Increasing Transparency through a Multiverse Analysis

Journal:	*Perspectives on Psychological Science*
Manuscript ID	PPS-15-234.R3
Manuscript Type:	Original Article
Date Submitted by the Author:	16-May-2016
Complete List of Authors:	Steegen, Sara; KU Leuven, Faculty of Psychology and Educational Sciences Tuerlinckx, Francis; KU Leuven, Faculty of Psychology and Educational Sciences Gelman, Andrew; Statistics Vanpaemel, Wolf; University of Leuven, Faculty of Psychology and Educational Sciences
Keywords:	Methodology: Quantitative, Methodology: Scientific
User Defined Keywords:	Multiverse analysis, Data processing, Good research practices, Fertility research

Increasing Transparency through a Multiverse Analysis

Sara Steegen
KU Leuven - University of Leuven

Francis Tuerlinckx
KU Leuven - University of Leuven

Andrew Gelman
Columbia University

Wolf Vanpaemel
KU Leuven - University of Leuven

Author Note

Please send correspondence to:
Email: wolf.vanpaemel@kuleuven.be

We thank Kristina Durante for making the data and the survey material available and for helpful clarifications. Richard Morey, Don van den Bergh and several anonymous reviewers have provided valuable suggestions. The data and the code can be found on `https://osf.io/zj68b/`. The research leading to the results reported in this paper was supported in part by the Research Fund of KU Leuven (GOA/15/003) and by the Interuniversity Attraction Poles programme financed by the Belgian government (IAP/P7/06).

Abstract

Empirical research inevitably includes constructing a dataset by processing raw data into a form ready for statistical analysis. Data processing often involves choices among several reasonable options for excluding, transforming and coding data. We suggest that instead of performing only one analysis, researchers could perform a multiverse analysis, which involves performing *all* analyses across the whole set of alternatively processed data sets corresponding to a large set of reasonable scenarios. Using a worked example focusing on the effect of fertility on religiosity and political attitudes, we show that analyzing a single data set can be misleading, and propose a multiverse analysis as an alternative practice. A multiverse analysis offers an idea of how much the conclusions change because of arbitrary choices in data construction, and gives pointers as to which choices are most consequential in the fragility of the result.

Keywords: multiverse analysis, arbitrary choices, data processing, good research practices, transparency, selective reporting

Increasing Transparency through a Multiverse Analysis

Introduction

Psychology has been stirred by dramatic revelations of questionable research practices (John, Loewenstein, & Prelec, 2012), implausible findings (Wagenmakers, Wetzels, Borsboom, & Van der Maas, 2011), and low reproducibility (Open Science Collaboration, 2015; Yong, 2012). The resulting crisis of confidence has led to a wide array of recommendations for improving research practices. Commonly cited advice includes replication, high power, co-piloting, adjusting the alpha level, focusing on estimation rather than on testing, and adopting Bayesian statistics (e.g., Asendorpf et al., 2013; Bakker, van Dijk, & Wicherts, 2012; Johnson, 2013; Wagenmakers et al., 2011). A major class of recommendations involves a call for increased transparency in reporting, including pre-registration of hypotheses and analyses, clearly distinguishing between confirmatory and exploratory findings, disclosing all conditions and measures, sharing data, and sharing research materials (e.g., Chambers, 2013; LeBel, Campbell, & Loving, in press; Morey et al., 2016; Nosek & Bar-Anan, 2012; Nosek et al., 2015; Simmons, Nelson, & Simonsohn, 2012; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). In this paper, we use a worked example to suggest that research transparency can further be increased by performing what we term a *multiverse analysis*.

A multiverse analysis starts from the observation that data used in an analysis are usually not just passively recorded in an experiment or an observational study. Rather, data are to a certain extent actively *constructed*. Data construction occurs when the raw data are converted into a form ready for analysis. When preparing their data for analysis, researchers often take several processing steps, such as discretization of variables into categories, combination of variables, transformation of variables, data exclusion, and so on. These processing steps typically come with many researcher degrees of freedom (Simmons, Nelson, & Simonsohn, 2011), as there are often several options in each step. As a result, raw data do not uniquely give rise to a single data set for analysis, but rather to multiple alternatively processed data sets, depending on the specific combination of choices—a

many worlds or *multiverse* of data sets. As each data set in this data multiverse can lead to a different statistical result, the data multiverse directly implies a multiverse of statistical results.

Researchers often select a single (or a few) data processing choices and then present this as the only analysis that ever would have been done. This practice of selective reporting would not be problematic if the single data set under consideration is processed based on sound and justifiable choices. However, choosing among the possibilities during data processing is often arbitrary, and justifications for the choices are typically lacking. For example, partitioning a variable into two or more discrete categories often involves an arbitrary split point; there can be various reasonable combinations or transformations of variables; and there are different sensible guidelines to determine which data points to exclude. This multiplicity of reasonable processing steps gives rise to a multiverse of reasonable data sets, which directly implies that there are several reasonable statistical results. Any arbitrariness that is present in the data construction is inherited by the statistical result.

When privileging a single arbitrary data set from the multiverse of possible data sets, the multiverse of statistical results is ignored. The inevitable arbitrariness in the data, and the sensitivity of the result, is hidden to the readers, which makes the interpretation of the single result hard at best and impossible at worst. In the light of this problem of selective reporting, we propose to use a *multiverse analysis* as an alternative to a single data set analysis. Such a multiverse analysis has two goals: it enhances transparency by providing a detailed picture of the robustness or fragility of statistical results, and it helps identifying the key choices that conclusions hinge on.

A multiverse analysis involves performing the analysis of interest across the whole set of data sets that arise from different reasonable choices for data processing. It can be seen as a systematic and organized extension of *outlier analysis* (see, e.g., Ramsey & Schafer, 2012; Simmons et al., 2011), which involves examining the robustness of one's conclusions with and without the elimination of outlying observations. A multiverse analysis displays

the stability or robustness of a finding, not only across different options for exclusion criteria, but across different options for *all* steps in data processing. It is closely related to the idea of *garden of forking paths* in data analysis (Gelman & Loken, 2014), which highlights that the one-to-many mapping from scientific theories to statistical hypotheses typically leads to an implicit, potential multiple comparison problem. The multiverse analysis focuses on one particular aspect of this multiple comparison issue, related to data processing.

In the remainder of this paper, we demonstrate a multiverse analysis using data from recently published research. We first describe the results of an analysis focusing on a single constructed data set only. Next, we describe a multiverse analysis based on the same raw data, and highlight how the multiverse analysis reveals the impact of arbitrary processing choices on the statistical results.

