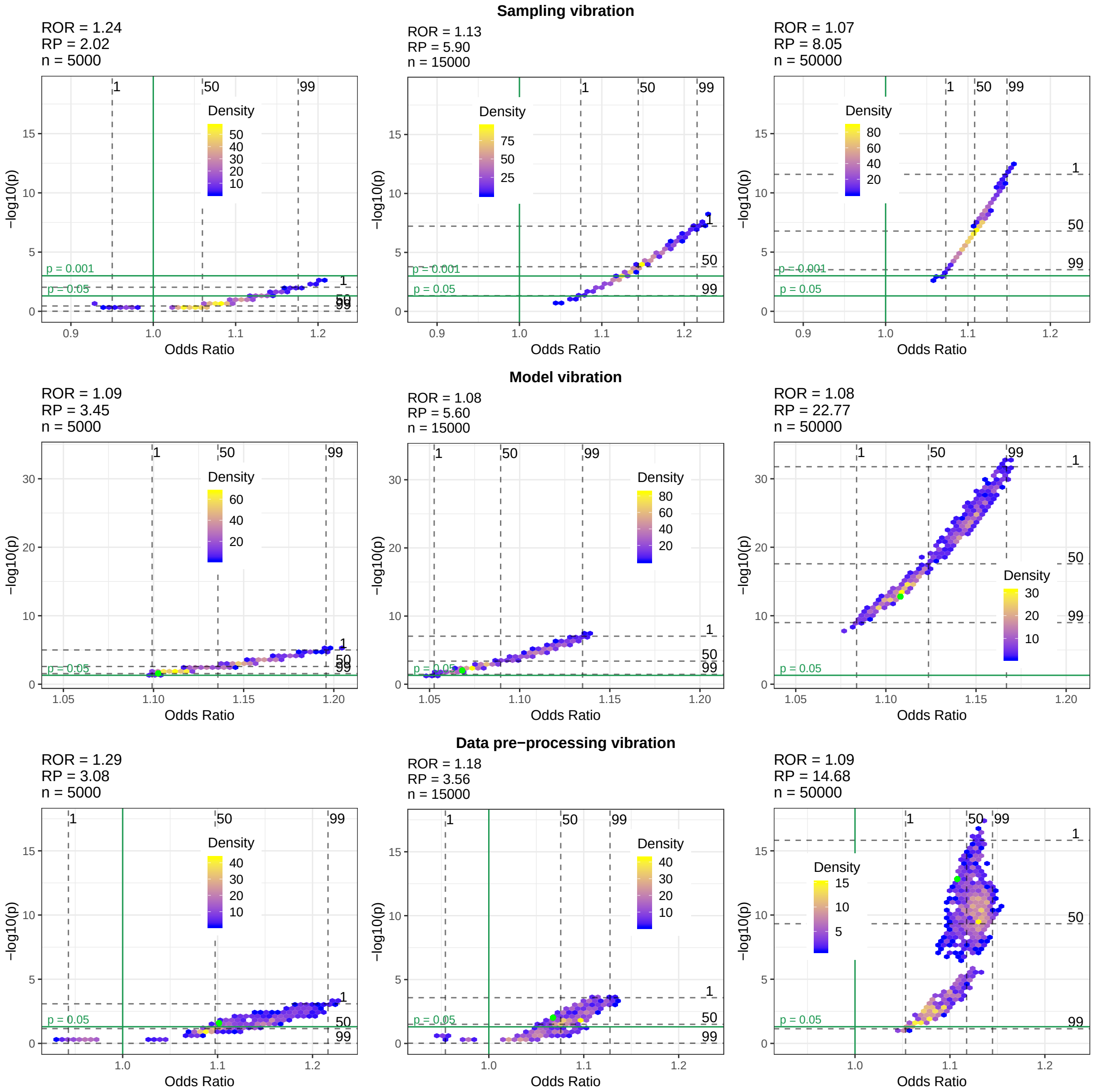


Sampling vibration

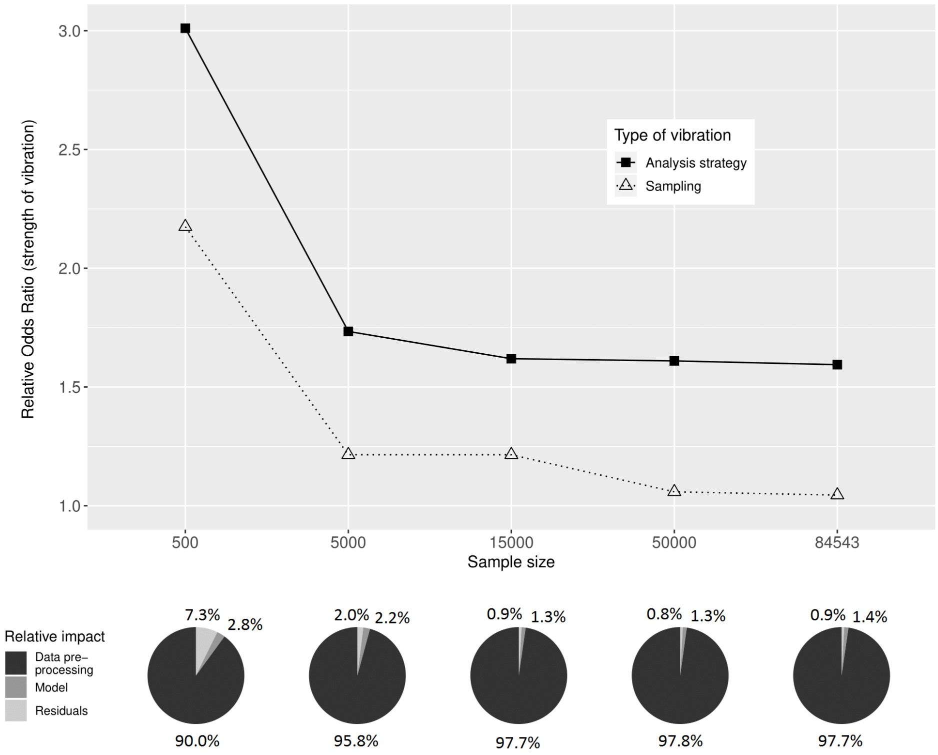
Model vibration

Data pre−processing vibration

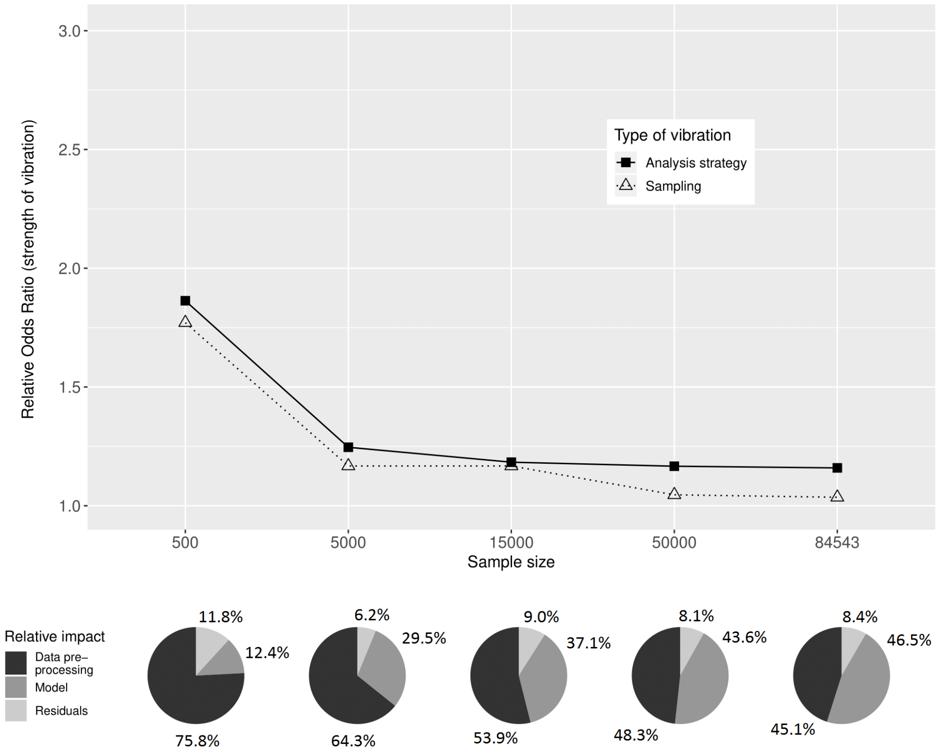
Supplementary Figure 7: Volcano plots for different types of vibration and different sample sizes (n) for the association between openness and physical activity. The summary measures ROR and RP indicate relative odds ratios and relative p-values, respectively. Green dots indicate results obtained with favorite model choices (middle row) and favorite data pre-processing choices (bottom row).



