Under peer review at Meta-Psychology. This is a revised version of the submitted paper. The editorial process, including decision letters and reviews, can be accessed at: http://tinyurl.com/mp-submissions

Anyone can contribute with Open Peer Review through hypothes.is directly on this preprint.

Multiverse analyses in the classroom

Tom Heyman[1] & Wolf Vanpaemel[2]

[1] Leiden University

[2] KU Leuven

Author Note

Correspondence concerning this article should be addressed to Tom Heyman, Methodology and Statistics Unit, Institute of Psychology, Leiden University, Wassenaarseweg 52, 2333 AK Leiden, The Netherlands. E-mail: t.d.p.heyman@fsw.leidenuniv.nl

Abstract

Most empirical papers in psychology involve statistical analyses performed on a new or existing dataset. Sometimes the robustness of a finding is demonstrated via data-analytical triangulation (e.g., obtaining comparable outcomes across different operationalizations of the dependent variable), but systematically considering the plethora of alternative analysis pathways is rather uncommon. However, researchers increasingly recognize the importance of establishing the robustness of a finding. The latter can be accomplished through a so-called multiverse analysis, which involves methodically examining the arbitrary choices pertaining to data processing and/or model building. In the present paper, we describe how the multiverse approach can be implemented in student research projects within psychology programs, drawing on our personal experience as instructors. Embedding a multiverse project in students' curricula addresses an important scientific need, as studies examining the robustness or fragility of phenomena are largely lacking in psychology. Additionally, it offers students an ideal opportunity to put various statistical methods into practice, thereby also raising awareness about the abundance and consequences of arbitrary decisions in data-analytic processing. An attractive practical feature is that one can reuse existing datasets, which proves especially useful when resources are limited, or when circumstances such as the COVID-19 lockdown measures restrict data collection possibilities.

Keywords: multiverse analysis; robustness; education; pedagogy; open science

Word count: 6173

An important part of many psychology students' (under)graduate programs are research-methods classes in which students are asked to complete their own (small-scale) research project (e.g., Kierniesky, 2005). Typically, the goal is to run through the entire empirical cycle, thus putting knowledge gained from previous theory-focused courses into practice. However, this can be quite challenging, as time and resources are often limited in such projects. As a consequence, students and instructors might (begrudgingly) take shortcuts, resulting in ill-designed or underpowered studies, poorly-motivated research questions, sloppy measurement practices, and so on. Perhaps the most devastating consequence of this approach is that students could come away with a wrong impression of what psychological research entails, and it might even instill bad habits in prospective researchers.

In the present paper, we suggest an alternative implementation of research-methods classes that addresses these concerns. In particular, we propose that completing a multiverse analysis project as part of such research methods classes has several important benefits. First, we explain what a multiverse analysis entails (see Steegen et al., 2016). Then, we describe the two main ingredients of a multiverse-in-the-classroom project: a suitable dataset and a solid (meta-)scientific background. Next, we give a worked example of such a project, based on our personal experience as instructors. Finally, we discuss the benefits and challenges of the multiverse-in-the-classroom.

What is a Multiverse Analysis?

Most empirical papers in psychology involve some kind of data analysis. Typically, there is no unique path from the raw data to the eventual conclusions of a paper. Researchers need to make a number of decisions along the way, such as whether and how to deal with outliers and missing data, whether and how to transform variables, and so on. In some cases, theoretical considerations provide a clear solution to such questions, yet, at times, researchers

have little to go on, so they turn to their gut feeling, lab habits, or field-specific standards, which are often poorly motivated. As a result, when processing and analyzing empirical data, researchers regularly face certain choices that are arbitrary in nature. These researcher degrees of freedom (Simmons et al., 2011) lead to a garden of forking paths (Gelman & Loken, 2014).

As an example, suppose that, for a given dataset, a researcher identifies four plausible ways to deal with outliers, three approaches to handle missing data, and two reasonable options to transform a particular variable. Assuming all combinations are sensible, this would lead to 4*3*2 = 24 unique paths, each with its own outcome (see also Bishop, 2016). However, researchers usually report the results for just one or a few of these paths, by picking only one or a few options out of the pool of plausible alternatives (e.g., deleting observations 2.5 standard deviations above the mean, listwise deletion when encountering missing values, and log-transforming a positively skewed variable; see also Elson, 2016, for a practical illustration).

In contrast, the idea of a multiverse analysis (Steegen et al., 2016) is to explore and report on a wide array of imaginable (combinations of) reasonable alternatives, each of which providing an answer to the same research question. By explicitly considering the results of several reasonable analyses, a multiverse analysis can give an idea about the robustness or fragility of a certain finding, and might even point to moderators of the effect in question (i.e., key choices regarding data processing and/or analysis that the conclusion depends on). A multiverse analysis can be applied to newly collected data (e.g., Kalokerinos et al., 2019), but also retrospectively using existing data (e.g., Moors & Hesselmann, 2019). For instance, Credé and Phillips (2017) conducted a multiverse analysis on data from Carney and colleagues (2010) examining the power pose effect, which is the (controversial) finding that holding a high-power body pose affects hormone levels. Their multiverse analysis revealed

that most alternative pathways yielded null effects, whereas the original single-pathway analysis produced a significant effect.

The importance of a multiverse analysis is also nicely illustrated by the study of Silberzahn and colleagues (2018), in which 29 research teams independently examined whether referees in soccer are more likely to give red cards to players with a darker skin tone compared to light-skin-toned players. All teams used the same dataset to answer this question, yet the conclusions varied considerably: 20 out of 29 teams (69%) found a positive relation (i.e., dark-skin-toned-players tended to receive more red cards), whereas 9 teams obtained a null effect, which was even numerically negative in two cases. These results underscore that there are often several ways to process and analyze a given dataset, and that picking a single pathway might be deceiving, which is why conducting a multiverse analysis can be very informative.