Demonstration

Our demonstration of a multiverse analysis focuses on data collected by Durante, Rae, and Griskevicius (2013). These authors conducted two studies investigating the effect of fertility on religiosity and political attitudes. We selected this paper simply to illustrate how a multiverse analysis can help researchers better understand the extent to which their results depend on various data processing choices. First, we describe the raw data that were collected in both studies. Next, we describe the single data set analysis reported by Durante et al. (2013). Finally, we show what these authors could have found, had they performed a multiverse analysis of their data rather than the single data set analysis. A more detailed description of the raw and processed data is provided in the online Supplemental Materials.

Data collection

A total of 275 women participated in Study 1. Each participant was asked to answer three religiosity items using a 9-point scale. Further, each participant was asked to indicate the typical length of her menstrual cycle, the start date of her last menstrual period, and

the start date of her previous menstrual period. Additionally, each woman indicated how sure she was about these two start dates, using a 9-point scale. Finally, each woman was asked to indicate her current romantic relationship status, with the following four response options: (1) *not dating/romantically involved with anyone*, (2) *dating or involved with only one partner*, (3) *engaged or living with my partner*, and (4) *married*.

Quite laudably, Durante et al. (2013) performed a second study to replicate the findings in Study 1, and to extend them to political attitudes. In Study 2, 502 women participated. The main difference with Study 1 was that participants were also asked to answer five items assessing fiscal political attitudes, five items assessing social political attitudes (using a 7-point scale for these ten items), one item assessing their voting preference (Mitt Romney or Barack Obama), and one item assessing their campaign donation preference (Mitt Romney or Barack Obama). Another difference with Study 1 was that participants also indicated the expected start date of their next menstrual period.

Single data set analysis

The data collected in the procedure described above are not ready for analysis yet. Preparing the data set for analysis requires several processing steps and decisions. We describe the different data processing steps taken by Durante et al. (2013) to construct a single data set for each study, and the main results and conclusions that follow from this data set. The results of these single data set analyses are identical to the ones reported by Durante et al. (2013).

Constructing the single data set.

Religiosity. The three religiosity items are averaged to create a religiosity score.

Fiscal and social political attitudes. The five fiscal political attitudes items are averaged to create a fiscal political attitudes score, and the five social political attitudes items are averaged to create a social political attitudes score.

Fertility. Participants are classified in a *high* versus *low* fertility group based on their cycle day. Participants with cycle days ranging from 7 to 14 are assigned to the high fertility group, whereas participants with cycle days ranging from 17 to 25 are assigned to

the low fertility group. A woman's cycle day is based on the number of days before next menstrual onset, which in turn is based on cycle length, which is computed as the difference between the start date of the woman's last menstrual period and the start date of the woman's previous menstrual period.

Relationship status. Participants are assigned to a *single* versus *committed relationship* group. Women who selected response option (1) or (2) on the relationship status item are assigned to the group of single women, whereas women who selected response option (3) or (4) are assigned to the group of women in committed relationships.

Exclusion criteria. The assignment of the participants to a high or low fertility group automatically excludes women whose cycle days are not in the high or low fertility range. Beyond this exclusion, no other participants are excluded.

Deriving the single statistical result. Based on this single data set, the effect of fertility on religiosity and political attitudes is examined, with relationship status as an interacting variable. For religiosity, an ANOVA reveals a fertility × relationship status interaction, in both studies ($F(1, 159) = 6.46$, $p = 0.012$, in Study 1; $F(1, 299) = 8.21$, $p = 0.004$ in Study 2), indicating that single women reported less religiosity if they were in the high-fertility group than if they were in the low-fertility group, whereas women in relationships reported more religiosity if they were in the high-fertility group than in the low-fertility group. Regarding fiscal political attitudes, an ANOVA reveals no significant effects of fertility status. Regarding social political attitudes, a fertility × relationship status interaction is found, $F(1, 299) = 12.26, p = .001$, indicating that single women reported less socially conservative attitudes if they were in the high-fertility group than if they were in the low-fertility group, whereas women in relationships showed the opposite pattern. Finally, logistic regression reveals a significant fertility × relationship status interaction both for voting preferences, $b = -1.62$, $Wald(1) = 8.35$, $p = .004$, and donation preferences, $b = -1.71$, $Wald(1) = 9.30$, $p = .002$, indicating that single women were more likely to vote and donate for Obama if they were in the high-fertility group than if they were in the low-fertility group, whereas women in relationships were more likely to vote and donate for Romney if they were in the high-fertility group than if they were in the

low-fertility group.

Multiverse analysis

The different data processing steps in the single data set analysis are far from the only reasonable ones (see also Harris, Pashler, & Mickes, 2014). This means that the data set used in the single data set analysis corresponds to just a single data set in a much larger multiverse of data sets. More importantly, this also means that the statistical result based on the single data set reflects only one possible outcome in a multiverse of possible outcomes. Without knowing which other statistical results could have reasonably be observed, it is impossible to evaluate the robustness of the finding. Transparency could be increased by performing, for each research question, the same analysis for *all* possible data sets, defined by the reasonable choices for data processing. This is the multiverse analysis.

We will first construct the multiverse of data sets, which consists of all data sets that could be obtained by combining different reasonable data processing choices. Then, we analyze each data set in this data multiverse separately, leading to the multiverse of statistical results. In this multiverse analysis, we consider choices in data processing that Durante et al. (2013) might themselves have considered, had they performed a multiverse analysis rather than a single data set analysis. To increase the likelihood that these authors would have considered these choices reasonable, the different processing choices we use are based on previously published studies by Durante and her collaborators, where possible. In the same spirit, we followed Durante et al. (2013) in dichotomizing the relationship status and fertility variables, although the practice of dichotomization is not without criticism (e.g., MacCallum, Zhang, Preacher, & Rucker, 2002).[1] Further, the vicarious character of our multiverse analysis implies that for the construction of the multiverse of results, we will adopt the statistical analyses that were used by Durante et al. (2013), including the focus on p-values and the adoption of 0.05 as the significance level.

[1]For one of the six analyses of interest, Durante et al. (2013) report an additional analysis that uses a continuous measure of fertility, conception probability, rather than the dichotomized one, maybe inspired by these criticisms (see also Gangestad et al., 2016). However, since the majority of their analyses uses a dichotomized assessment of fertility, we will do so here as well.

We stress that this is only a hypothetical illustration of a multiverse analysis. Our multiverse is only a subset of a larger multiverse of possible data-analytic choices, and we can not rule out that Durante et al.'s (2013) actual multiverse might have been different.

Constructing the data multiverse. The first step involves listing the different reasonable choices during each step of data processing. Box 1 summarizes five arbitrary choices in data processing, both in Study 1 and 2, and for each arbitrary choice, the alternative reasonable options we will consider.