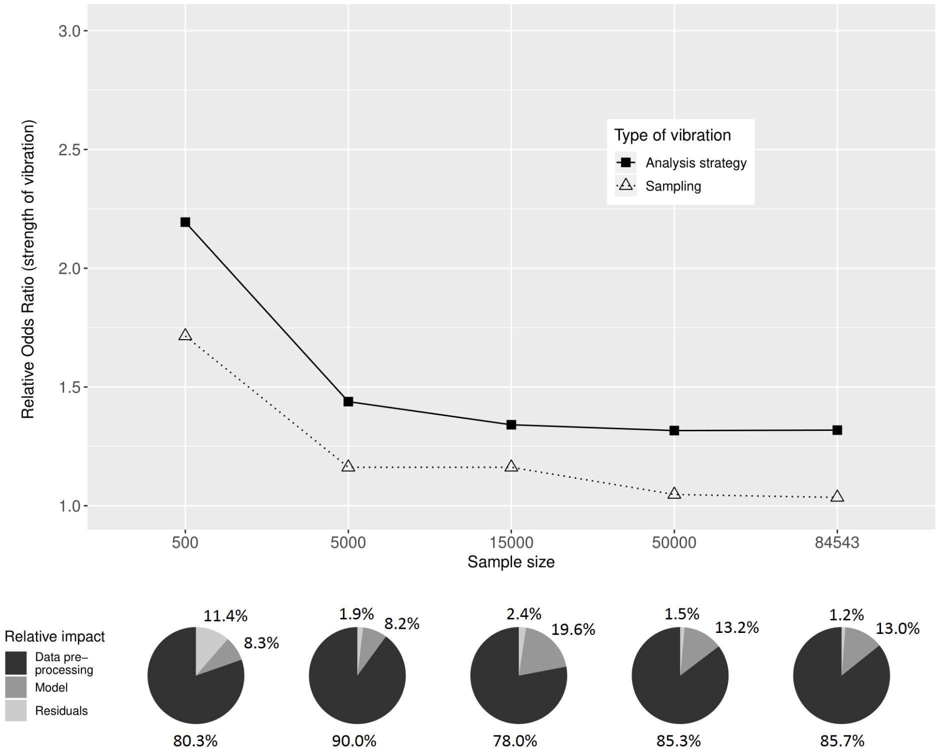
Supplementary Figure 8: Volcano plots for different types of vibration and different sample sizes (n) for the association between neuroticism and obesity. The summary measures ROR and RP indicate relative odds ratios and relative p-values, respectively. Green dots indicate results obtained with favorite model choices (middle row) and favorite data pre-processing choices (bottom row).