Ideas similar to that of a multiverse analysis have been proposed under different names, such as specification curve analysis (Simonsohn et al., 2020), vibration of effects analysis (Patel et al., 2015), multimodel analysis (Young & Holsteen, 2017), and the many analysts approach by Silberzahn et al. (2018) discussed above (though, in contrast with the other approaches, in the latter approach the different choices are distributed over research teams rather than being performed by the same team). Multiverse-style analyses are increasingly being recognized as providing crucial information, and researchers have also proposed various extensions and refinements. For instance, multiverse analyses have been applied in the context of meta-analyses (Voracek, Kossmeier, & Tran, 2019), suggested as an approach to deal with different random effect structures of multi-level models (Harder, 2020), and used in combination with so-called explorable explanations allowing readers of a paper to dynamically move through the multiverse (Dragicevic et al., 2019). In addition, Liu and colleagues (2020) recently developed a programming tool called Boba, which helps

researchers to conduct and visualize multiverse analyses, whereas others have developed specific R packages to facilitate multiverse analyses (e.g., Masur & Scharkow, 2019; Sarma & Kay, 2019).

Teaching Multiverse Analyses

The key message of this paper is that multiverse analyses are ideally suited to be included in laboratory or research-methods classes. In line with its general theme, there are a multitude of ways in which multiverse analyses can be incorporated in research-methods classes, taking into account the available time, place in the curriculum, and learning objectives. Yet, they all require two essential ingredients: a suitable dataset and a solid (meta-)scientific background. Both of these elements will be discussed in turn, including some guidance based on personal experience.

A suitable dataset

A multiverse analysis can be conducted on newly gathered data, or one could reuse an existing dataset. From an educational point-of-view, the former option is fairly comparable to a typical student research project, though the eventual statistical analyses will be considerably more elaborate, sophisticated and time-consuming. Focusing on existing data is perhaps more unusual in the context of a research methods class, in that it involves finding a suitable dataset and isolating the hypotheses of interest, rather than designing a study to test a hypothesis and collecting data. For short projects or when students are relatively inexperienced, the instructor could select one or a few suitable studies, thus assuring that students can hit the ground running. Alternatively, students with a stronger background could be given the opportunity to find a suitable study themselves.

Selecting a study from the literature for a multiverse analysis comes with several challenges. One obvious requirement for such a study is that it should have publically available data, or that the original authors share their data for the agreed-upon purposes. This already narrows down the pool of studies, as psychological scientists are often not able or willing to share research data (Vanpaemel et al., 2015; Wicherts et al., 2006), though since the start of the open science movement and its various initiatives (e.g., Morey et al., 2016), there has been an increase in data availability (Kidwell et al., 2016). Furthermore, even if data are available, it does not necessarily imply that they are amenable to a multiverse analysis. It might, for instance, be unclear what a certain variable measures, or how a data file is structured (Hardwicke et al., 2018). Obviously, the multiverse-using-existing-data-approach is only feasible when one has access to reusable data.

Another important criterion is that the study should afford plausible alternative data-analytic pathways, tailored to the students' capabilities, to test the hypothesis of interest. We suspect that many studies in psychology meet this requirement, by for example focusing on outlier detection, dichotomization of variables, covariate inclusion, and so on. However, the data need to be available at a level raw enough to allow the construction of different pathways. If one only has access to the processed data (e.g., after dichotomization), rather than to the raw data, certain reasonable alternative processing and analysis options can not be explored.

A final issue to consider is analytical reproducibility (i.e., conducting the same analyses on the same dataset and obtaining the same results). Ideally, one selects a study of which the (most important) results are reproducible, or, at minimum, that the reason for non-reproducibility is clear. This requirement restricts the pool of possible target studies even further, as analytic reproducibility within psychological research has been shown to be far from ideal. For example, Hardwicke and colleagues (2018) were able to independently

reproduce the key results from only 11 out of 35 articles with reusable data published in the journal *Cognition*. More surprisingly, even with the help of the original authors, key results of 13 articles could not be reproduced. Artner et al. (2020) describe similar struggles in their attempt to reproduce 232 key statistical claims from 46 articles, based on the raw data, without help from the original authors (see also Wicherts et al., 2011). Although reproducibility is not strictly necessary in order to conduct a multiverse analysis, it does provide some reassurance that the data were processed and interpreted in the way intended by the authors. For example, before conducting their multiverse analysis, Steegen et al. (2016) had to correct various minor reporting errors in the original data, which were discovered only by first attempting to reproduce the results (see their supplemental materials).

If the original results are not (entirely) reproducible, but the source of the inconsistencies is easily identifiable (e.g., use of dummy coding rather than effect coding or correctable typos in the data file), one can still be reasonably confident in one's understanding of the data-analysis, and the study might be a suitable target for the type of research project described here. In fact, such cases can be especially interesting from an educational point of view, as they demonstrate the project's relevance, and illustrate that even accomplished researchers might struggle with data analysis at times. Yet, when there is no discernable explanation for non-reproducible results, undertaking a multiverse analysis is potentially fruitless, especially when the discrepancies are substantial, because one might have misinterpreted the data. Of course, it is also possible that the original authors made a mistake, but it can be time-consuming to figure this out, and the authors might not be able or willing to help clear up any discrepancies.

Finding a study meeting all these requirements can be quite challenging, for students and instructors alike. A useful starting point for this search process is the article library on curatescience.org, which provides the possibility of filtering articles based on the availability

of data (LeBel et al., 2018). Furthermore, one could browse repositories like the Open Science Framework (Soderberg, 2018) for articles with open data. Consulting recent issues of journals using badges to signal articles with open data and open materials (https://www.cos.io/our-services/badges; Kidwell et al., 2016) is another excellent option. Of course, the instructor could provide a dataset of their own or one they are already familiar with. This could either be the primary or only option (see Example Application below), or as a back-up in case (some) students wouldn't be able to find a suitable dataset themselves. Based on our experience, both of these approaches work well.