Box 1. Processing choices

1. Assessment of fertility (F) (high vs low)
   (a) F1: high = cycle days 7–14; low = cycle days 17–25
   (b) F2: high = cycle days 6–14; low = cycle days 17–27
   (c) F3: high = cycle days 9–17; low = cycle days 18–25
   (d) F4: high = cycle days 8–14; low = cycle days 1–7 and 15–28
   (e) F5: high = cycle days 9–17; low = cycle days 1–8 and 18–28
2. Next menstrual onset (NMO)
   (a) NMO1: reported start date previous menstrual onset + computed cycle length
   (b) NMO2: reported start date previous menstrual onset + reported cycle length
   (c) NMO3: reported estimate of next menstrual onset
3. Assessment of relationship status (R) (single vs relationship)
   (a) R1: single = response options 1 and 2; relationship = response options 3 and 4
   (b) R2: single = response option 1; relationship = response options 2, 3, and 4
   (c) R3: single = response option 1; relationship = response options 3 and 4
4. Exclusion of women based on cycle length (ECL)
   (a) ECL1: no exclusion based on cycle length
   (b) ECL2: exclusion of participants with computed cycle length < 25 or > 35 days
   (c) ECL3: exclusion of participants with reported cycle length < 25 or > 35 days
5. Exclusion of women based on certainty ratings of start dates of two previous menstrual periods (EC)
   (a) EC1: no exclusion based on certainty ratings
   (b) EC2: exclusion of participants who are not certain about at least one start date (i.e., sure < 6)

Fertility. First, the classification of women into a high or low fertility group based on cycle day can be done using several reasonable alternatives: assigning women with cycle days 6–14 to the high fertility group and women with cycle days 17–27 to the low fertility group (e.g., Durante, Griskevicius, Hill, Perilloux, & Li, 2011); days 9–17 for high fertility and 18–25 for low fertility (Durante, Griskevicius, Simpson, Cantú, & Li, 2012); days 8–14 for high fertility and 1–7 and 15–28 for low fertility (Durante, Griskevicius, Cantú, & Simpson, 2014); and days 9–17 for high fertility and 1–8 and 18–28 for low fertility (Durante & Arsena, 2015).

Second, there are different reasonable alternatives for estimating a woman's next menstrual onset, which is an intermediate step in determining cycle day. A reasonable way to estimate next menstrual onset is based on the women's reported estimate of their typical cycle length (e.g., Thornhill & Gangestad, 1999). Another reasonable strategy for determining the onset of the next period involves using the self-reported expected start date of the next menstrual period (e.g., Haselton & Miller, 2006).[2]

Relationship status. There are at least two reasonable alternative options to the dichotomization of women's relationship status, stemming from the ambiguous nature of response option (2) *dating or involved with only one partner*. This option can cover both single women (*dating*) or women in relationships (*involved with only one partner*). Thus, women who select this response could reasonably be classified as being either in committed relationships or as being single. A third option involves discarding participants who select this ambiguous response option, and only classifying participants selecting option (1) as single women, and participants selecting option (3) or (4) as women in relationships.

Exclusion criteria. First, it is not unreasonable to exclude participants with irregular cycle lengths in fertility research. This could amount to only including women with cycle lengths 25 to 35 (e.g., Durante et al., 2012). This exclusion criterion can be instantiated in two reasonable ways, using either a woman's computed cycle length or a woman's self-reported typical cycle length.

[2]The fact that typical cycle length and the expected start date of the next period were collected by Durante et al. (2013) suggests that they considered this option at least somewhat reasonable.

Second, another justifiable exclusion criterion concerns women's reported certainty ratings of the start dates of their last two menstrual periods. It is reasonable to exclude participants who were not sufficiently confident about their report, and to consider only data from participants with a certainty rating above the midpoint for both dates (e.g., Durante, Arsena, & Griskevicius, 2014).

Based on this tabulation of choices, the multiverse of data sets is constructed by considering all combinations of reasonable choices in data processing, and deriving a data set for each of the different choice combinations. In Study 1, there are $5 \times 2 \times 3 \times 3 \times 2 = 180$ choice combinations (see Box 1; for the estimation of next menstrual onset, option (c) cannot be applied to Study 1, as the expected start date of the next menstrual period was not collected in this study). Some of the choice combinations are inconsistent: when participants are excluded based on reported or computed cycle length, we do not consider next menstrual onset based on computed or reported cycle length, respectively. After excluding these inconsistent combinations, we are left with $180 - 2 \times (5 \times 1 \times 3 \times 1 \times 2) = 120$ choice combinations. Similarly, in Study 2, there are $5 \times 3 \times 3 \times 3 \times 2 = 270$ choice combinations, but after excluding inconsistent combinations, $270 - 2 \times (5 \times 1 \times 3 \times 1 \times 2) = 210$ choice combinations remain.

Deriving the multiverse of statistical results. After constructing the data multiverse, the analysis of interest (in this case, an ANOVA or a logistic regression) is performed across all the alternatively constructed data sets.[3] The results are shown in Panels A-F of Figure 1, each showing a histogram of the p-values of the fertility $\times$ relationship interaction effect.

For two variables —religiosity in Study 1 (Panel A) and fiscal political attitudes (Panel C) — the multiverse analysis reveals a near-uniform distribution, indicating that the p-value for the interaction effect between fertility and relationship varies widely across the multiverse. For religiosity, seven out of the 120 choice combinations lead to a significant interaction effect, whereas the remaining 94% lead to p-values ranging from 0.05

[3]Due to coding errors in the data, there were some missing data (see online Supplemental Materials for details). In our analyses, incomplete cases are discarded.

to 1. For fiscal political attitudes, 8% of the 210 choice combinations lead to a significant interaction ($p < 0.05$), whereas the remaining choice combinations lead to p-values across the entire range from 0.05 to 1.

For the remaining four variables, roughly half of the choice combinations lead to a significant interaction effect. In particular, for religiosity in Study 2 (Panel B), 88 out of the 210 choice combinations (42%) lead to a p-value smaller than 0.05. Regarding social political attitudes (Panel D), 49% of the p-values is smaller than 0.05. Finally, 46% and 57% of the p-values are smaller than 0.05 for voting (Panel E) and donation (Panel F) preferences, respectively. In these cases, it is informative to display the multiverse in greater detail by showing which constellation of choices corresponds to which statistical result. This allows to identify the key choices in data processing that are most consequential in the fluctuation of the statistical results.

