PRE-PRINT

What should a preregistration contain?

Jonathon McPhetres
Massachusetts Institute of Technology
University of Regina

A large amount of variation exists in beliefs about the purpose and benefits of preregistration, making it difficult to implement and evaluate, and limiting its usefulness. Additionally, no single resource exists to describe what a preregistration should contain or how it should be used. In this paper, I describe what an effective preregistration should contain and when it should be used. Specifically, preregistration should 1) restrict as many researcher degrees of freedom as possible, 2) detail all aspects of a study's method and analysis, 3) detail information on decisions made during the planning stages, and 4) specify how the results will be used and interpreted. Further, a preregistration must be publicly verifiable and permanent. Finally, I argue that preregistration should be used in any situation where researchers intend to collect data in order to make a claim, description, decision, or inference based on that data. I also note that preregistrations which do not address each of these points do more harm than good by falsely signalling credibility and quality.

Practical significance: This manuscript also provides general guidelines as to what an effective preregistration should include and when it should be used. It is also argued that preregistration should be applied to all research activities, including descriptive and exploratory studies.

Keywords: Preregistration; meta-science, hypothesis testing; transparency; planning.

"*But what is so novel about this? This is* the *method of science and always has been; why give it a special name? The reason is that many of us have almost forgotten it…How many of us write down our alternatives and crucial experiments every day, focusing on the* exclusion *of a hypothesis?" -Platt, 1964, pp. 347-348.*

Preregistration was initially suggested as a way to control false-positive error rates on probability testing when examining a hypothesis (Nelson, Simmons, & Simonsohn, 2018; Simmons, Nelson, & Simonsohn, 2011) and to curb the presenting of exploratory results as confirmatory. (e.g. HARKing; Kerr, 1998; Mayrhofer, 2017; Rubin, 2017). A literature search would even give the impression that preregistration need only be used when a hypothesis is tested and p-values are reported Nosek, Ebersole, DeHaven, & Mellor, 2018; van 't (Buchanan & Hvizdak, 2009; Lakens, 2019; Veer & Giner-Sorolla, 2016; Wagenmakers & Dutilh, 2016). However, no single resource exists to describe what an effective should contain or what level of comprehensiveness should be attained.

This manuscript will outline what a preregistration should contain to achieve that level of comprehensiveness. I begin by discussing the current state of preregistration and then move on to discussing the contents of an ideal preregistration.

The current state of preregistration.

If a naive researcher were to search the literature on pre-registration, it would be reasonable to come to the conclusion that pre-registration is only useful (but still optional) under two conditions: 1) when a hypothesis is tested, and 2) when a *p*-value is reported. Indeed, the majority of literature on the values, benefits, and criticisms of preregistration centres around these two activities (Forstmeier, Wagenmakers, & Parker, 2017; Ledgerwood, 2018; Moore, 2016; Nelson et al., 2018; Nosek, Ebersole, DeHaven, et al., 2018; Szollosi et al., 2020; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012).

It would also be easy to come to the conclusion that planning and restricting freedom in the design

Jonathon McPhetres, Sloan School, Massachusetts Institute of Technology and Hill/Levene Schools of Business, University of Regina.
Correspondence concerning this paper should be addressed to Jonathon McPhetres. Email: jon.mcphetres@gmail.com

and analysis of a study is not a necessary aspect of preregistration. Indeed, many efforts towards preregistration focus on making the practice fast, easy, and accessible to newcomers.

These misconceptions limit the usefulness of preregistration and can cause confusion for reviewers, editors, and readers who wish to evaluate the quality of the planning put into a research report.

When should preregistration be used? Preregistration should be used for all social science research activities. That is, preregistration should be used any time a researcher plans to collect data in order to make a claim, description, decision, or inference based on that data. The reason for this is due to the flexibility inherent in the research process. The flexibility begins when one *conceptualizes* a construct and then makes decisions on how to measure and analyse that construct.

Thus, two important qualities of preregistration should be kept in mind when reading the rest of this paper and thinking about the goals of preregistration:

1. Preregistration is not limited to situations where a hypothesis is tested.
2. Preregistration is not limited to situations where a *p*-value is reported.

1. Preregistration is not limited to situations where a hypothesis is tested. Hypotheses are not necessary for preregistration because the pitfalls that preregistration helps avoid are still present in exploratory and descriptive research. These pitfalls are things like reporting false positives, twisting results to support a previously held conclusion, accidentally duping one's self into confirming an expectation, interpreting a faulty experiment, and reporting dishonest results. The only way to solve this problem is to plan out every aspect of a study ahead of time and to then follow that plan, whether it be for confirmatory, exploratory, or descriptive research.

Additionally, theories and hypotheses in psychology are often too vague and poorly specified to make clear predictions (Fiedler, 2018; Szollosi & Donkin, 2019; Yarkoni, 2019), let alone quantitative and point-specific predictions (Meehl, 1967). If there existed psychological theories that were specific enough to guide hypotheses, operationalizations, methodologies, and interpretations, we wouldn't need preregistration to begin with.

2. Preregistration is not limited to situations where a p-value is reported. One does not need *p*-values to make claims based on data, to use data to make decisions, or to report data descriptively. **Box 1** gives an example of one such situation where preregistration is useful but does not require *p*-values. The same amount of flexibility exists in a pilot project, a pre-test, a "descriptive" study, or a standard "confirmatory" study; this flexibility can equally influence the results and one's interpretation of those results. Specifically, research has shown that arbitrary decisions about the design and analysis of a study often have dramatic consequences for results (Gelman & Loken, 2013; Landy et al., 2020; Silberzahn et al., 2015) and that flexibility in the methods, analysis, and reporting of a study are responsible for increasing false-positive results (Simmons et al., 2011).

Additionally, researchers can preregister Bayes factors (van 't Veer & Giner-Sorolla, 2016; Wicherts et al., 2016), effect sizes (Lakens, 2014, 2017), qualitative themes (L. Haven & Van Grootel, 2019) or any other inference criteria they will implement.

Further, the claim that exploratory *p*-values are uninterpretable has not prevented researchers from conducting exploratory analyses and reporting and interpreting the *p*-values associated with those analyses alongside, and with the same veracity as, preregistered *p*-values. It seems that psychologists want their cake and want to be able to eat it, and to publish it, too.

The only way a preregistration can help diagnose the veracity of a *p*-value is by limiting as many researcher degrees of freedom as possible. So, the real culprit here is not the *p*-value itself, but rather the amount of flexibility inherent in the research process. Preregistration should be seen as a tool to limit that flexibility as much as possible.

What should an effective preregistration contain?

In this section, I lay out four pieces of information that a preregistration should detail (see https://osf.io/6qv2b for a comprehensive preregistration form). In addressing these four pieces of

Box 1: Pre-testing a set of stimuli

Situation: Some researchers wish to run an experiment involving a set of images as stimuli. To do this study, they need to run a pre-test from a larger set of possible images (let's say 20 images) to make sure people think the images are funny. Thus, there are (at least) two possible studies here: the pre-test (a so-called "descriptive" study) in which they select the stimuli, and the so-called "confirmatory" test in which they test a hypothesis using the stimuli from the pre-test.

The researchers briefly discuss (verbally) that they will run a survey where subjects rate the images on funniness and sadness (and perhaps a few other ratings just for good measure). The idea is to pick the top 5 images with the highest 'funniness ratings' and the bottom 5 images with the lowest funniness ratings; those 10 images will be used for the final survey.