A solid (meta-)scientific background

It is important to build a solid meta-scientific framework, and provide students with sufficient background information about multiverse analyses at the beginning of the project (unless they are already familiar with these concepts from other courses). For example, one could cover some insightful meta-scientific articles such as Simmons and colleagues' paper (2011) about researcher degrees of freedom and their effect on the false positive rate, Gelman and Loken's paper (2014) describing how data-analysis can be conceived as a garden of forking paths, and Steegen et al.'s (2016) introduction to multiverse analyses. That way, students are gently introduced to the concept of a multiverse analysis and the rationale behind it. In addition, it serves to foster critical thinking and demonstrates the relevance of such (meta-)scientific studies, including their own.

Besides these more general meta-scientific articles, students could benefit from several (published) examples of a multiverse analysis (e.g., Credé & Phillips, 2017; Moors & Hesselmann, 2019), to give them an idea of what it concretely entails. This serves two purposes. One, it provides guidance on how to summarize and interpret the outcome of a multiverse analysis (e.g., plotting a distribution of p-values, or creating a heatmap with p-

values as a function of the various analytic pathways). Two, it stimulates students in recognizing potentially arbitrary choices, thus giving them inspiration for their own multiverse.

Still, it can be quite challenging and overwhelming for students to generate alternative data-analytic pathways. A useful source, besides the papers mentioned above, is the work of Wicherts and colleagues (2016), which offers a comprehensive overview of researcher degrees of freedom. Moreover, one could also encourage students to look for alternative pathways in related work. In particular, when the project involves re-analysis of a published study, students could critically assess the rationale behind the article's data-analytic choices, or examine papers cited in the target article as well as previous publications from the same authors on the same topic. To facilitate this, the instructor could organize a (group) discussion about the paper in question and point out some potentially relevant or remarkable choices. Students could (or should) also try to reproduce the original findings, if they haven't done so already as part of the process to select the target study (see above). That way, students familiarize themselves with the target study and its data, which might give them ideas for their eventual multiverse.

Throughout the project, strong guidance is needed. It is critical to inform students about the expectations regarding a multiverse analysis, and to tackle misconceptions. For one, the goal should not be to merely devise as many paths as possible. The key is that the alternatives are properly motivated --- quality over quantity (Del Giudice & Gangestad, in press). Furthermore, when multiple students use the same dataset, it is perfectly plausible to end up with different paths, and thus potentially a seemingly-contradicting answer to the same research question. This does not mean that someone made a mistake, rather it shows the ubiquity of arbitrary decisions. Clear communication about these issues is important to avoid any confusion among students. Providing feedback to students, particularly when it comes to

the construction and implementation of the multiverse analysis, is also instrumental to make the project a success. Some students may come up with poorly motivated alternative pathways, in which case the supervisor should steer them in the right direction or encourage them to carefully (re)consider the rationale for their choices. Feedback could also take the form of a group discussion at a later stage of the project, to address the different pathways students came up with and compare their outcomes.

Though not strictly necessary, basic knowledge of R (i.e., a programming language primarily used for data analysis and visualization; R Core Team, 2020), or even R Markdown (i.e., an environment to create dynamic, reproducible reports; Allaire et al., 2016), can help students in running their analyses and reporting their results, yet there is quite a steep learning curve. Multiverse analyses involve combining different options (e.g., different outlier criteria for different dependent variables that are transformed in various ways). Especially when this amounts to many individual pathways, it will be more efficient to integrate them instead of performing each analysis separately, yet that does require some programming experience or training.

Example Application

This section describes an actual implementation of the multiverse-in-the-classroom approach in the context of an undergraduate research project (see Table 1 for a summary of the syllabus). Besides illustrating the viability of the approach, we hope that it can inspire instructors, course coordinators, and program directors who would consider including multiverse analyses in their research-methods classes. Of course, there are many alternative ways to implement the multiverse-in-the-classroom approach, taking into account aspects such as timing, group size, students' prior knowledge, learning objectives, and so on.

The project took place in the 2020 spring semester with the first author as the instructor, and was inspired by a course jointly-taught by both authors in previous years. It was embedded in a course called Bachelorproject, which spans 17 weeks, and is organized for students in the final year of their undergraduate psychology program. These students have already followed several statistics and methods courses, typically amounting to 30 European Credits (EC). The Bachelorproject represents a study load of 15 EC, during which students need to write an individual thesis describing the outcomes of a research project. The course is mandatory for all undergraduate psychology students, but they are divided in small groups each with a different instructor and a different research topic (e.g., mental health in university students, examining people's interest in psychedelics, individual differences in the attentional bias towards emotion,....). The multiverse-in-the-classroom approach described here was used in one such group, consisting of eight students. Students ultimately had to write a thesis about their project following the typical Introduction-Method-Results-Discussion structure. The resulting products were evaluated on the same criteria as other research projects within the course by two independent graders (including the instructor). In addition, the instructor also graded the process as a whole.

The project involved the re-analysis of an existing dataset, which was provided by the instructor. The selected target article was a study by Smith and colleagues (2019), examining the influence of acute stress on semantic memory retrieval. Smith et al. found that participants performed better on an open-ended trivia questionnaire after experiencing acute stress, and when they showed a stronger stress response. The study met all of the above criteria: reusable processed data were available in detailed enough format (the underlying raw data were, at the time, available upon request, and are now publically available; see Smith, 2020); the results were reproducible (except for one easily-identifiable deviation); and the data processing and analysis steps afforded various alternative pathways.

In a first meeting with the students of +- 1 hour the general topic of the thesis was introduced by the instructor. This included a short description of the target study as well as a brief introduction to the concept of a multiverse analysis. In the next meeting (+- 2 hours), the target article was examined in detail through a journal club, in which the instructor led the discussion. Students were expected to read the article in advance, and were encouraged to pay special attention to methodological and data-analytical choices. Furthermore, any aspects of the paper that were unclear to the students were addressed during the meeting. From this point onwards, students were encouraged to start thinking about alternative analysis pathways, inspired by the group discussion, through searching for literature around the same topic, etc.

The third meeting (+- 1.5 hours) consisted of an interactive lecture on data sharing (including ethical issues such as protecting the privacy of participants), reproducibility, and scientific integrity (including a discussion of questionable research practices). The idea is to introduce some concepts that are directly relevant for their thesis (e.g., reproducibility) as well as to give students a broad overview of meta-scientific topics.