Such a closer inspection is provided in Figure 2, showing a grid of p-values for each of these four variables. In each panel, the cells show the different p-values that can be obtained across all choice combinations for data processing. Depending on whether the p-value is smaller or larger than the α-level, the cells are colored gray or white, respectively. For religiosity in Study 2 (Panel A), most data sets constructed under the second option for relationship assessment (R2) yield a non-significant interaction effect. The first and third options (R1 and R3) consistently lead to a significant interaction effect in combination with the first and second option for fertility assessment (F1 and F2) and to a non-significant interaction effect in combination with F5, whereas data sets constructed under R1 or R3 in combination with F3 or F4 lead to more fluctuating conclusions, depending on the other choices for data processing. The different exclusion criteria and cycle day estimation options do not seem to have a large impact on fluctuation in the statistical conclusion. For social political attitudes (Panel B), the statistical conclusion is highly robust for the first and second option for relationship status assessment (significant for R1 and non-significant for R2). Using the third option for relationship status assessment (R3) leads to more fluctuation, depending on the choices for the other processing steps. Finally, for voting and

donation preferences (Panels C and D, respectively), it is hard to extract a consistent pattern of fluctuation across the different choice combinations. It seems that all arbitrary choices for data processing can have an impact on whether the obtained data set will lead to a significant or a non-significant outcome.

Discussion

Converting a set of observations into a data set that is suitable for statistical analysis usually requires active data construction. If there are strong grounds to justify the necessary processing steps, the raw observations uniquely translate into a single data set for analysis. In many cases, however, the intermediate processing steps involve arbitrary or, as Leamer (1983) calls them, whimsical, choices, so that the single set of observations does not uniquely lead to a single data set. Rather, it spawns a multiverse of data sets, and thus does not admit a unique conclusion. Yet, researchers often analyze, or at least report, only one (or a few) data sets that are the result of one (or a few) outcomes of this chain of arbitrary choices. To the extent their single data set is based on arbitrary processing choices, their statistical result is arbitrary. We suggest that, if several processing choices are defensible, researchers should perform a multiverse analysis instead of a single data set analysis. This involves considering *all* different reasonable data sets, except those arising under inconsistent choice combinations. A multiverse analysis is a way to avoid or at least reduce the problem of selective reporting, by making the fragility or robustness of the results transparent, and helps the identification of the most consequential choices.

In our demonstration, we started from a single set of raw data and performed both a single data set analysis as well as a multiverse analysis. Comparison of both types of analysis highlights the dramatic impact of going beyond an $N = 1$ sample from the multiverse. For religiosity in Study 1, the arbitrary data processing choices made in the single data set analysis led to a significant result. Placing this significant result in the multiverse of statistical results illustrates the risk of running a single data set analysis. The multiverse analysis revealed that almost all choice combinations for data processing lead to large p-values. As such non-statistically-significant findings in general represent nothing

more than uncertainty, this pattern of results clearly raises serious questions regarding the finding on the effect of fertility found in the single data set analysis, and should make a researcher hesitant to trust the single data set finding. The effect of fertility on religion seems too sensitive to arbitrary choices, and thus too fragile to be taken seriously.

For most other variables, there was considerable ambiguity: The interaction seemed to be significant across about half of the arbitrary choice combinations. In these cases, the conclusion on the effect of fertility strongly depends on the evaluation of the different processing options. Both the authors performing the multiverse analysis and the readers of the research can construct arguments in favor or against certain choices, and the validity of these arguments will help drawing the conclusion. For example, if additional information suggests that the fifth option of assessing fertility is clearly superior, then Panel A of Figure 2 indicates that there is little evidence for an effect of fertility on religiosity in Study 2. On the other hand, if additional information suggests that the second option of assessing fertility is clearly superior, then most choice combinations lead to a significant interaction effect.

If no strong arguments can be made for certain choices, we are left with many branches of the multiverse that have large p-values. In these cases, the only reasonable conclusion on the effect of fertility is that there is considerable scientific uncertainty. One should reserve judgment and acknowledge that the data are not strong enough to draw a conclusion on the effect of fertility. The real conclusion of the multiverse analysis is that there is a gaping hole in theory or in measurement, and that researchers interested in studying the effect of fertility should work hard to *deflate* the multiverse. The multiverse analysis gives useful directions in this regard.

In general, deflating the multiverse involves developing a better and more complete theorizing of the constructs of interest, and improving their measurement. Both routes for deflating the multiverse are illustrated in our case study. A first approach involves improving the experimental material and design. For example, the detailed multiverse examination shown in Figure 2 revealed that a lot of fluctuation hinged on the different

choices for relationship status assessment. Thus, apparently, this type of research could benefit from a better way of assessing relationship status. Looking at the alternative options for assessing relationship status, it seems that the ambiguous response option (2) in the relationship status question could be formulated more precisely, so that relationship status assessment is no longer an arbitrary choice. This would have narrowed down the multiverses to 40 and 70 choice combinations in Study 1 and 2, respectively.

A second approach for deflating the multiverse involves developing more complete and more precise theory, in such a way that some options are theoretically superior than others, and should be preferred when constructing data sets. For example, a great deal of variation in the results appeared to be driven by the different options for assessing fertility. Clearly, for this type of research, developing and applying a more precise way of assessing fertility should become a research priority. The availability of different reasonable options for estimating next menstrual onset or for classifying women into a high or low fertility group based on their cycle day stems from the fact that a precise theoretical foundation is lacking (Harris, 2013). The development of elaborated theories concerning these issues would narrow down the number of alternative options and deflate fluctuation. Recently, Gangestad et al. (2016) have recommended assessing fertility based on the detection of surges in luteinizing hormone, ideally in a within-subjects design. It is of note that this alternative strategy of assessing fertility was used in several papers by Durante (e.g., Durante et al., 2011, 2012).

Pre-registration (e.g., Chambers, 2013; Wagenmakers et al., 2012) or blind analysis (e.g., MacCoun & Perlmutter, 2015) are not useful strategies for deflating the multiverse. By pre-registering a study, all analytical choices—including the arbitrary ones—are made ahead of time, before collecting the data. Similarly, in a blind analysis, all analytical choices are made using a data set with temporarily removed data labels. The appeal of both strategies is that the choices cannot be made conditional on the (real) data. However, the considered results are still just the results given one choice combination, albeit pre-registered or blindly made, and their robustness across other reasonable choice

alternatives remains hidden from view. Thus, pre-registration or blind analysis do not preclude a multiverse analysis, as they do not annihilate the arbitrariness in data preparation.

As is evident from our demonstration, a multiverse analysis is highly context-specific and inherently subjective. Listing the alternative options for data construction requires judgement about which options can be considered reasonable, and will typically depend on the experimental design, the research question and the researchers performing the research. Whereas this subjectivity may seem undesirable, presenting results given only a single combination of reasonable options is much more misleading; indeed one of the sources of the current crisis in scientific replication is that researchers traditionally have taken p-values at face value without considering the multiplicity of choices in data construction. A related point is that not all options are necessarily exactly interchangeable. Some options might seem better than others, at least for some researchers. If such is the case, this knowledge can be used to construct arguments for interpreting results such as those shown in Figure 2. However, a multiverse analysis should involve all plausible construction alternatives, not just the most plausible ones. Only when one choice is clearly and unambiguously the most appropriate one, variation across this choice is uninformative.