It does not particularly matter to them which images are the funniest, so the researchers have not written down a declarative statement with specific guess of which items will be the funniest. As they also do not plan to run any significance tests, they have not preregistered the study.

Should the researchers preregister the pre-test? The traditional view of preregistration would suggest that, because there is no specific hypothesis and no *p*-values, the research should not be preregistered. However, there are still many, many areas for flexibility in the interpretation of this data that can have an impact on the results of the confirmatory study they will soon run, and this flexibility can be limited through the use of preregistration. Thus, there are compelling reasons for preregistration. To illustrate them, let's consider a few possible outcomes from this research.

***Outcome 1*:** The data come in and there are actually multiple images rated at the highest point on the scale, meaning that there are actually more than 5 images with the "highest" funniness ratings. What do the researchers do? They could have thought of a contingency plan to use a second variable in this case (e.g., selecting images with the highest funniness and lowest sadness ratings). In this case, a pre-planned set of guidelines on how they will interpret and use the results would have been useful. A bit of careful planning could have yielded a few possible outcomes where the researchers could have decided what to do ahead of time.

***Outcome 2*:** Suppose though that all 20 images had relatively the same funniness means from the pre-test, varying only at the decimal level. They run the second "confirmatory" study, but a reviewer asks whether the results are idiosyncratically dependent on the chosen set of images. The researchers respond with the information about their pre-test, indicating that these were the funniest images from a pre-test selection of 20 images. Thus, an arbitrary set of information is used to make a claim to support the conclusions of a research study.

***Outcome 3*:** Suppose all 20 items differ very little in terms of "funniness" means. So, the researchers decide to replace some set of items and run another pre-test. That is, they discard the data from the initial pre-test. What goes unnoticed is that the means of some of the items change drastically—some of the images that were rated as funny in the first pre-test have much lower means in this second pre-test. Thus, an arbitrary set of information is used to justify claims made as a result of the second confirmatory study.

***Outcome 4*:** The researchers realize that it is difficult to select the items from these results, so, they decide to do a significance test to examine whether some set of images "don't differ from each other on funniness" and use that subset: a process which could have benefitted equally from the traditional view of preregistration.

***Outcome 5*:** The researchers see that the images differ very little on the average ratings. So, they decide on a cut-off criterion, such as "X% of participants will rate the image above the mid-point of the scale." Alternatively, the researchers could decide on an effect size, a significance value, or open-ended descriptions to select the images.

***Outcome 6*:** The research determine that it is difficult to determine which stimuli to use, so they decide to ignore the pre-test altogether and simply select some items that they like best. The pre-test data is discarded and the results are not used or reported anywhere.

***Outcome 7*:** The researchers determine that it is difficult to select stimuli based only on funniness ratings. They elect to explore a variety of combinations of ratings to determine which items to select. In the end, they decide to first exclude items which have a rating of greater than 1 on *scary,* and then they select the items which have the great absolute difference between *funny* and *sad.*

Consider: Each of the above outcomes could have been avoided (or solved) by simply preregistering a study design, analysis plan, and considering how the results would be reported. Further, a preregistered study in a permanent repository would remain there for the researchers, reviewers, and readers to access and evaluate, thus avoiding the situation in Outcome 6.

information, a researcher will have carefully planned out their research, anticipated any problems, determined that the analyses can yield the answers they seek, and determined possible interpretations of those analyses. Further, planning out possible interpretations and detailing uses of the data will help avoid issues and criticisms when making claims based on the data.

Specifically, a preregistration should:

a) Restrict as many RDFs as possible
b) Detail all aspects of a study's method and analysis
c) Detail information on decisions made during the planning stages
d) Specify how the results will be used and interpreted

A preregistration should restrict as many RDFs as possible. Limiting flexibility in the analysis and reporting of results is the only way to help curb incorrect or misleading interpretations and false positives, and can increase confidence in the statistical and descriptive results. A checklist of RDFs can be found elsewhere (Wicherts et al., 2016) and I recommend that researchers look over this list when writing up a preregistration to ensure they can address each of these points. Additionally, this is why a longer, more detailed preregistration form is useful: it helps researchers anticipate areas of flexibility and decisions they may not have thought of. There is also some evidence that these longer forms are better at restricting RDFs (Veldkamp et al., 2017).

However, it's important to note that these RDFs don't apply only to situations where a hypothesis is tested or *p*-values are reported. Instead, as argued previously, these areas of flexibility apply to any situation in which a researcher intends to make a claim, description, decision, or inference based on some data. **Box 2** gives an example of one research context with such flexibility. Researchers often use descriptive data to bolster claims or design decisions; thus intentions for the uses of the data should be outlined, which requires that researchers consider different possible outcomes. Careful and exhaustive preregistration should be exercised even for so-called "exploratory" research.

A preregistration should detail all aspects of a study's method and analysis. As with above, a preregistration should detail all aspects of a study. While some of these points may be identified by looking at the list of RDFs, some of these points may not be obvious, such as:

All measures and materials included in a study: Surveys often include extra measures used for other purposes or items intended to be used for multiple projects, papers, or tests. Some tasks may be included so that they can be used in multiple studies; these decisions are economical and are perfectly reasonable. However, it's easy to ignore, include, or rationalize the use of some measures post-hoc when they aren't specified in a plan ahead of time. Identifying all the measures included in a given survey or data collection session will help prevent unintended or exploratory findings being passed off as confirmatory.

Thus, in a given preregistration, researchers should identify 1) all measures included in a survey/session, and 2) which will be used to test the present question or hypothesis, and 3) which will not. This way, evaluators have access to all the measures included in a given data session and the researcher can specify ahead of time which measures will be used for which purpose.

All details of item calculations, construct measurement, or task designs: Most psychological methods and theories are too poorly defined to guide a specific test (Fiedler, 2018; Szollosi & Donkin, 2019; Yarkoni, 2019) and, in the social sciences, nearly everything is correlated (Meehl, 1990). As a result, almost any task or measure could be used to make a claim about the world or to provide support for a hypothesis or a theory. Thus, all details about operationalization, measurement, item selection, scale calculations, or task design should be specified. This is also a great place to indicate decisions about why which set of measures, tasks, algorithms, calculations, or statistical procedures were chosen.

The actual analysis scripts used to analyse and visualize the data. Researchers should run through the actual analysis, interpretation, and reporting of their results prior to conducting the study using toy data. This can help in identifying problems with the analysis, survey, or study design, and can also aid in double-checking that the analysis will provide the kind of information and answers one seeks. Additionally, attaching an analysis script to the preregistration provides an

Box 2: Flexibility in how the results are analysed and used

Situation: Some researchers have reported the results of a (preregistered) survey in which they examine correlations between personality-level variables and a set of 10 items that are supposedly politically divisive (for example, climate change). The researchers thought carefully about the items and chose items they deliberately assumed to be politically divisive.

In asking for feedback on the paper, a colleague suggests that they obtain data demonstrating that the items are indeed politically divisive (i.e. rather than making that decision themselves). So, the researchers run a short survey where they simply present the items and ask people to rate whether or not the items are divisive (responses are Yes or No). They intend to look at the percentage ratings for each item to determine whether each item is sufficiently divisive. However, the results of this study have the potential to affect other results already reported: if it turns out that some of the items are not sufficiently divisive, they will need to redo some key analyses. This means the preregistered hypotheses would be invalidated and would now need to be labelled as exploratory.

Let's assume they don't have a "hypothesis" per se; that is, they have not formed a guess as to what will happen and have not written it down in a declarative statement. Let's they assume they will just interpret the results fairly (whatever that means) and report whatever the results are.