The next four meetings (+- 2 hours each) involved journal clubs around articles on, respectively, data sharing and reproducibility (i.e., Wicherts et al., 2011 and Hardwicke et al., 2018), researcher degrees of freedom (i.e., Simmons et al., 2011), and multiverse analysis (Steegen et al., 2016). Each time, two students led the discussion, but everyone was supposed to read the paper in advance and take part in the discussion. The instructor intervened sporadically if something was unclear or to point out relevant aspects. The purpose of these meetings was three-fold. First, it served to build a solid meta-scientific background, and to give students inspiration for their own multiverse analysis. Second, writing the introduction section for a thesis about multiverse analyses can be challenging as it differs somewhat from that of a “regular” empirical study. Hence, discussing a few key articles puts them on the right

track. Finally, these journal clubs were also meant to improve students' presentation and discussion skills.

The four final collective meetings (+- 2 hours each) served to introduce the students to R and R Markdown. Students were guided through a custom-made script showing how to read in data, transform and combine datasets, use conditional statements and loops, make graphs, and perform all the analyses that were used in the target paper. The script already used the data from the target paper to make sure that students understood what the variables meant. Even though the script introduced all the procedures needed to reproduce the results of the target paper, they were illustrated using different variables. As a take-home exercise, students then tried to independently reproduce the key outcomes of the target paper using R, which they later embedded in an R Markdown document. This guaranteed that all students were (eventually) able to follow the processing and analysis pathways outlined by Smith and colleagues (2019). Note that it wasn't required from students to write their thesis in R Markdown, or even use R for their eventual analyses. In the end, all eight students conducted their multiverse analysis in R, and two of them wrote their final paper using R Markdown.

From that point onwards, each student had four individual feedback meetings with the instructor in which their research proposal (i.e., rationale for the different pathways), analysis plan, code, results, and write-up were discussed. Seventeen weeks after the start of the course, they were expected to submit their final thesis and accompanying analysis script.

An exhaustive overview of all the alternatives students came up with would take us too far, but the following examples serve to illustrate the versatility of a multiverse approach to (under)graduate research projects. For instance, Smith and colleagues considered responses to a trivia questionnaire as being correct if they completely matched the correct answer, were misspelled but easily extrapolated, were inappropriately pluralized or capitalized, were

common synonyms of the correct answer, or if the first four or more letters matched the correct answer. However, students considered various reasonable alternatives to this coding scheme, such as treating incomplete responses as incorrect, regardless of how many letters matched the correct answer (Boere, 2020; De Jong, 2020; Hoogeterp, 2020; Kraaijenbrink, 2020; Kuipers, 2020; Van Dijk, 2020; Van Rijn, 2020; Van Wijk, 2020). Exploring this variation was only possible because students had access to the raw data (i.e., responses of each participant to each question), as the processed data only contained accuracy scores per participant based on the original coding scheme. Furthermore, some students redefined the construct *reactivity to stress*. In the original paper, it was operationalized as the change in cortisol levels relative to a baseline, whereas students also considered the change in the psychological stress response measured through the State-Trait Inventory for Cognitive and Somatic Anxiety, described in Grös, Antony, Simms, and McCabe, 2007 (Hoogeterp, 2020; Van Rijn, 2020). Additionally, some students added covariates to the analyses (e.g., age; Kraaijenbrink, 2020), or removed covariates (i.e., gender; Hoogeterp, 2020; Kuipers, 2020; Van Wijk, 2020). Yet other pathways involved imputing missing values (Kraaijenbrink, 2020), or removing observations (e.g., excluding participants who did not display an elevated cortisol level after stress-induction; Boere, 2020; Van Dijk, 2020).

Although there was some overlap in data-analytic choices between students, each individual project featured unique pathways, which were based on existing literature (e.g., Merz et al., 2016), statistical arguments, and/or a critical appraisal of the original study. The breadth of options is illustrated in Figure 1, showing the distribution of p-values for Smith et al.'s (2019) main finding resulting from each students' multiverse analysis (see https://osf.io/rtayk/ for the underlying R code). On average, students' multiverse analyses comprised 78 paths (minimum 18, maximum 160).

This outcome highlights the feasibility and potential of undergraduate research projects incorporating multiverse analyses. We hasten to add that it does not serve as a way to evaluate the robustness of Smith and colleagues' main finding, because certain data-analytic choices explored by the students were insufficiently motivated. The work done by the students, however, offers an ideal starting point for a more thorough multiverse analysis of the finding (see Heyman et al., in preparation).

Benefits of the Multiverse-in-the-Classroom Approach

Incorporating multiverse analyses in (under)graduate research projects (or other courses) has many benefits for students as well as (psychological) science in general.

One strength of the multiverse-in-the-classroom approach is that it can be flexibly adapted to the course's learning objectives, classroom size, time frame, background of the students, and so on. For instance, one can conduct a multiverse analysis reusing an existing dataset, like in the example described above, or one could use newly gathered data. Because the latter option involves an additional step compared to a typical research project, it is well-suited for situations where something extra is required from students (e.g., students enrolled in an honours program), whereas the former option can be applied more broadly. Importantly, as there is no need to design a new study, or to collect any data, the students' overall time-investment is comparable to that of a regular research project. Moreover, an adapted version of such a multiverse project can be used in a more statistics-oriented course rather than a research-methods-oriented course. Both authors have used a similar approach as part of a 13-week graduate statistics course within a psychology research master track for a number of years. There, the +- 40 students were instructed to write a report about the multiverse analysis they conducted in small groups using existing data. Because these graduate students are well-versed in statistical analyses and programming, and due to the group-nature of the project, it

can easily fit in a 13-week course as compared to the 17-week undergraduate research project described above.