The richness of possibilities for different data processing choices present in the raw data made the case study exceptionally suitable for the demonstration of a multiverse analysis. We do not expect that all multiverses will consist of such a numerous amount of data sets. The fact that more typical multiverses will tend to be smaller does not make a multiverse analysis less necessary. Even when confronted with only one arbitrary data processing choice, researchers should be transparent about it and reveal the sensitivity of the result to this choice.

We aimed to show the multiverse analysis we think Durante et al. (2013) could have done, instead of their single data set analysis. Since their single data set analysis used p-values, our demonstration of the multiverse analysis did too. There is, however, nothing inherently special about p-values from a multiverse perspective. Increasing the

transparency in reporting through a multiverse analysis is valuable, regardless of the inferential framework (frequentist or Bayesian), and regardless of the specific way uncertainty is quantified — a *p*-value, an effect size, a confidence (Cumming, 2013) or credibility (Kruschke, 2010) interval or a Bayes factor (Morey & Rouder, 2011).

The primary goal of a multiverse analysis is to enhance research transparency. Unlike, for example, a *p*-curve analysis (Simonsohn, Nelson, & Simmons, 2014), it is not a formal test of questionable research practices such as selective reporting, or a method to estimate the strength of the evidence for an effect. The multiverse analysis does not produce a single value summarizing the evidential value of the data, nor does it imply a threshold for an effect to reach to be declared robustly significant. Nevertheless, one might try to summarize the multiverse analysis more formally. One reasonable first step is to simply average the *p*-values in the multiverse, in this case averaging all the numbers displayed in Figure 1 or 2. This mean value can be considered as the *p*-value of a hypothetical pre-registered study with conditions chosen at random among the possibilities in the multiverse and seems like a fair measurement in a setting where all of the possible data processing choices seem plausible (as in the example presented here, where the different options are drawn from other papers in the relevant literature).

We have focused on the multiverse of statistical results originating from the data multiverse, i.e., the different reasonable choices in data processing. We have ignored arbitrary choices occurring at the level of statistical models used in data analysis. Choices at the model level include choosing among different statistical approaches (e.g., a repeated-measures ANOVA or a hierarchical linear model), focusing on main effects or interactions, approximating errors normally, assuming random effects, assuming homoscedasticity, assuming linearity, choosing between a parametric and a non-parametric approach, and so on. One specific analysis thus corresponds to a single sample from a *model multiverse*. If the choice for a single model specification out of the model multiverse can not be justified, a model multiverse analysis can be performed to reveal the effect of this arbitrary choice on the statistical result.

A compelling example of such a model multiverse is provided in Patel, Burford, and Ioannidis (2015), focusing on the choices in deciding which predictors and covariates to include. Such a model multiverse analysis is related to *perturbation analysis* (Geisser, 1993) and to *sensitivity analysis* in economics (e.g., Leamer, 1985) and in Bayesian statistics (e.g., Kass & Raftery, 1995), all of which involve investigating the influence of arbitrary modeling assumptions on the results, such as using a normal error distribution or a t-distribution, the inclusion of different variables, or using different reasonable priors. In a more complete analysis, the multiverse of data sets could be crossed with the multiverse of models to further reveal the multiverse of statistical results. Thus, the multiverse analysis as demonstrated here is a minimal attempt at establishing a range of analyses consistent with a research hypothesis. To the extent that there are arbitrary choices not only in data preparation but also in data analysis or model choice, this motivates encompassing analyses of multiple predictors, interactions, or outcomes in a hierarchical model so as to reduce problems of multiple comparisons (Gelman, Hill, & Yajima, 2012).

Our demonstration of the multiverse analysis should serve as a cautionary tale. We hope it raises awareness that, in the light of the multiverse of statistical results, isolating a single statistical result stemming from a chain of arbitrary choices can be highly misleading. Readers of research need to get a sense of sensitivity of conclusions to arbitrary decisions in data preparation, and thus of the fragility or robustness of a claimed effect. We believe that it should become standard practice to go beyond a single data set analysis, and to acknowledge the multiverse of statistical results. Admittedly, performing a multiverse analysis will often be difficult, and to a large extent subjective, but that does not change the fact that it is a necessary step for increasing transparency.

References

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., ... others (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*, 108–119.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*, 543–554.

Chambers, C. D. (2013). Registered reports: A new publishing initiative at cortex. *Cortex*, *49*(3), 609-610.

Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Durante, K. M., & Arsena, A. R. (2015). Playing the field: The effect of fertility on women's desire for variety. *Journal of Consumer Research*, *41*, 1372–1391.

Durante, K. M., Arsena, A. R., & Griskevicius, V. (2014). Fertility can have different effects on single and nonsingle women: Reply to Harris and Mickes (2014). *Psychological Science*, *25*, 1150–1152.

Durante, K. M., Griskevicius, V., Cantú, S. M., & Simpson, J. A. (2014). Money, status, and the ovulatory cycle. *Journal of Marketing Research*, *51*, 27–39.

Durante, K. M., Griskevicius, V., Hill, S. E., Perilloux, C., & Li, N. P. (2011). Ovulation, female competition, and product choice: Hormonal influences on consumer behavior. *Journal of Consumer Research*, *37*, 921–934.

Durante, K. M., Griskevicius, V., Simpson, J. A., Cantú, S. M., & Li, N. P. (2012). Ovulation leads women to perceive sexy cads as good dads. *Journal of Personality and Social Psychology*, *103*, 292–305.

Durante, K. M., Rae, A., & Griskevicius, V. (2013). The fluctuating female vote: Politics, religion, and the ovulatory cycle. *Psychological Science*, *24*, 1007–1016.

Gangestad, S. W., Haselton, M. G., Welling, L. L., Gildersleeve, K., Pillsworth, E. G., Burriss, R. P., ... Puts, D. A. (2016, mar). How valid are assessments of conception probability in ovulatory cycle research? Evaluations, recommendations, and

theoretical implications. *Evolution and Human Behavior*, *37*(2), 85–96.

Geisser, S. (1993). *Predictive inference: An introduction*. New York: Chapman & Hall.

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, *5*(2), 189–211.

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

Harris, C. R. (2013). Shifts in masculinity preferences across the menstrual cycle: Still not there. *Sex Roles*, *69*, 507–515.