Questions: Do the researchers need a hypothesis and/or a *p*-value? Should they preregister this study?

Do they need a hypothesis? First, instead of considering whether the researchers need a hypothesis, let's consider the research question and whether it is effectively different from what a "hypothesis" is usually considered to be. The research question is effectively "Are all, some, or none of these items perceived as politically divisive?"

This research question could easily be answered with a statistical analysis such as a chi-square to examine whether the responses are equally distributed. Such a process would not change the goal of the study or how the data will be used, yet, according to traditional descriptions of preregistration, this action would make the "analysis" suitable for preregistration (because it now involves the interpretation of *p*-values).

Another option is to simply reword the research question to something like "Each of the items are politically divisive." This, in practice, is no different from a hypothesis such as "Each item will yield greater than 50% agreement that it is politically divisive." Thus, the researchers will use this information to make an inference about some quality of those statements and then to make a claim based on the data they've collected. They will also then use that information to make decisions about what to do with their data.

Is preregistration useful here? Next, let's consider whether they should or could preregister the study and, if yes, what would be preregistered. One might say "Well, this is a descriptive study. We don't need to preregister it because there's only one way to interpret the results." However, a quick glance at the list of researcher degrees of freedom (Wicherts, 2016) highlights many possible ways this data can be flexibly interpreted—note that much of this flexibility has nothing to do with p-values. Consider each of the following possibilities for flexibility:

***Option 1*:** What is considered agreement? Greater than 50%? what about 49.6% or 50.5%? This may seem like a trivial point, but it could have large implications for how the originally reported results are changed.

***Option 2*:** What do they do with the results of this survey: report it in a supplement or a footnote? Modify the results already reported in the paper? What do they do if only one of their items is not rated as politically divisive? What if 4 of their 10 items are not rated as politically divisive? One option is to later decide "Well, 9 of the 10 items were rated as strongly divisive. So, we'll just leave the results as is and report this in a supplement. If half of them were not divisive then we'd have to do something about the results." Another option is to decide to report both sets of analyses: one in a supplement and one in the main text. The problem, though, is that a decision to focus on the results which accord with their preferred outcome may be the main focus of the paper.

***Option 3*:** What about the quality of the study, itself? Consider a situation where almost all of the items come back with very low ratings: say only ~20% of the sample thinks each item is politically divisive. One could easily reason that the question was "asked wrong" or that the purpose of the survey was unclear. It's easy to think of a different way to ask the question and run this survey again. The first survey then gets discarded as a 'failed pilot.'

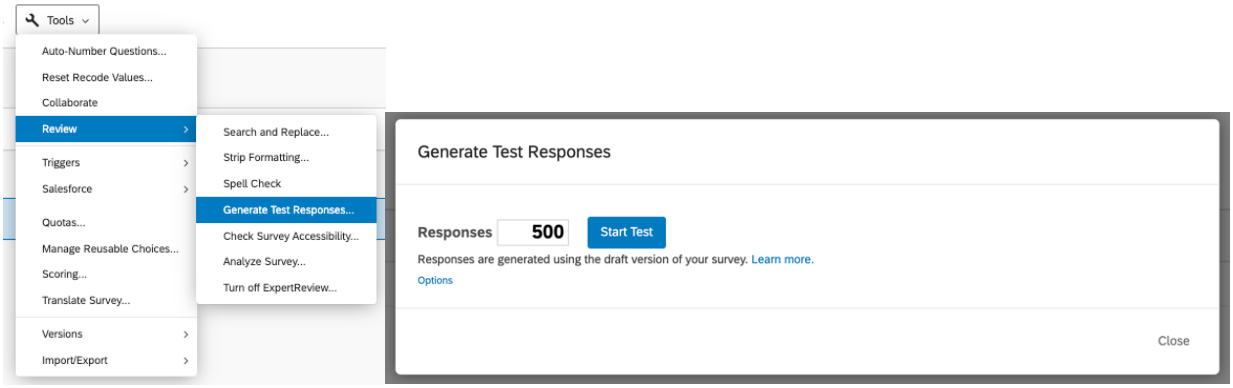
***Option 4*:** What about missing data? A survey respondent may have accidentally skipped over an item or two. Excluding this respondent listwise might push one or two of the items over the 50% barrier; excluding them pairwise may not have this effect.

Consider: The choices made by the researchers may seem obvious to them at the time, but other researchers may have made different decisions. Further, hindsight bias is strong: readers and reviewers may question why this item analysis was not done a priori but, because the researchers do not record their decisions in an open-notebook or lab log, it is unclear even to them. Finally, this is an example why it is important to preregister *how the results will be used.*

Box 3: Generating toy data using Qualtrics and R

Using Qualtrics

Generating toy data from Qualtrics is simple once a survey is already designed. The resulting "test data" can then be imported into a statistical analysis program where it can be analysed. From the menu inside a completed survey, choose 'Tools' → 'Review' → 'Generate Test Responses.' The menu will ask you to determine the number of test responses and, after a few moments, the test data can be downloaded and analysed via your preferred software.



Using R.

For those unfamiliar with R, the simple script provided below generates seven variables ($N = 100$) one might measure in a hypothetical laboratory study. These variables can be renamed, expanded, or modified to fit other study needs. Once these data are generated, the researcher can then write up an analysis script and analyse them to prepare for the actual study. The script can be copied and pasted into R or R-studio and run as-is. The analysis script can then be written as though this were real data.

```
```{r}
x <- rnorm(100) #generates a normally distributed continuous variable with M = 0, SD = 1. This could be used as an individual difference variable.

z <- rnorm(100, mean = 3, sd = 1) #generates a normally distributed continuous variable with M = 3, SD = 1. This could be used as an individual difference variable that would be different from variable x.

m <- rpois(100, 1) #generates a variable on a Poisson distribution; for example, response time.

a <- rbinom(25, size=1, prob=0.45) #generates a binomial distribution with an 45% chance of having a value of 1. This could be used for a variable like gender, where a sample might have slightly more females than males.

b <- sample(0:1, size = 100, replace = T, prob = c(.80, .20)) #generates a dichotomous variables with an 80% chance of being 0. This could be used for a variable like experimental condition.

c <- sample(c('a', 'b', 'c'), size = 100, replace = T) #generates a categorical variable with three levels. This could be used for a variable like experimental condition or a personality level variable such as political orientation.

y <- 2*x + .5*z*x + 3*m + rnorm(100) #generate an outcome variable predicted by x, z, m, and random error. This could be used as a dependent variable, as it is correlated with three variables and an interaction.

df <- data.frame(x, z, m, a, b, c, y) #combine the variables into a data frame.
```
```

unambiguous description of how the data will be treated.

Toy data can be generated easily through multiple means (see **Box 3**). Qualtrics allows users to generate "test data" after a survey is designed. This data can then be downloaded and analysed as though it were real. For those not familiar with R, a very simple R-script gives an example of how one can generate and run analyses based on toy data. This is easily modifiable based on different needs and, of course, much more complicated designs exist. For example, the `sim.multilevel()` function from the `{psych}` package (Revelle, 2019), and the `sim.icc()` function from the `{multilevel}` package (Bliese, 2016) both allow for the simulation of multilevel data. These packages also allow for generating sets of correlated variables.

Going through the motions with toy data—including writing up the analysis script, running the analysis, practicing the interpretation of the results, and generating figures and tables—can be extremely beneficial to the research process. Investigators may realize a confound or a problem, they can think about what information the study will yield and what it will not yield, or it may simply save some time later.