As a multiverse project does not necessarily require collecting new data, one could effectively save a lot of resources (i.e., time of participants and students, money to pay participants,…). Therefore, it is ideal for situations where collecting new data is impractical or impossible, for instance, because special equipment or expertise is required, getting ethical approval takes too much time, or when one does not have access to a participant pool or money to pay participants. This proved to be especially relevant in the lockdown situation due to COVID-19 in spring 2020. Indeed, the lockdown measures, which involved suspending all in-vivo data collection and required classes to be taught online, had very little impact on the project discussed above, with all students meeting the original deadline.

The flexibility also applies to the selection of a target study. Each student could focus on a separate paper, or, as was the case in the example above, each student independently construes their own data-analytic pathways for the same data set. The latter option is comparable to the many-analysts-one-dataset approach used by Silberzahn and colleagues (2018), augmented with the additional requirement that every analyst (i.e., student) should consider several plausible alternatives rather than a single one. We believe the many-multiverses-one-dataset option is the most interesting of the two, because any given multiverse will rarely (if ever) exhaust *all* reasonable options, hence it makes sense to adopt a form of data-analytic triangulation. In other words, there is a multiverse of multiverse analyses, which can be captured to some degree by asking different students to focus (semi-)independently on the same overarching topic. Although it is unrealistic to expect that every individual project will be of the same quality, it can be enlightening to see the variability, or lack thereof, in outcomes. Indeed, as Figure 1 shows, it is possible that some

multiverse analyses suggest the effect in question to be quite robust, whereas others suggest the effect to be rather fragile.

Despite bridging an important gap in psychological science by showing the robustness or fragility of findings, multiverse analyses are relatively rare, owing perhaps to their apparent complexity and/or their perceived lack of novelty. In that sense, one can draw a parallel to replication studies: once rare in psychology (Makel et al., 2012), they are now becoming more mainstream through various initiatives (see Zwaan et al., 2018). Moreover, Frank and colleagues (Frank & Saxe, 2012; Hawkins et al., 2018) promoted conducting replication studies in student research projects (see also Grahe et al., 2012 and Wagge et al., 2019). The current proposal seeks to accomplish a similar goal for multiverse analyses. Note that both approaches can complement each other, in that one can conduct a multiverse analysis on replication data, either as part of the same project or across different iterations of the course (e.g., one group conducting the replication study, and another group performing multiverse analyses, possibly the following semester or academic year).

Another major benefit of adopting the multiverse-in-the-classroom approach, besides its flexibility, is that it gives students the opportunity to make a tangible contribution to psychological science, something that might not always occur with (under)graduate research projects. Moreover, under some conditions, the work done by students and instructor(s) can be solidified in a joint research paper, suitable for publication, as was the case for the example application. The classroom phase then serves as an elicitation step of possible reasonable variations, which, in a second step, are evaluated for adoption in a multiverse analysis by a domain expert. Such a two-step multiverse analysis - where data-analytical pathways are first elicited from different sources which then get synthesized and applied to the data - can even yield more comprehensive and less biased results compared to a regular multiverse analysis.

The multiverse in the classroom approach also provides ample pedagogical opportunities. Conducting a multiverse analysis typically requires students to perform a number of different statistical analyses. It thereby addresses an often-heard complaint from psychology students regarding the relevance of statistics. Even though most research projects involve the practical application of statistics, it rarely is a focal point (in some cases, the analysis part might actually be considered a nuisance). Furthermore, a multiverse project may help students to better understand the intricacies of statistical analyses. Importantly, it is not a purely methodological or statistical project, as it also involves an empirical research question such as "to what extent does power posing have an effect on hormone levels" or "what is the effect of stress on semantic memory". Hence, there is still the thrill of discovery, which helps fuel students' engagement.

At a more abstract level, a multiverse project also allows students to gain first-hand experience with the importance of open science, reproducibility, proper documentation of data, and so on. In addition, it teaches them to critically evaluate the rationale behind a study, especially its methodology, and it gives them an idea about the imperfections of psychological science. As future consumers of research, it is relevant to recognize that arbitrary decisions abound in research, and to realize their consequences. A multiverse-in-the-classroom project really drives this point home. Moreover, for those students aspiring to become producers of research, it is paramount to adopt responsible research practices, such as assessing the robustness of key outcomes. In fact, students spontaneously mentioned these aspects in an informal evaluation of the course[1] (e.g., "it has really changed my perspective on research… and sparked my interest" or "it was interesting to see what happens to the p-values when conducting different analyses"). One could argue that typical research projects, in which

[1] Formal evaluations were not collected at the first author's institution due to the COVID-19 outbreak.

students are required to develop a new hypothesis, design a new study and collect data, teaches them bad habits or even questionable research practices as it is rather difficult to accomplish all this in a rigorous manner within the, usually limited, timeframe.

Challenges and Objections

A multiverse-in-the-classroom project can involve designing a new study, but that might not be feasible within the confines of a single semester, because developing and conducting such an analysis in itself is rather time-consuming. The option involving existing data is more readily applicable, yet one potential objection is that such a project does not cover the entire empirical cycle. Although a multiverse project requires a thorough literature search, motivating a research question, and a comprehensive data-analysis of which the results ought to be interpreted and discussed, students may miss out on learning specific skills (e.g., regarding data collection). When the development of such skills is a central objective of the course, one might need to look for a creative solution. For instance, in the example application described above, the absence of a data collection phase was addressed by having students recode the participants' responses to the trivia questionnaire. Note though that one can raise similar concerns about more widely-applied projects such as those involving online data collection. In fact, there is often quite a bit of variability in what is demanded of students across projects within the same (under)graduate program. More fundamentally, accreditation guidelines for research projects in psychology often explicitly mention the possibility to conduct secondary data analyses (e.g., Australian Psychology Accreditation Council, n.d.; The British Psychological Society, 2019).

Another challenge of conducting a multiverse analysis is that it requires combining various alternatives (e.g., three different outlier criteria and four different data transformations yield 12 outcomes). In principle, every analysis can be conducted separately, but this becomes

unwieldy quite quickly, so one could use a script to increase efficiency. Depending on the students' background, the latter option might prove to be unattainable unless one would include some programming classes in the curriculum (e.g., teaching the language R).