Harris, C. R., Pashler, H., & Mickes, L. (2014). Elastic analysis procedures: An incurable (but preventable) problem in the fertility effect literature. Comment on Gildersleeve, Haselton, and Fales (2014). *Psychological Bulletin*, *140*, 1260–1264.

Haselton, M. G., & Miller, G. F. (2006). Women's fertility across the cycle increases the short-term attractiveness of creative intelligence. *Human Nature*, *17*, 50-73.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524–532.

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences*, *110*, 19313–19317.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, *14*, 293–300.

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

Leamer, E. E. (1985). Sensitivity analyses would help. *The American Economic Review*, *75*(3), 308–313.

LeBel, E. P., Campbell, L., & Loving, T. J. (in press). Benefits of open and high-powered

research outweigh costs. *Journal of Personality and Social Psychology*.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*(1), 19.

MacCoun, R., & Perlmutter, S. (2015). Blind analysis: Hide results to seek the truth. *Nature*, *526*(7572), 187–189.

Morey, R. D., Chambers, C. D., Etchells, P. J., Harris, C. R., Hoekstra, R., Lakens, D., ... others (2016). The peer reviewers' openness initiative: incentivizing open research practices through peer review. *Royal Society Open Science*, *3*(1), 150547.

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*, 406–419.

Nosek, B. A., Alter, G., Banks, G., Borsboom, D., Bowman, S., Breckler, S., ... others (2015). Promoting an open research culture: Author guidelines for journals could help to promote transparency, openness, and reproducibility. *Science (New York, NY)*, *348*(6242), 1422.

Nosek, B. A., & Bar-Anan, Y. (2012). Scientific utopia: I. Opening scientific communication. *Psychological Inquiry*, *23*, 217–243.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Patel, C. J., Burford, B., & Ioannidis, J. P. (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.

Ramsey, F., & Schafer, D. (2012). *The statistical sleuth: A course in methods of data analysis* (3rd ed.). Stanford: Cengage Learning.

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*, 1359–1366.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 word solution. *Available at SSRN 2160588*.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*, 534–547.

Thornhill, R., & Gangestad, S. W. (1999). The scent of symmetry: A human sex pheromone that signals fitness? *Evolution and Human Behavior*, *20*, 175–201.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & Van der Maas, H. L. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*(3), 426–432.

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*, 632–638.

Yong, E. (2012). In the wake of high profile controversies, psychologists are facing up to problems with replication. *Nature*, *483*, 298–300.

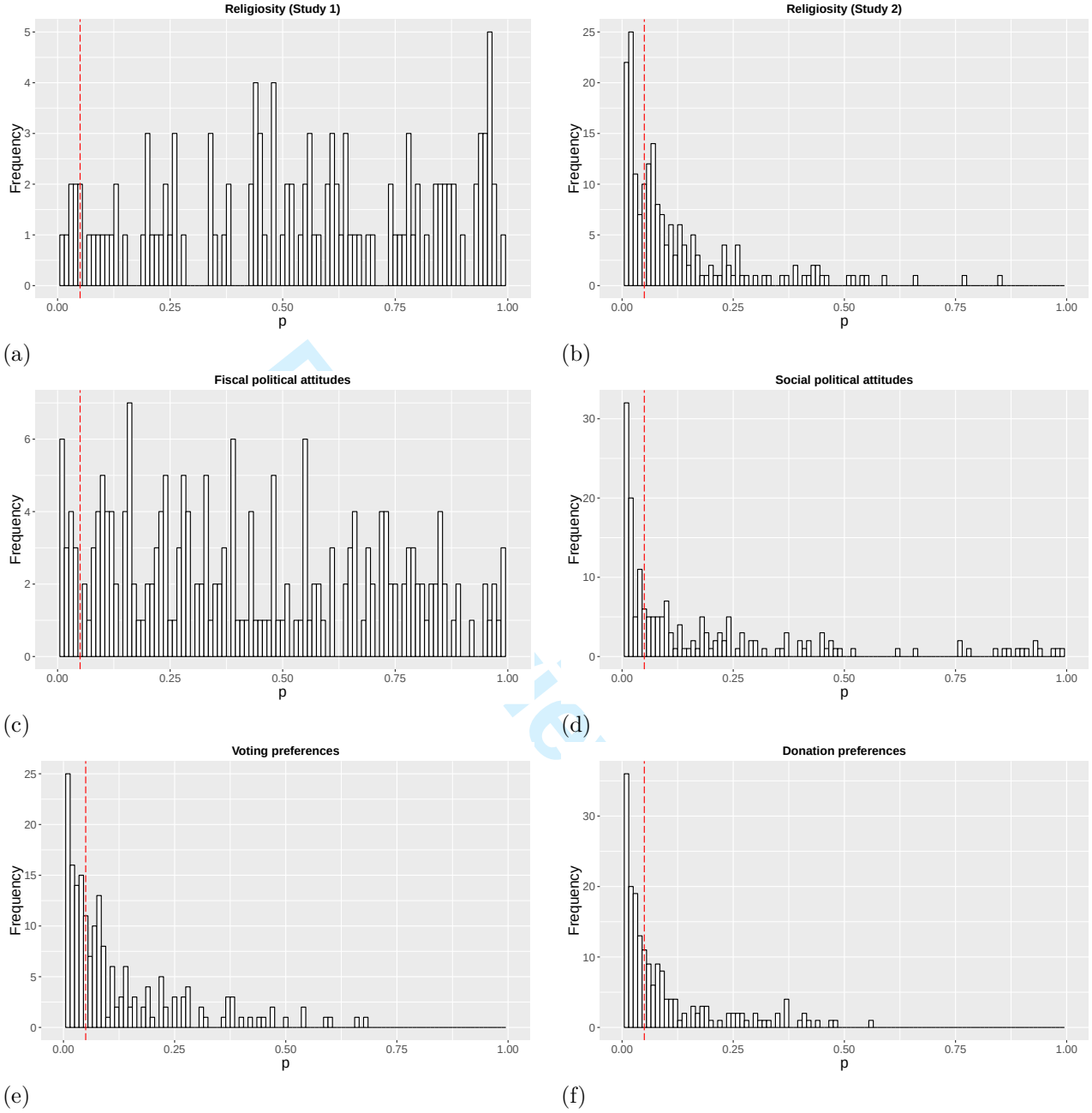
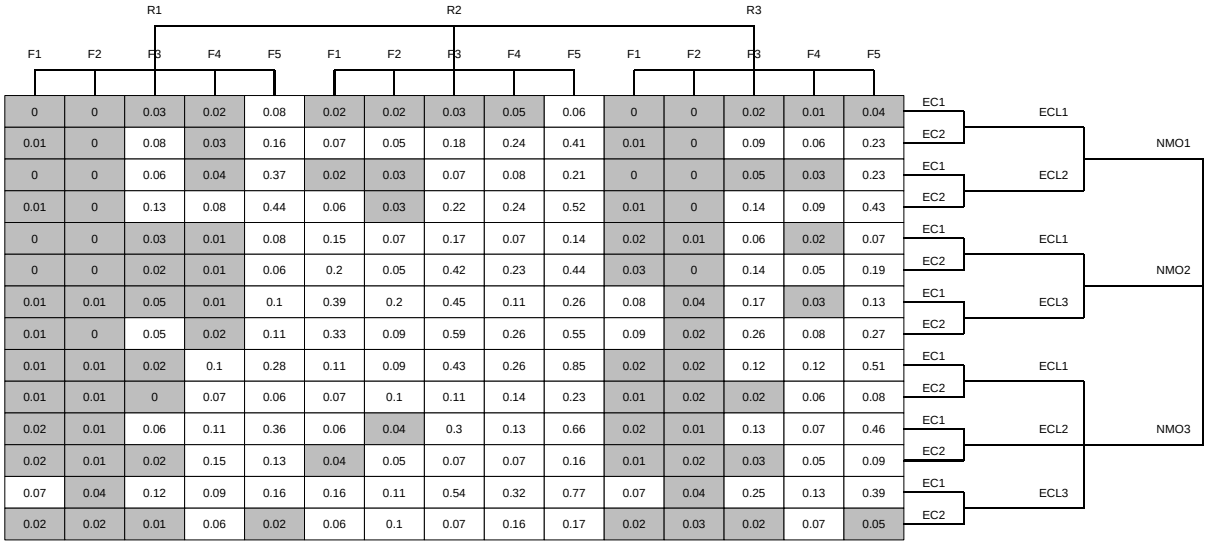