Further, in the event that the study does not yield a "publishable" finding, the results can be easily reported and posted on a public repository (e.g. the Open Science Framework) to avoid adding to one's file drawer. Such reports need not be overly detailed because all of the information is already present in the preregistration; a simple set of results would suffice.

Avoid preregistering underspecified "exploratory" analyses. Everything that is planned (or considered as a possibility) at the time of the preregistration should be included in the preregistration. Often, researchers may wish to preregister only their main analysis, but then include several vague statements about secondary analyses or measures. Alternatively, researchers may have other ideas for analyses and deem these "exploratory", excluding them from the preregistration.

However, this undermines the purpose of preregistration for all of the reasons detailed in the preceding sections. Most notably, poorly specified analyses and ideas also inherently carry with them a large number of RDFs—including the option to include or exclude them from the final report depending on the outcome.

To deal with this, researchers should avoid including underspecified analyses labelled as exploratory or secondary in their preregistration forms. An example of this is seen in a call for registered reports at Royal Society Open Science (Chambers, 2020), which explicitly disallows exploratory analyses in their registered report submissions (https://osf.io/93znh/).

Researchers are encouraged to either fully plan out and preregister all analyses (at which point they are no longer exploratory) or to explicitly state that there are no plans for a given measure. Further, researchers should explicitly identify which measures are being included for other purposes so that there is a record and a commitment to interpreting these as no more than exploratory.

Exploratory analyses (or modified analyses) that arise after conducting the planned analyses can still be included, but these should be verified in a preregistered replication prior to publication. If the preregistration plans are registered in a public repository (e.g., the Open Science Framework), these plans will eventually become public. This ensures that exploratory studies that did not yield favourable results cannot be hidden forever.

A preregistration should detail information on decisions made during the planning stages. There are several ways researchers can document the decisions they make during the planning stages of research and these can easily be incorporated into the preregistration process. For example, a lab notebook can be shared and notes, reasons for decisions, or future plans can be summarised. Including this information in a preregistration form is critical because readers and reviewers may ask why a particular decision was made and the researchers may have forgotten!

Answers to questions during the review process, for example, can easily yield post-hoc rationalisations that are inaccurate or only serve to defend the results as presently reported. Further, they provide opportunities for researchers to deviate from their plans in order to satisfy a reviewer and increase chances of publication. Detailing this information ahead of time can also ensure that a given study design answers the questions one is interested in or actually provides a clear test, or a *severe test* (Mayo, 1991), of a hypothesis.

A preregistration should include information on how the results will be used and interpreted. It's generally assumed that researchers have at least two hypotheses (read: "guesses"): the alternative hypothesis (e.g., "it worked!") and the null hypothesis (e.g., "it didn't work."). In reality, the results of psychological studies can support any number of conclusions about the world, regardless of whether these conclusions are based in theory or not (or, rather, conclusions can be generated to support any number of results). The reason for this is that psychological theories are too poorly developed to be able to quantify the specific situations or constraints under which a given hypothesis should be predicted, or even what specific results should be considered support for a hypothesis. (Meehl, 1967). The result is (unfortunately) not more careful predictions, but rather more liberal interpretations.

Thus, more detail is necessary in what one expects from a study, or in considering what *could possibly* come out of a study in order to facilitate planning and aid in interpretation. Below, I discuss how researchers should define their inference criteria, consider all possible outcomes of the study, and determine a priori how each of those outcomes would be interpreted.

At a minimum, a preregistration should define inference criteria. Prior to conducting a study, researchers should think carefully about what would constitute support or lack of support for an effect and define this explicitly. There are many possible ways to determine this (Dienes, 2020). At the very least, researchers should outline the specific criteria they will use to make an inference—for example, a *p*-value or an effect size. However, it's important to note that choosing a *p*-value of less than .05 is simply an indirect (and perhaps accidental) way of determining the minimum effect size (because a sample size will also be determined). Thus, there are many more informative ways of choosing inference criteria—for example, researchers could identify the smallest effect size that they are willing to consider (Anvari & Lakens, 2019; Lakens, Scheel, & Isager, 2018).

Consider a typical study in which the main product is the result of a correlation analysis. This study has three possible outcomes: the null result, a positive correlation, and a negative correlation. Importantly, there are many possible *r*-values which could constitute each outcome. If a researcher was expecting a positive correlation, for example, are they willing to accept *any* positive correlation? Values of $r = .05$, $r = .50$, and $r = .90$ are all positive correlations, yet they would yield very different interpretations, and not all of those results would be interpreted as a "success". The interpretations also depend on the constructs, the operationalisations, and the reason a positive correlation is expected in the first place.

Different people may also make different decisions on whether a specific *r*-value belongs in the "expected result" bin or in the "null result" bin. Again, such decisions also depend on the constructs in question, but researchers generally have an idea of what convincing and unconvincing evidence would look like prior to embarking on a study, and they design a study that is likely to yield a recognizable result.

One issue with neglecting to determine this criterion a priori is that researchers may (unknowingly) have different criteria at different points in time, or they may (unknowingly) change their criteria after they see the result. While this constitutes a form of bias, it need not be intentional—after all, researchers are excellent at duping themselves (Nuzzo, 2015). To avoid the ambiguity of overclaiming results or coming up with a post-hoc rationalisation of why some result was obtained, researchers should determine an evidence threshold. In the case where a result was obtained through exploratory analyses, the procedures yielding that result can be directly preregistered in a new study to examine its replicability.

At a minimum, researchers should specify the range of inference criteria by stating what set of results would constitute support and lack of support for an effect. A simple example of this is presented in **Box 4**. Of course, the outcomes of many research projects can be much more complicated; a slightly more complicated example is presented in **Box 5**. However, this is still a very manageable task and one that can provide the researcher with insights to refine the design of their experiment.

Preregistration is ideal for determining possible outcomes. Why commit to one hypothesis which, ultimately, will just be our "intellectual child" that we become attached to and want to protect (Chamberlin, 1897 p. 358-359; Platt, 1964)? A better way of dealing with ambiguity is, instead of committing to a specific hypothesis,

Box 4: Identifying Evidence Thresholds

Situation: Whenever a researcher wants to run a study, they generally begin with some limited pieces of information. The inspiration for a given study could have come from a specific theoretically-derived prediction, from a previous exploratory analysis, or it could have been thought up in a dream. No matter the source of the idea, the researcher has at least 4 pieces of information:

1. An understanding of the relevant concepts.
2. An operationalization of the concept (e.g. a measured construct).
3. An idea (or theory) guiding why those two measures should (or should not) be correlated.
4. A threshold of evidence, whether implicitly assumed or explicitly stated.

The study, itself, has at least 3 possible outcomes:

Outcome 1: A positive result, confirming the researcher's hypothesis or expectation
Outcome 2: A negative result, opposite to the researcher's hypothesis or expectation
Outcome 3: No result, or the "null" result.

Problem: There are many study "results" (e.g., *r*-values, means, effect sizes, etc) that could constitute support for each of the three possible outcomes. But, what is convincing evidence? What kind of evidence would cause one to abandon a research program? Of course, these may vary some from person to person, but a few careful considerations can yield some decisions. Those decisions can then help to limit the inferences drawn from a set of results, and can also help refine the study so that it is more (or less) likely to yield a results within the range of interest. This is especially true when attempting to replicate a result from an exploratory analysis.