Another potential hurdle for students (and instructors alike) revolves around the interpretation of a multiverse analysis. In contrast to a typical research project, one does not end up with a single outcome, but with a collection of outcomes. This elicits questions such as when should a finding be considered robust, when is it presumably a fluke, and how should the results be summarized and presented. Indeed, published papers involving multiverse analyses typically eyeball the pattern of results, for instance, by plotting the distribution of p-values. Steegen et al. (2016) tentatively suggest to focus inference on the average p-value, but beyond that, there is little guidance as to how to synthesize a multiverse analysis (but see Simonsohn, Simmons, & Nelson, 2020).

A more fundamental objection could be that approaches such as pre-registration are more desirable, so that students should spend their time learning about pre-registration rather than about multiverse analyses. Pre-registration entails that one specifies the analysis plan before knowing its results, if possible even before starting the data collection (Nosek et al., 2018). As such, pre-registration makes transparent which choices could be data-driven and which are not. However, if a researcher pre-registers one or few analytic pathways, one is still left in the dark about how robust or fragile the effect is, or about whether certain choices are more critical than others (for a similar argument see Steegen et al., 2016). To that end, one would need to conduct a multiverse-style analysis. Of course, one could pre-register a multiverse analysis to combine the strengths of both approaches, but this increases the complexity of the project.

Finally, one should be cautious that students do not completely lose faith in (psychological) science. Indeed, whereas the goal is to make students critical consumers of scientific output, and, as a result, careful producers of scientific output, they should not come away with the idea that science is inherently flawed or that all researchers are opportunistic or fraudulent. Along the same lines, students should be made aware that not all hypotheses can necessarily be tested in a myriad of ways. Based on the informal evaluation mentioned above, students did not come away with such incorrect notions, but future research on the effectiveness of the multiverse-in-the-classroom approach should determine whether this is indeed the case.

Conclusion

The present paper proposes to implement multiverse analyses in student research projects, and provides a practical demonstration that we hope will encourage, help and inspire instructors to adopt it in their own courses. Because multiverse analyses speak to the robustness of a (published) finding, it can fulfill an important need in psychological science, thus making the results of such projects truly relevant. Furthermore, it is an excellent way to put statistics in practice, it fosters critical thinking, and raises awareness about the prevalence and consequences of arbitrary data-analytic decisions. Finally, the flexibility of the multiverse-in-the-classroom approach makes it suitable for all kinds of projects, even when data collection is not feasible.

Author Contributions

Both authors conceptualized the idea. TH wrote the first draft of the manuscript and WV provided extensive feedback. Both authors approved the final version for submission.

Competing Interests

The authors declare that there were no conflicts of interest with respect to the authorship or the publication of this article.

References

Allaire, J. J., Cheng, J., Xie, Y., McPherson, J., Chang, W., Allen, J., Wickham, H., Atkins, A., & Hyndman, R. (2016).rmarkdown: Dynamic Documents for R. Retrieved from https://CRAN.R-project.org/package=rmarkdown

Artner, R., Verliefde, T., Steegen, S., Gomes, S., Traets, F., Tuerlinckx, F., & Vanpaemel, W. (2020). The reproducibility of statistical results in psychological research: An investigation using unpublished raw data. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000365

Australian Psychology Accreditation Council. (n.d.). *Accreditation Standards for Psychology Programs Evidence Guide (Version 1.2).* https://www.psychologycouncil.org.au/sites/default/files/public/Evidence_Guide_20180718_V1.2_0.pdf

Bishop, D. V. M. (2016). Open research practices: Unintended consequences and suggestions for averting them.(commentary on the peer reviewers' openness initiative). *Royal Society Open Science*, *3*(4), 160109. https://doi.org/10.1098/rsos.160109

Boere, R. (2020). *Het belang van reproduceerbare en transparante wetenschap: Een multiverse benadering.* [The importance of reproducible and transparent science: A multiverse approach]. [Unpublished bachelor's thesis]. Leiden University.

Carney, D. R., Cuddy, A. J., & Yap, A. J. (2010). Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological Science, 21* (10), 1363-1368. https://doi.org/10.1177/0956797610383437

Credé, M., & Phillips, L. A. (2017). Revisiting the power pose effect: How robust are the results reported by Carney, Cuddy, and Yap (2010) to data analytic decisions? *Social Psychological and Personality Science, 8* (5), 493-499. https://doi.org/10.1177/1948550617714584

De Jong, S. (2020). *Het effect van stress op het semantisch geheugen: Een multiverse benadering.* [The effect of stress on semantic memory: A multiverse approach]. [Unpublished bachelor's thesis]. Leiden University.

Del Giudice, M., & Gangestad, S. W. (in press). A traveler's guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. *Advances in Methods and Practices in Psychological Science*.

Dragicevic, P., Jansen, Y., Sarma, A., Kay, M., & Chevalier, F. (2019, May). Increasing the transparency of research papers with explorable multiverse analyses. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-15).

Elson, M. (2016). Flexibility in methods & measures of social science. https://www.flexiblemeasures.com/

Frank, M. C., & Saxe, R. (2012). Teaching replication. *Perspectives on Psychological Science, 7* (6), 600-604. https://doi.org/10.1177/1745691612460686

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

Grahe, J. E., Reifman, A., Hermann, A. D., Walker, M., Oleson, K. C., Nario-Redmond, M., & Wiebe, R. P. (2012). Harnessing the undiscovered resource of student research

projects. *Perspectives on Psychological Science*, *7*(6), 605-607. https://doi.org/10.1177/1745691612459057

Grös, D. F., Antony, M. M., Simms, L. J., & McCabe, R. E. (2007). Psychometric properties of the State-Trait Inventory for Cognitive and Somatic Anxiety (STICSA): Comparison to the State-Trait Anxiety Inventory (STAI). *Psychological Assessment*, *19*(4), 369-381. https://doi.org/10.1037/1040-3590.19.4.369

Harder, J. A. (2020). The Multiverse of Methods: Extending the multiverse analysis to address data-collection decisions. *Perspectives on Psychological Science*, 15(5), 1158-1177. https://doi.org/10.1177/1745691620917678