Figure 1. Histogram of p-values for the interaction effect between fertility and relationship status on religiosity for the multiverse of 120 data sets in Study 1 and 210 data sets in Study 2 (panels A and B), on fiscal and social political attitudes for the multiverse of 210 data sets in Study 2 (panels C and D), and on voting and donation preferences for the multiverse of 210 data sets in Study 2 (panels E and F). The dashed line indicates $p = 0.05$.

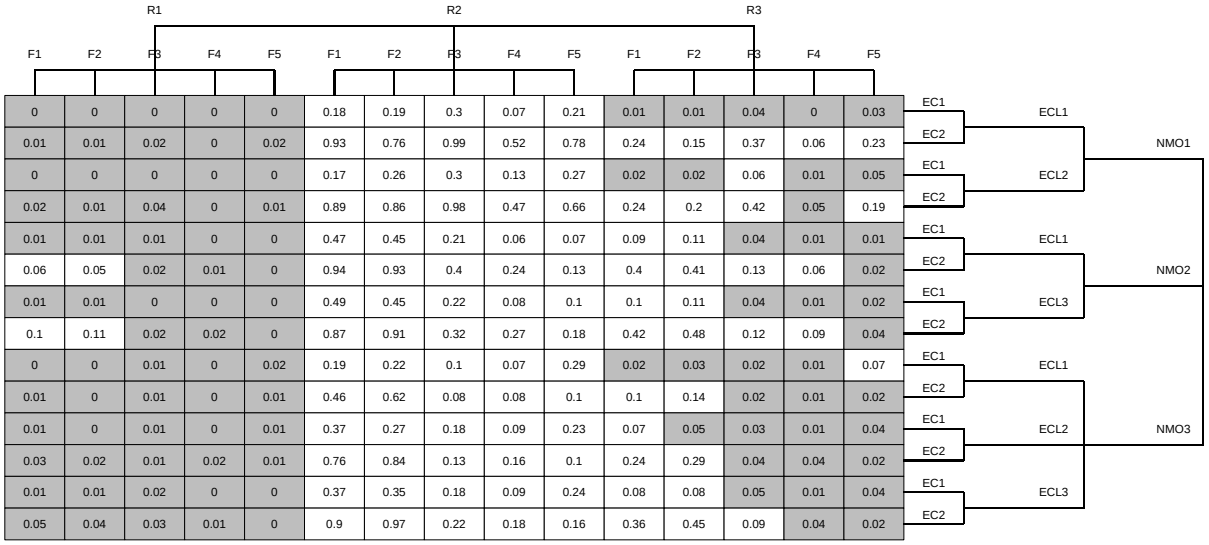
Religiosity (Study 2)



0	0	0.03	0.02	0.08	0.02	0.02	0.03	0.05	0.06	0	0	0.02	0.01	0.04
0.01	0	0.08	0.03	0.16	0.07	0.05	0.18	0.24	0.41	0.01	0	0.09	0.06	0.23
0	0	0.06	0.04	0.37	0.02	0.03	0.07	0.08	0.21	0	0	0.05	0.03	0.23
0.01	0	0.13	0.08	0.44	0.06	0.03	0.22	0.24	0.52	0.01	0	0.14	0.09	0.43
0	0	0.03	0.01	0.08	0.15	0.07	0.17	0.07	0.14	0.02	0.01	0.06	0.02	0.07
0	0	0.02	0.01	0.06	0.2	0.05	0.42	0.23	0.44	0.03	0	0.14	0.05	0.19
0.01	0.01	0.05	0.01	0.1	0.39	0.2	0.45	0.11	0.26	0.08	0.04	0.17	0.03	0.13
0.01	0	0.05	0.02	0.11	0.33	0.09	0.59	0.26	0.55	0.09	0.02	0.26	0.08	0.27
0.01	0.01	0.02	0.1	0.28	0.11	0.09	0.43	0.26	0.85	0.02	0.02	0.12	0.12	0.51
0.01	0.01	0	0.07	0.06	0.07	0.1	0.11	0.14	0.23	0.01	0.02	0.02	0.06	0.08
0.02	0.01	0.06	0.11	0.36	0.06	0.04	0.3	0.13	0.66	0.02	0.01	0.13	0.07	0.46
0.02	0.01	0.02	0.15	0.13	0.04	0.05	0.07	0.07	0.16	0.01	0.02	0.03	0.05	0.09
0.07	0.04	0.12	0.09	0.16	0.16	0.11	0.54	0.32	0.77	0.07	0.04	0.25	0.13	0.39
0.02	0.02	0.01	0.06	0.02	0.06	0.1	0.07	0.16	0.17	0.02	0.03	0.02	0.07	0.05