Solution: The goal here is to use the information from points 1-3, above, to inform the decision in point 4 (the evidence threshold), and then to explicitly describe what positive, negative, or "null" evidence would look like. Our researcher considers the concepts, the reasoning behind the investigation, and how the concepts are operationalised and measured. Using this information, the researcher comes up with the following three categories. This information is then described in the preregistration form—in many cases, it may be useful to include graphical depictions of these results to further reduce ambiguity. The top half of Table 1 describes hypothetical results from a correlation study. The bottom half of Table 1 describes hypothetical results from a study with two conditions (Group A and Group B).

Table 1. Example of simple categorisation of possible outcomes from two hypothetical studies.

	Verbal Hypothesis	Criteria for Positive Result	No Result	Negative Result
		(Supports researcher's guess, hypothesis, etc)	(Null result, uninteresting, too small, etc)	(Opposite to the researcher's guess or hypothesis)
Correlation Study	X and Y will be positively correlated	r = .20 or greater	Within the range of r = .19 to r = -.10	r = -.10 or stronger
Two-group study	Group A will have a higher mean than Group B	d = .30 or greater AND Mean values for Group A must be above 4 (on a 7-point scale)	d = .29 or smaller OR Means for Group A are below 4 (on a 7-point scale)	Group B means are above 4 (on a 7-point scale) AND d = .20 or greater

Consider: Upon considering the above information, the researcher may have decided that the measurement and operationalisation of the concepts was not specific enough and would likely contain too much error to yield the larger effect sizes they would require. As a result, the researcher could revisit the study design and refine the measurement.

Consider also that the inference criteria need not be limited specifically to the interpretation of standardized effect sizes; it could be based on the actual interval-level values of the concept being measures (for example, heart rate or temperature), on a meaningful real-world interpretation of the measured construct (for example, the dollar amount constituting an income gap), or even on the labels used for Likert-type rating scales. There are many ways to decide on inference criteria (Anvari & Lakens, 2019) and researchers should make use of different approaches under different circumstances.

simply outline all possible outcomes and interpretations.

As just noted, there are always many possible outcomes for a study and researchers should consider what the results *could* mean. **Box 5** describes a situation where a seemingly simple study can yield many possible results and even more verbal explanations of those results.

Conventional practices would suggest that researchers observe data *and then* make sense of the data—a task which many researchers enjoy doing, no doubt. However, I argue here that these post-hoc explanations of data trends *sans theory* (and its constraints) are often useless for drawing inferences about the world. If the data reflect the real world in any way, then only a single interpretation (along with a single set of results) could be true. However, there are many reasons a set of data might not reflect the real world (e.g. biased sampling, poor measurement, etc). Researchers should have a set of constraints to guide how results are interpreted—ideally, researchers would rely on theory but, as noted, these are all but absent in the psychological sciences (Fiedler, 2018; Gervais, 2020; Meehl, 1967, 1978; Muthukrishna & Henrich, 2019).

Thus, instead of specifying a hypothesis, researchers can consider each possible outcome that could be yielded and then determine how they would interpret those outcomes. These plans can be preregistered along with the rest of their study. Doing this will achieve at least 5 things.

First, it will allow researchers to determine if the conditions, constraints, or measures included are necessary and sufficient to answer the research question. Second, it will allow researchers to include conditions or measures which would be needed to clarify ambiguity in the results. Third, it would allow researchers to remove conditions or measures that are not needed which then increases power and precision. Fourth, it forces researchers to commit to the results and explanations that *could* possibly be yielded by the study design they have set forth and prevents hypothesizing, theorizing, interpreting, and misusing results after data has been analysed. Fifth, it requires that results which do not conform to expectations would need to be verified in an additional preregistered replication study in order to determine their replicability *prior* to committing them to the publication record.

Preregistration is ideal for testing competing sets of hypotheses. More in line with what Platt (1964) had in mind, researchers could also develop a limited and specific set of competing hypotheses to test. This can be done very easily and naturally during the planning stages of research, through discussions with lab members, or while testing out analysis scripts with toy data.

Again, in the simplest form, each study has at least three outcomes: the null result, the expected result, and the opposite of the expected result. Consider a simple two-group study with Groups A and B. The null result is when the two groups means are the same (or close to the same), alternative hypothesis number one is when Group A has a higher mean, and alternative hypothesis number 2 is when Group B has a higher mean.

Psychologists can purposefully design a study with competing hypotheses made by different people (or different models, or when the researchers was in different moods on different days). The researcher would then outline the conditions that would support each hypothesis and determine how each result would be interpreted.

Another way is to state a conditional interpretation of support. For example, “In the case that at least 3 of the 5 measures demonstrate large effect sizes (e.g. d = .70 or greater), we will infer that ___. If only one or two measures show large effect sizes, it could be because of P or it could indicate that Q happened. However, if only measure X differs between the two groups, then this could possibly indicate M and the next step would be to _____.”

Another way is to state something about the description of the data: “In order to determine that the experimental condition is having no effect across the four conditions, then we would expect data in each condition to be normally distributed and the means to be around the mid-point. However, in the case that the mean score of each group increases linearly from Group A to Group D with [specific description of magnitude, effect size, Bayes factor, or alpha level], we will infer that ___.” Other descriptive claims could also be used in this way. For example, “We expect that at least 20% of our sample will yield scores higher than a 4.0 on a 7-point scale. Failure to obtain this distribution would indicate __.”

Box 5: Considering all possible outcomes of a study

Situation: Consider a hypothetical study on political attitudes: researchers wish to provide participants with articles about political topics in order to measure how people receive political information. Specifically, the researchers are interested in whether political partisans would be more receptive to a single piece of counter-attitudinal information if it were embedded in a larger, pro-attitudinal article.

The design is simple: the researchers randomly assign subjects to one of 3 conditions where they vary the ratio of counter-attitude:pro-attitude arguments in an article about a political topic. The ratio conditions are 0:4 (e.g., 0 counter-attitude arguments and 4 pro-attitude arguments), 1:3, and 2:2. After reading the article, subjects rate how "fair" they think the article was in its treatment of the political topic.

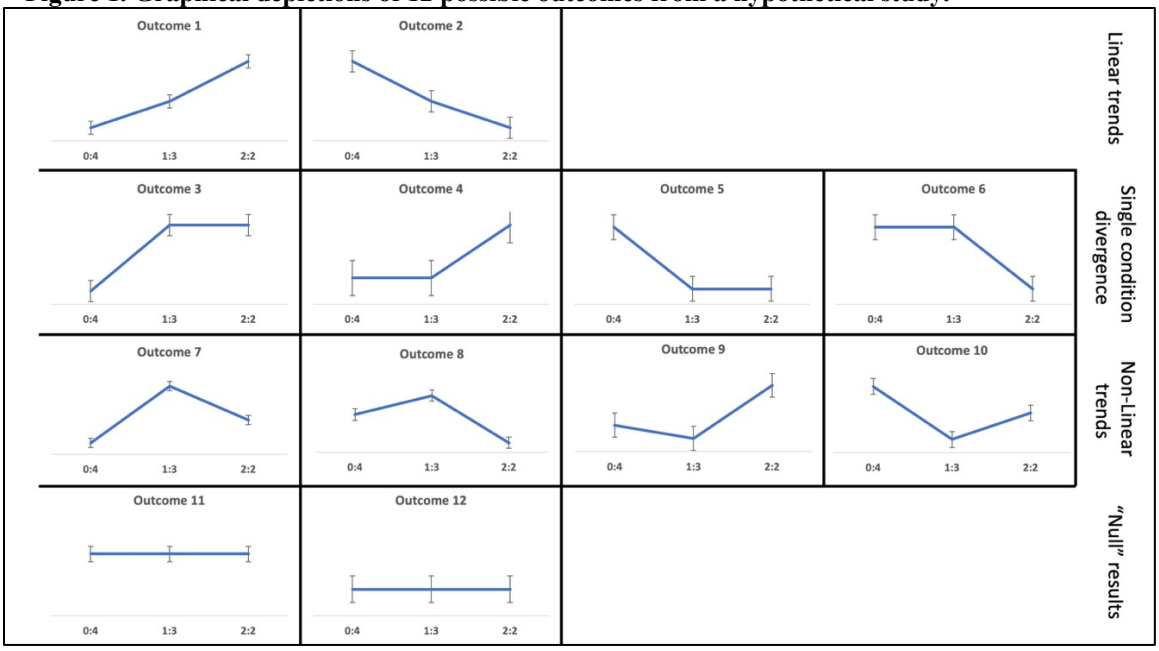
Problem: Right away, one can see that there can be many hypotheses (read: guesses) developed about which condition would yield the highest *fairness* ratings and even more explanations as to why such a result was obtained. To name a few:

- H1: The 2:2 condition should be seen as the most fair (because this is the only condition in which the arguments are actually balanced).
- H2: The 2:2 condition should be seen as the most fair by those who identify closer to mid-points of political ideology scales (because they are moderates and, thus, are less susceptible to partisan influence).
- H3: The 0:4 condition should be seen as the most fair (because people are subject to my-side bias).
- H4: The 0:4 condition would be seen as the most fair by political partisans, but not by moderates (i.e., those who score near the mid-point of a political ideology measure).
- H5: The 1:3 condition should be seen as the most fair (because people wish to think of themselves as objective, smart, and informed, and exposure to a *single* counter-attitudinal perspective reinforces this belief. However, the 2:2 condition would be too much counter-attitudinal information and would "backfire").
- And so on…

Consider: As it turns out, the researchers actually expect a result consistent with H5. However, what happens if they receive some other result? What do they do with their study—do they make sense of the findings and publish them?

One way to limit how such results would be used is to plan ahead. Our team of researchers can consider each possible outcome to determine what the results would mean and why they could occur. As an exercise to illustrate this, consider the panel of graphs below. For each outcome below, there are many possible "explanations" and interpretations. I will not go through them all here but, one can easily imagine how different people can develop different explanations. There are of course a few other ways the data could come out, and even more when one considers variance (as seen in the "null" results shown in Outcomes 11 and 12) and possible interactions (as in H2 and H4).

Figure 1. Graphical depictions of 12 possible outcomes from a hypothetical study.



These hypotheses needn't focus on statistical significance. For example, researchers could determine descriptive differences between groups (e.g., "If at least 55% of people agree…), or the effect sizes of interest ("We will select images with effect sizes greater than a Cohen's d = .30…"). Both of these strategies achieve the level of restriction and inference criteria that would be needed to make decisions based on the data.

Isn't that what hypotheses are for? Historically and, as is the case in many other scientific disciplines, the role of a hypothesis was to identify and constrain a specific set of conditions under which a prediction from a theory would be supported or not. Ideally, predictions derived from theories are context-specific, unambiguous, and detail a specific set of analyses, and point-specific results. However, psychological theories and hypotheses rarely (never?) achieve these goals. Again, psychologists are not often constrained to any specific method, calculations, or analysis procedure; we get to decide what we want to do with each new study! Thus, without specifying different sets of outcomes and what those results will mean, a set of data can be used to justify almost any claim (Gigerenzer, 1998).

Put differently, preregistration can serve the role of constraining methods, predictions, and interpretations in place of theory—but only if the study adequately constrains all possible sources of flexibility. Thinking about what the results might look under different circumstances and what each of those sets of results might mean can result in a better planned study. This can be a difficult task and, indeed, will take time. Doing so will allow one to adjust the study to avoid possible problems or to tweak the study design so as to focus more directly on the desired phenomena. Such a practice will also yield detailed information on how those results will be used.

Specify how the data will be used. Researchers should specify how the data will be interpreted or used, even in absence of a hypothesis. For example, in a pre-test or pilot situation, researchers may wish to select a sub-set of stimuli based on specific results. The results of these pre-tests may even be used in the final report to substantiate claims or assuage concerns about a set of materials. This is a situation where researchers are again making claims, decisions, and inferences based on some set of results that may lack a hypothesis or a *p*-value.

Researchers can make these decisions in the form of a decision tree, in an explicit set of If/Then statements, or by simply stating some possible outcomes. The more detail here, the better the resulting report will be because the researcher will have resolved any possible ambiguity.

In the vignette described in **Box 1**, a group of researchers pilot test a set of images in order to choose a subset of them for an experiment. In determining the selection criteria (e.g. effect sizes, percentages, significance tests, etc…) the researchers can then specify what the images will be used for. For example, a simple statement will suffice in most cases: "The images that differ significantly from the neutral set will be used in an experiment to test X, Y, and Z." Of course, having that follow-up experiment planned out would be the best course of action here, and the pre-test and confirmatory test could be rolled into one set of plans.

This prevents file-drawering "pilot" studies. Another benefit of specifying how the results will be used is that, if all studies are registered in an open and publicly available repository, unfavourable results can't be re-designated as pilot studies. Historically, researchers may run studies until they find something that works—previous unsuccessful experiments were labelled as "pilots" and were rationalized as the experimenter finding all the boundary conditions and moderators under which a specific prediction was inaccurate (e.g., Bem, 2011). Once they got that significant p-value: voila, discovery!

If all studies—exploratory, pilot, pre-testing, descriptive, qualitative, etc—are registered in a public repository, this cannot happen because eventually all registrations become public (e.g. on the Open Science Framework registrations can be embargoed for a maximum of four years). Of course, this requires that studies are preregistered in a proper registry where registrations are guaranteed to become public at some future date. In a situation where a preregistration could remain private forever, this benefit cannot exist. This also requires that reviewers actually view preregistrations as part of the review process, something which could be delegated by editors.

Summary. An effective preregistration should be exhaustive and should result from the careful

planning of a study. An effective preregistration will clearly lay out an unambiguous plan for collecting, analysing, and interpreting data. As can be seen, there is likely no easy way to address all of these pieces of information, except through careful planning, which takes time.

There are, however, many possible preregistration forms which will guide a researcher through many aspects of a study they might not have considered. Research suggests that the more detailed forms are better at restricting more degrees of freedom (Claesen et al., 2019; Veldkamp et al., 2017). However, I suggest simply thinking of the preregistration form as the methods section of your completed report.

Some things to keep in mind.

In digesting the above information, there are a few things one should keep in mind when approaching a preregistration. These may be especially important for students and for researchers who are new to the practice.

This is difficult and will take time. This is probably the most important thing to keep in mind. Planning out a study is not something that can be done quickly. It will take time, it will be difficult, and it will take practice. Employing these practices will slow down the workflow, but it will be worth it.

Preregistrations that do not satisfy all of the above requirements can be harmful. Labelling a poorly designed study as preregistered can increase confidence in the quality of the results or in the veracity of the claims. Indeed, a majority of preregistrations do not clearly identify deviations from the preregistered and most preregistrations do not adequately restrict all RDFs (Claesen et al., 2019). Further, exploratory analyses are currently accepted alongside preregistered analyses without the requirement that they are then verified in a preregistered replication attempt. This undermines the benefits of preregistration and renders them effectively the same as a non-preregistered study. If the implementation is so laissez-faire, what is the point of preregistering at all?