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., Hofelich Mohr, A., Clayton, E., Yoon, E. J., Tessler, M. H., Lenne, R. L., Altman, S., Long, B., & Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal *Cognition*. *Royal Society Open Science*, *5* (8), 180448. https://doi.org/10.1098/rsos.180448

Hawkins, R. X. D., Smith, E. N., Au, C., Arias, J. M., Catapano, R., Hermann, E., Keil, M., Lampinen, A., Raposo, S., Reynolds, J., Salehi, S., Salloum, J., Tan, J., & Frank, M. C. (2018). Improving the replicability of psychological science through pedagogy. *Advances in Methods and Practices in Psychological Science*, *1* (1), 7-18. https://doi.org/10.1177/2515245917740427

Hoogeterp, L. (2020). *Het effect van stress op het semantisch geheugen: Een multiverse benadering*. [The effect of stress in semantic memory: A multiverse approach]. [Unpublished bachelor's thesis]. Leiden University.

Kalokerinos, E. K., Erbas, Y., Ceulemans, E., & Kuppens, P. (2019). Differentiate to regulate: Low negative emotion differentiation is associated with ineffective use but not selection of emotion-regulation strategies. *Psychological Science*, *30*(6), 863-879. https://doi.org/10.1177/0956797619838763

Kidwell, M. C., Lazarevic, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., Kennett, C., Slowik, A., Sonnleitner, C., Hess-Holden, C., Errington T. M., Fiedler, S., & Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLoS Biology*, *14* (5), e1002456. https://doi.org/10.1371/journal.pbio.1002456

Kierniesky, N. C. (2005). Undergraduate research in small psychology departments: Two decades later. *Teaching of Psychology*, *32*(2), 84-90. https://doi.org/10.1207/s15328023top3202_1

Kraaijenbrink, J. (2020). *The effect of stress on the semantic memory: A multiverse approach* [Unpublished bachelor's thesis]. Leiden University.

Kuipers, C. (2020). *The effect of stress on the semantic memory: A multiverse approach* [Unpublished bachelor's thesis]. Leiden University.

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A unified framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science*, *1*(3), 389–402. https://doi.org/10.1177/2515245918787489

Liu, Y., Kale, A., Althoff, T., & Heer, J. (2020). Boba: Authoring and visualizing multiverse analyses. arXiv:2007.05551.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7* (6), 537–542. https://doi.org/10.1177/1745691612460688

Masur, P. K., & Scharkow, M. (2019). specr: Statistical functions for conducting specification curve analyses. Available from https://github.com/masurp/specr

Merz, C. J., Dietsch, F., & Schneider, M. (2016). The impact of psychosocial stress on conceptual knowledge retrieval. *Neurobiology of Learning and Memory*, *134*, 392-399. https://doi.org/10.1016/j.nlm.2016.08.020

Moors, P., & Hesselmann, G. (2019). Unconscious arithmetic: Assessing the robustness of the results reported by Karpinski, Briggs, and Yale (2018). *Consciousness and Cognition*, *68*, 97-106. https://doi.org/10.1016/j.concog.2019.01.003

Morey, R. D., Chambers, C. D., Etchells, P. J., Harris, C. R., Hoekstra, R., Lakens, D., Lewandowsky, S., Coker Morey, C., Newman, D. P., Schönbrodt, F. D., Vanpaemel, W., Wagenmakers, E-J., Zwaan, R. A. (2016). The Peer Reviewers' Openness Initiative: Incentivizing open research practices through peer review. *Royal Society Open Science*, *3*(1), 150547. https://doi.org/10.1098/rsos.150547

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600-2606. https://doi.org/10.1073/pnas.1708274114

R Core Team. (2020). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Sarma, A., & Kay, M. (2019). multiverse: Explorable Multiverse data analysis and reports in R. R package version 0.1.4.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., ... & Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1* (3), 337-356. https://doi.org/10.1177/2515245917747646

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

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*, 1208-1214. https://doi.org/10.1038/s41562-020-0912-z

Smith, A. M. (2020, June 30). Acute stress enhances general-knowledge semantic memory. https://doi.org/10.17605/OSF.IO/EQ8SY

Smith, A. M., Hughes, G. I., Davis, F. C., & Thomas, A. K. (2019). Acute stress enhances general-knowledge semantic memory. *Hormones and Behavior*, *109*, 38-43. https://doi.org/10.1016/j.yhbeh.2019.02.003

Soderberg, C. K. (2018). Using OSF to share data: A step-by-step guide. *Advances in Methods and Practices in Psychological Science*, *1*(1), 115-120. https://doi.org/10.1177/2515245918757689

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

The British Psychological Society. (2019). *Standards for the accreditation of undergraduate, conversion and integrated Masters programmes in psychology*. https://www.bps.org.uk/sites/bps.org.uk/files/Accreditation/Undergraduate%20Accreditation%20Handbook%202019.pdf

Van Dijk, M. (2020). *Acute stress enhances semantic memory: The robustness of the findings of Smith, Hughes, Davis, and Thomas (2019).* [Unpublished bachelor's thesis]. Leiden University.

Vanpaemel, W., Vermorgen, M., Deriemaecker, L., & Storms, G. (2015). Are we wasting a good crisis? The availability of psychological research data after the storm. *Collabra, 1* (1), 1-5. https://doi.org/10.1525/collabra.13

Van Rijn, L. (2020). *The effect of stress on semantic memory: A multiverse approach* [Unpublished bachelor's thesis]. Leiden University.

Van Wijk, T. (2020). *Het effect van stress op semantisch geheugen: Een multiverse benadering*. [The effect of stress on semantic memory: A multiverse approach]. [Unpublished bachelor's thesis]. Leiden University.