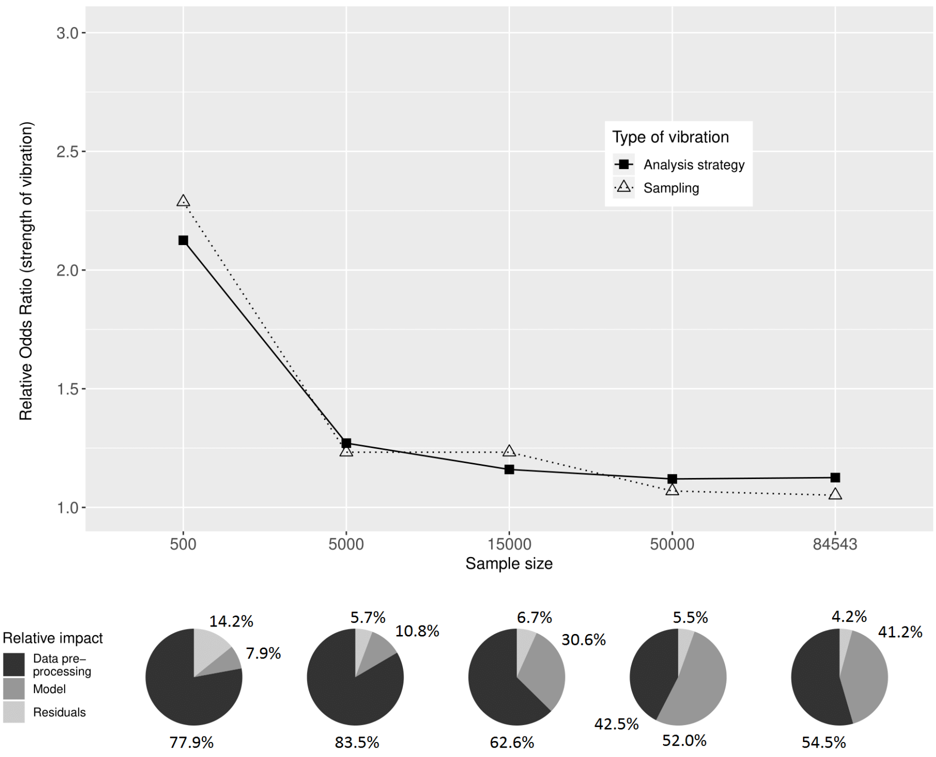
Supplementary Figure 9: Cumulative model and data pre-processing vibration ('analysis strategy') compared to sampling vibration (top panel), and relative impact of model and data pre-processing vibration for different sample sizes (bottom panel) for the association between conscientiousness and education.



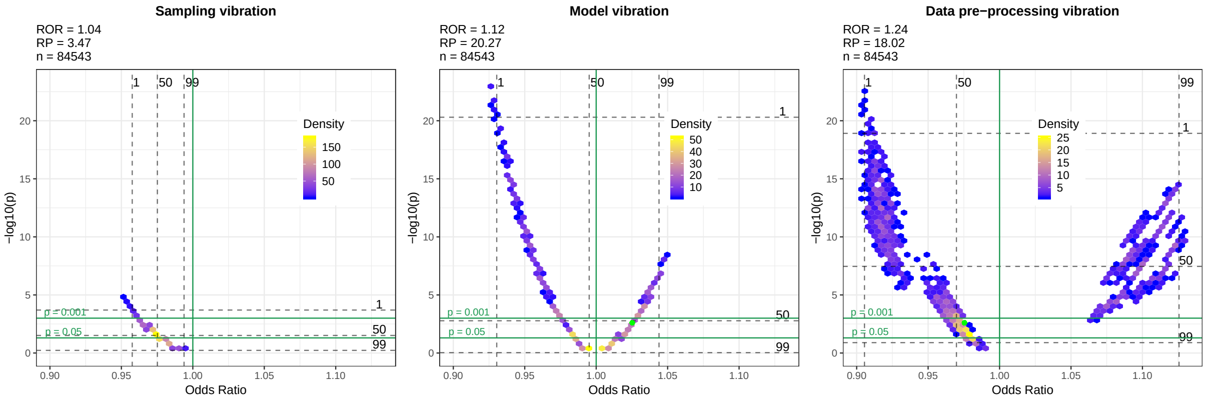
Supplementary Figure 10: Cumulative model and data pre-processing vibration (‘analysis strategy’) compared to sampling vibration (top panel), and relative impact of model and data pre-processing vibration for different sample sizes (bottom panel) for the association between agreeableness and smoking.



Supplementary Figure 11: 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 openness and physical activity.



Supplementary Figure 12: 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 obesity.



Supplementary Figure 13: Example how to use the vibration of effects as a tool for the association between neuroticism and relationship status.