One reason for this may be the equivalent of statistical rituals (Gigerenzer, 2018)—that is, researchers being taught to perform a specific action rather than the reasons for that action. Many researchers may hold a misconception of what preregistration actually does (for example, allows the interpretation of *p*-values), so labelling their study as preregistered makes them more confident in the results for the wrong reasons. If psychological science is to improve, these concerns must be taken seriously and researchers must take the time to carefully plan out their studies, and journals must move towards requiring at least this minimum level of preregistration.

Mistakes are not the end of the world. There are plenty of situations where a researcher may make a mistake or may not have anticipated some outcomes. This is not the end of the world: the researcher should clearly and boldly identify these deviations. However, it's also necessary to keep in mind that preregistration goes hand-in-hand with replication: deviations and exploratory claims should be replicated if they are to be taken seriously. **Boxes 1 and 6** give some examples of simple situations in which things change after collecting data.

Additionally, a research study may yield an unanticipated set of results—this is usually very exciting, but could also cause problems for the use and interpretation of the data. Such situations need to be handled with care but, generally, complete openness and transparency will assuage any concerns. Providing readers with a roadmap of how the results were arrived at and providing unencumbered access to the data is really the only way to instil confidence in a set of results.

Finally, some research takes an incredibly long time to carry out. Even over a single year, field-wide norms and recommendations may change, rendering a preregistered analysis decision unsuitable. While such a drastic change is unlikely to occur with any regularity, the possibility exists. However, in these situations, a clear and transparent plan outlining how the research was carried out and the reasons for those decisions will be paramount in evaluating this kind of deviation from the original plan.

Changes can be made. In the ideal situation, a preregistration would be followed with no deviations. However, there are inevitably problems, changes, or decisions that may require the plan be changed. In rare situations, new data may come in which would require reinterpretation or re-analysis of previously reported results. In a situation where each of these steps and changes were noted

Box 6: What if things change after running a preregistered study?

Situation: A group of researchers have a composite of 10 items they use as a DV. They preregister a study where they create a composite of the 10 items and conduct a t-test.
Problem: During review, a reviewer raises concerns about some of the items comprising the composite and requests that the researchers examine the 10 items further. Any modifications to the analyses they already reported would constitute a deviation from the preregistered analyses, typically requiring that the analyses now be designated as exploratory.
Result: So, following the reviewer's request, they preregister a study to examine the items further. In their preregistration they include details on how the results will be used: any problematic items (via carefully specified criteria) are removed and the same, original, analysis will then be run sans the problematic items.
They run the new study as requested and some result comes out (it does not matter the result; but consider both possibilities: a) they remove some items from the composite, b) they do not remove some items from the composite).
Question: Does this project still count as preregistered?
Consider: Both studies are preregistered. There are no steps in between that are not preregistered**.** There are no deviations in either analysis plan. How does this affect the status of *pre-registered*?
Consider also that the original pre-registered analysis concerned only one test (i.e. a single p-value), but things could have easily been more complicated. For example, what if the researchers had pre-registered a trial in which a large number of participants dropped out (reducing power dramatically)? What if a decision hinges on the result of a certain analyses or assumptions (e.g., homogeneity of variances, model fit indices)? What about data "cleaning" for particularly onerous physiological recordings?

Several things could have been improved in this situation (careful planning, backup plans, etc). The researchers will probably know that certain decisions will need to be made at certain points. Researchers can anticipate these and develop a decision tree to outline those decisions and possible options. More importantly, this also requires that researchers carefully consider possible outcomes of their study, as well as possible problems.

transparently, pre-registered, and the data are made publicly available, such changes should not have an impact because readers are able to evaluate *the research* rather than *the results.* Consider the example in **Box 6** and think about whether documenting changes prior to implementing them would affect a study's preregistered status. This is also relevant to the dilemma sometimes posed: should one rigidly adhere to a preregistration at all costs or should one change their analysis to a better one?

You can preregister uncertainty. Some people may have questions about what is suitable for preregistration. This is perhaps because preregistration has been thought of as "knowing all the answers ahead of time." In reality, we often don't know the true effect size, if the model will fit, which stimuli we will select, or even how many participants will enrol in our trial. These things may also influence other decisions that will be made, such as what analysis we will use.

The solution to this problem is simple: preregister it. That is, you can specify under which circumstances a certain action will be taken. This could take the form of decision trees, a list, or even two plans—for example, "If X criteria is satisfied, we will use Plan A. If Y happens, we will use Plan B…"

Preregistration should be the rule and not the exception. There are obviously situations where preregistration and replication are not possible. For example, time-sensitive studies around natural disasters that could not possibly be anticipated, or longitudinal data collection that just so happens to flank an unexpected election outcome. Of course, researchers should still take the time to slowly and carefully plan out research, no matter the situation (Scheel, 2020): I am not arguing that time constraints are an excuse to not preregister.

However, it may be the case that a researcher realizes (at a later date or in the middle of an event) that they have data relevant to some question. Researchers may also return to previously collected data to examine results before and after some event or to examine changes within individuals over time. These are precisely the situations in which it is critical to have a preregistration that describes the circumstances under which the data were collected and its original intended use. This

would provide the opportunity for readers to understand exactly why data were collected, when and how they were collected and, then to determine if such further use of them is warranted.

Obviously, the initial data collection itself would be registered under another project, but the secondary use of such data could also be registered as long as it is done completely transparently. Though, I suggest here that these are examples of situations where preregistration may not be required because it presents an impossibility. However, this means that preregistration should be the rule, not the exception.

Evaluation should be simple. Preregistrations may be difficult to evaluate because the names of variables, the order of tests, and the workflow is not followed exactly. As a result, reviewers and readers may have to hunt through several documents located in several places (supplementary materials, websites, etc) in order to determine if a research followed a preregistered plan.

The solution to this is simple: think of your preregistration like the methods section of the research report. Preregistration is like a roadmap to be followed: anyone should be able to follow it and come to the same end result. Further, if everything—the materials, preregistration, analysis script, and data—are all located in a single repository (e.g., the Open Science Framework), this task becomes easier for the editors, authors, and readers.

A note about exploratory analyses. Exploration is critical to the process of science (Rozin, 2001). Preregistration does not prevent exploratory analyses from being conducted and such was never the intention. However, the present proposal (and the included preregistration form) puts greater restrictions on what can be considered "confirmatory" results and what can be done with "exploratory" results.

Note that the attached preregistration form can be filled out for a completely exploratory study. At the least, a researcher should create an "empty" preregistration form which identifies the included measures and states explicitly that there are no planned analyses so that there is a record of the exploration. This commits the data to being exploratory and prevents unwarranted claims.

Most importantly, a combination of four behaviours prevents exploratory analyses from being passed off as confirmatory and prevents the results of exploratory analyses from being overstated: 1) explicitly identifying measures as exploratory and positively stating that there is no analysis plan for a given measure, 2) banning the inclusion of poorly specified exploratory analyses and research questions, 3) requiring permanent and public registry (e.g. the Open Science Framework) of study plans, and 4) requiring preregistered replication of exploratory analyses.

On their own, each of these four behaviours do little. In combination, though, these four behaviours formalise the process of exploratory analysis and create a record of the research process, tying one's hands and reducing file drawer issues.

So, continue: explore, run tests, measure new things, and try new procedures. However, ensure that the results are replicable before you make claims. Upon finding some result, the entire exploratory process can then be directly replicated in a new preregistered study which meets all of the aforementioned requirements.