Voracek, M., Kossmeier, M., & Tran, U. S. (2019). Which data to meta-analyze, and how? A specification-curve and multiverse-analysis approach to meta-analysis. *Zeitschrift für Psychologie*, 227, 64-82. https://doi.org/10.1027/2151-2604/a000357

Wagge, J. R., Baciu, C., Banas, K., Nadler, J. T., Schwarz, S., Weisberg, Y., IJzerman, H., Legate, N., & Grahe, J. (2019). A demonstration of the Collaborative Replication and Education Project: Replication attempts of the red-romance effect. *Collabra: Psychology*, *5*(1), 5. https://doi.org/10.1525/collabra.177

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PloS One*, *6* (11), e26828. https://doi.org/10.1371/journal.pone.0026828

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61* (7), 726-728. https://doi.org/10.1037/0003-066X.61.7.726

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, 7, 1832.

Young, C., & Holsteen, K. (2017). Model uncertainty and robustness: A computational framework for multimodel analysis. *Sociological Methods & Research*, *46*(1), 3-40. https://doi.org/10.1177/0049124115610347

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*, e120. https://doi.org/10.1017/S0140525X17001972

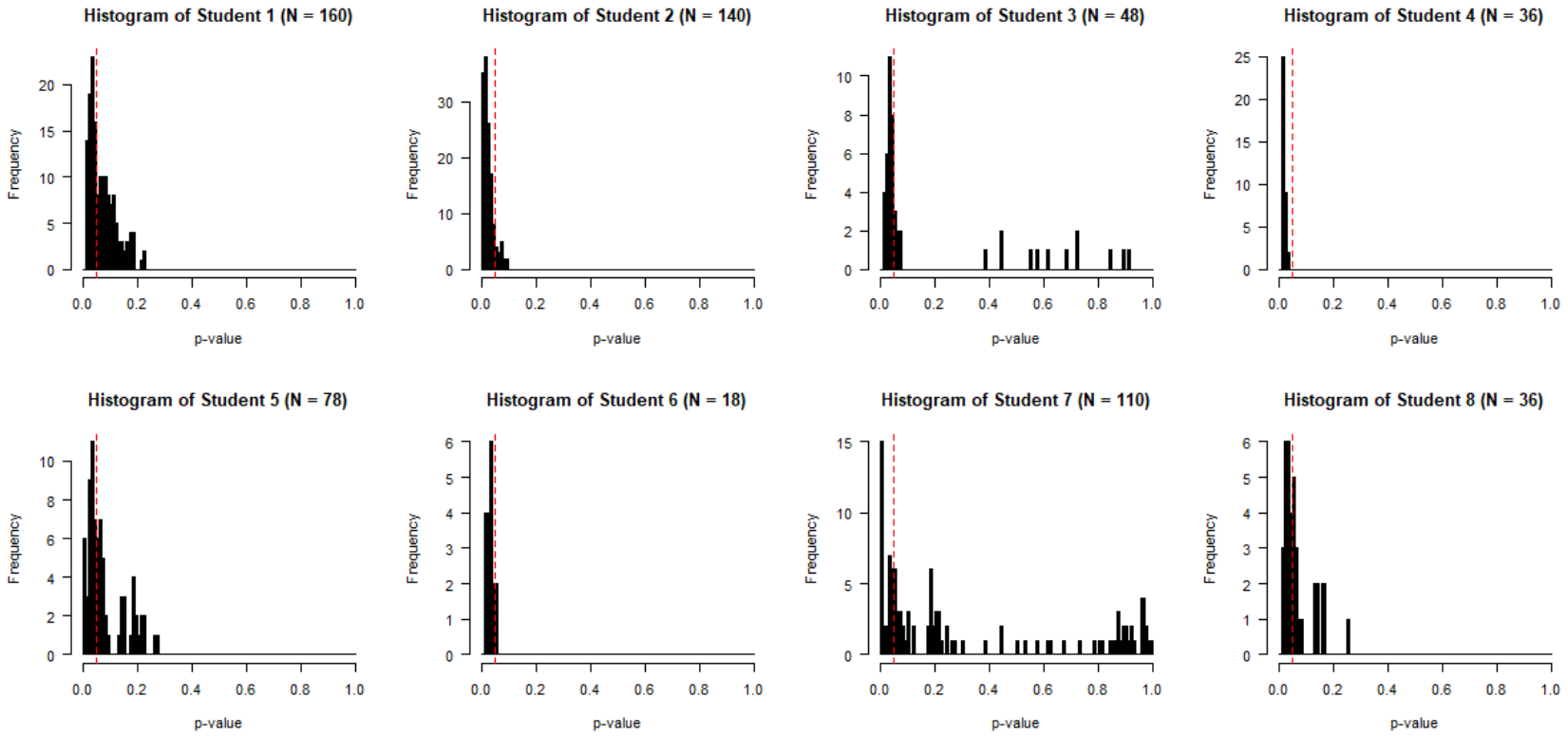
Table 1

Summary of the Syllabus for the Undergraduate Research Project Involving a Multiverse Analysis

Timing	Activity	Primary learning objective(s)
Week 1	General introduction	Understand the topic of the thesis
Week 2	Group discussion of target article (i.e., Smith et al., 2019) Class on ethics, data sharing, and reproducibility	Engage in critical thinking about the target article Understand the importance of data sharing and reproducibility
Week 3	Group discussion of Wicherts et al. (2011) Group discussion of Hardwicke et al. (2018)	Understand the importance of data sharing and reproducibility Understand the importance of data sharing and reproducibility
Week 4	Group discussion of Simmons et al. (2011) Group discussion of Steegen et al. (2016)	Recognize researchers' degrees of freedom and realize their impact Understand what a multiverse analysis entails, how to conduct one, and see how the results could be presented
Week 5	R intro	Perform data processing, visualization, and plotting in R
Week 6	R Markdown intro	Write reproducible and dynamic reports
Week 7-17	Conduct multiverse analysis and write thesis (including four opportunities for individual feedback)	Write a thesis incorporating relevant feedback

Figure 1

Distribution of P-values for Smith et al.'s (2019) Main Finding Resulting from Each Students' Multiverse Analysis



Note. The red dotted line indicates a p-value of .05. The figure in brackets indicates the number of pathways in each student's multiverse. Remark that not all pathways were properly motivated, so these results should not be considered an evaluation of the robustness of Smith and colleagues' main finding.