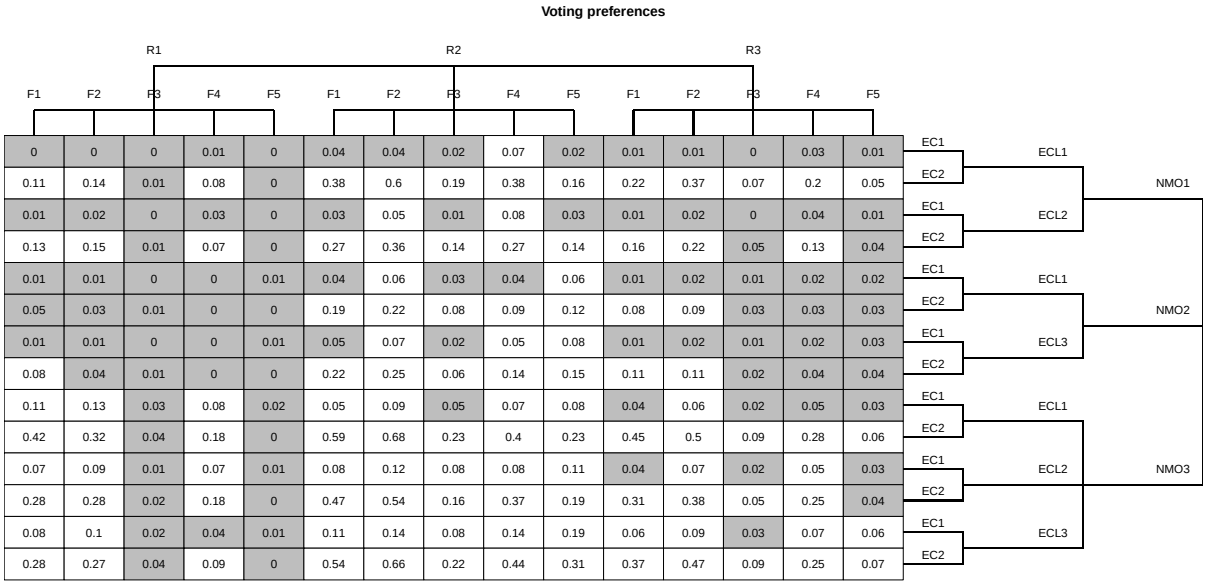
(a)

Social political attitudes

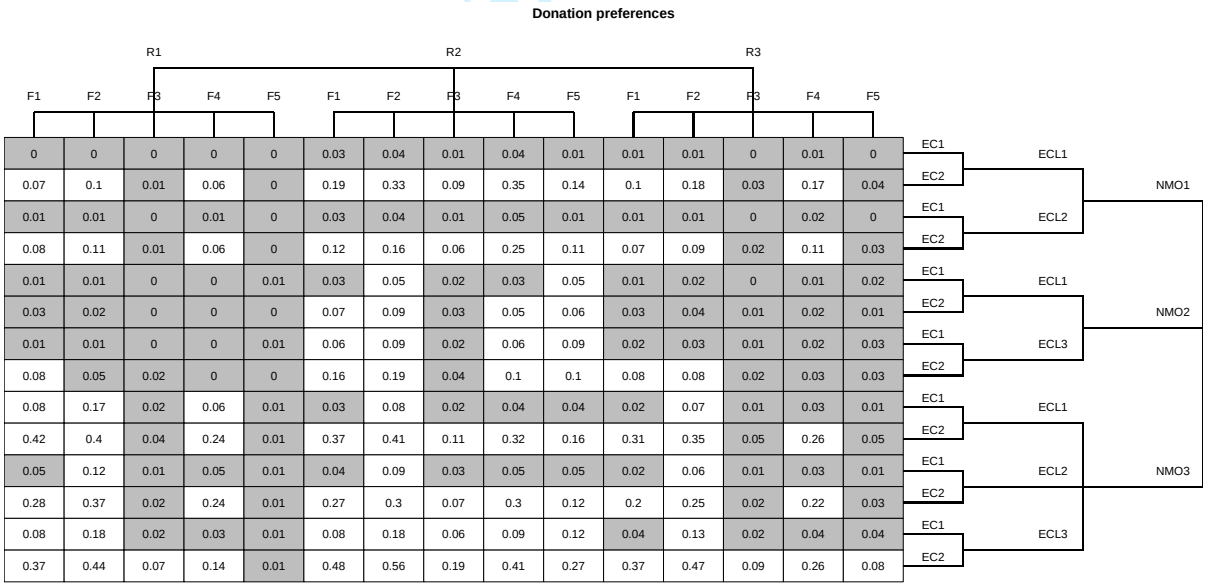


0	0	0	0	0	0.18	0.19	0.3	0.07	0.21	0.01	0.01	0.04	0	0.03
0.01	0.01	0.02	0	0.02	0.93	0.76	0.99	0.52	0.78	0.24	0.15	0.37	0.06	0.23
0	0	0	0	0	0.17	0.26	0.3	0.13	0.27	0.02	0.02	0.06	0.01	0.05
0.02	0.01	0.04	0	0.01	0.89	0.86	0.98	0.47	0.66	0.24	0.2	0.42	0.05	0.19
0.01	0.01	0.01	0	0	0.47	0.45	0.21	0.06	0.07	0.09	0.11	0.04	0.01	0.01
0.06	0.05	0.02	0.01	0	0.94	0.93	0.4	0.24	0.13	0.4	0.41	0.13	0.06	0.02
0.01	0.01	0	0	0	0.49	0.45	0.22	0.08	0.1	0.1	0.11	0.04	0.01	0.02
0.1	0.11	0.02	0.02	0	0.87	0.91	0.32	0.27	0.18	0.42	0.48	0.12	0.09	0.04
0	0	0.01	0	0.02	0.19	0.22	0.1	0.07	0.29	0.02	0.03	0.02	0.01	0.07
0.01	0	0.01	0	0.01	0.46	0.62	0.08	0.08	0.1	0.1	0.14	0.02	0.01	0.02
0.01	0	0.01	0	0.01	0.37	0.27	0.18	0.09	0.23	0.07	0.05	0.03	0.01	0.04
0.03	0.02	0.01	0.02	0.01	0.76	0.84	0.13	0.16	0.1	0.24	0.29	0.04	0.04	0.02
0.01	0.01	0.02	0	0	0.37	0.35	0.18	0.09	0.24	0.08	0.08	0.05	0.01	0.04
0.05	0.04	0.03	0.01	0	0.9	0.97	0.22	0.18	0.16	0.36	0.45	0.09	0.04	0.02

(b)



(c)



(d)

Figure 2. Visualization of the multiverse of *p*-values of the fertility × relationship status interaction on religiosity (Panel A), on social political attitudes (Panel B), on voting preferences (Panel C) and on donation preferences (Panel D) in Study 2, showing the dependence of the results on data processing choices. See Box 1 for an explanation of the acronyms.