The value of preregistration is different for different types of studies. I am arguing that preregistration should be used for all types of studies because people can hack, misrepresent, misuse, and file-drawer descriptive results just as easily as inferential statistics. However, it is clear that the benefit of preregistration varies for different types of studies.

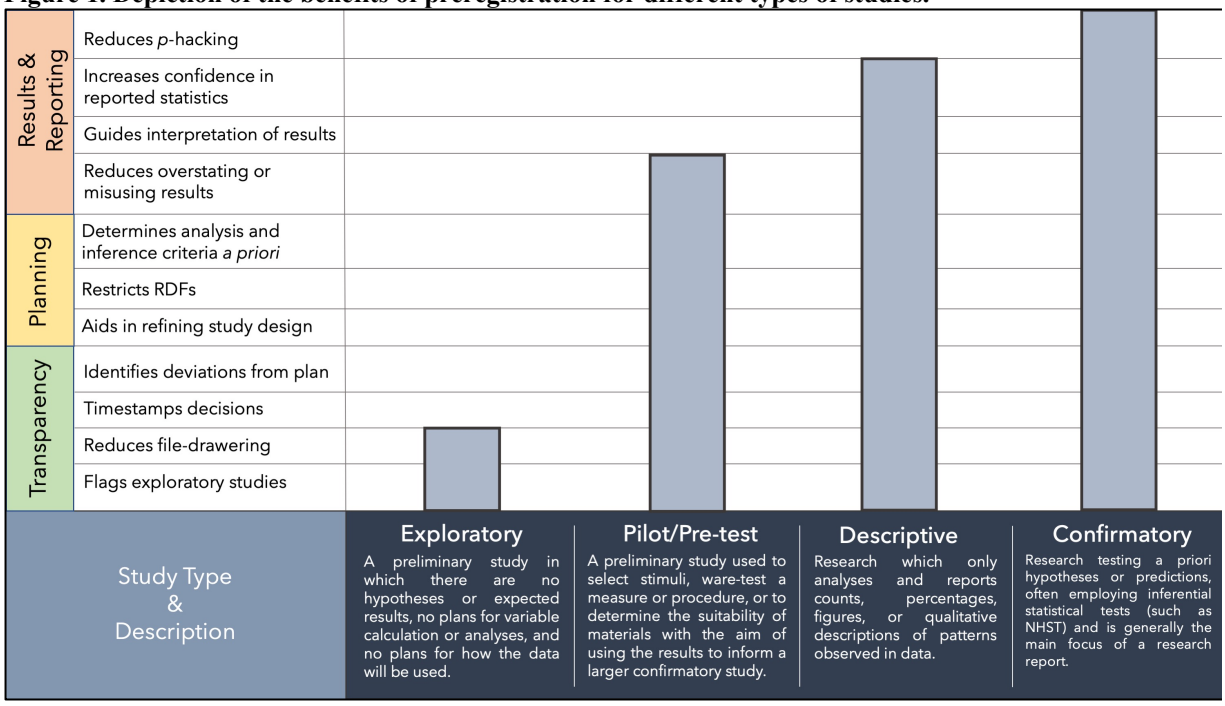
Box 7 depicts the different numbers and types of benefits achieved when preregistering different study types. For example, while preregistering exploratory studies does offer the benefits of increasing transparency by clearly flagging exploratory studies and preventing file-drawering of studies, more benefits are achieved when preregistering other types of studies (because there are more RDFs involved). Preregistering a pilot study is useful to determine analyses and interpretations of results (see **Boxes 1 and 2**). Preregistering a descriptive study can also increase confidence in how variables were calculated and the analyses that were conducted. Preregistering a confirmatory study, when RDFs are restricted as much as possible, can increase confidence in reported inferential statistics.

Box 7. Different types of studies receive different benefits from preregistration.

The benefits gained by preregistering a study come mainly from the planning aspect: planning helps one limit RDFs, detail measurements and analysis, anticipate problems and decisions, and guides the interpretation of results. Different types of studies involve different levels of planning. Thus, the specific benefits vary depending on the type of study. Of course, having only a few benefits—as with exploratory studies—does not mean one should not use preregistration. These benefits gained by preregistering exploratory studies (flagging exploratory studies as exploratory and reducing file-drawering) are as important as any of the other outcomes gained when registering other types of studies.

It's also important to note that the types of studies are defined in Figure 1, below. Of note is that an exploratory study is defined as any study without a specific a priori plan for calculation or analysis. It is important to distinguish this from a study in which there is some level of plan, but that plan is not fully realized or thought out. Using the proposed preregistration form requires that one either commits to a fully exploratory study or fully develops their analysis plan—which means it is no longer exploratory.

Figure 1. Depiction of the benefits of preregistration for different types of studies.



Some possible objections

Readers may have some objections to the things I've suggested. I've tried to anticipate some of them here and provide answers. I also hope that my suggestions continue to inspire discussion and debate around ways to improve psychological science: if people are not discussing and debating, then people do not care enough.

"Science is too slow; this makes it slower." One argument is that science already moves too slow and, by imposing requirements on people, it necessarily limits the net output of information. Put differently, more data is always better and encumbering people with requirements for planning and reporting reduces the amount of research the field generates. However, I think the answer to this concern is as follows: if one doesn't have time to carefully plan out all aspects of a study, they don't have time to do that research.

"Preregistrations are already too long; this makes them longer." Another argument is that a preregistration form should be easy and short so that people will use it and readers will read it. This is related to, but slightly different from, the previous objection. The concern here is that a difficult and complicated preregistration form may

discourage researchers from using it and reviewers from reviewing it.

However, in order to be effective, preregistration must be detailed. Science is not easy (Nosek et al., 2019); it takes time, careful planning, and consideration to yield useful products. Arguing that some process should be fast and easy does not improve the quality of information generated.

"Longer, more detailed preregistrations add responsibilities to editors and reviewers." This is true. However, rethinking the review process might be useful at this step. One possibility is that editors could assign reviewers to review certain aspects of a paper—for example, one reviewer could be assigned to assess consistency with the preregistration, one reviewer could re-run the analyses, and so on.

However, another possibility is to simply use the preregistration form as the methods section of the completed research report. This would eliminate the need to find and review two documents and would ensure consistency between the two products.

"There can be no perfect preregistration, so why even try?" "Haste makes waste" is an old proverb that especially applies here, because a carefully planned study is always better than a hasty study. Quick studies that are poorly designed or rushed are necessarily limited in the information they can provide: more information is not better if the information is incomplete, biased, or simply wrong.

"Preregistration stifles exploratory research." The concern that preregistration stifles or prevents discovery is often voiced because, after all, "shouldn't we let the data guide our decisions" (Goldin-Meadow, 2016)? However, many arguments have been made that our analysis choices and decisions should not be contingent on the data (Gelman & Loken, 2013; Landy, et al., 2018; Simmons, Nelson, & Simonsohn, 2018).

To be clear, I think that we should be asking questions and exploring and observing (Rozin, 2001). But, once we notice some pattern in the data, our first instinct should not be to report it as a "discovery." I think that our next step should be to plan out a study to examine what we think we saw and test it again with a clear and careful plan. Put differently, every new "discovery" should be replicated before it can be deemed a discovery. However, even if data did guide our decisions, I see no reason why those decisions can't be considered ahead of time and preregistered.

"Where is the line between pilot, exploratory, and confirmatory?" A researcher may run countless pilots and exploratory studies in which they measure a construct a variety of ways; some may be successful, and some may not. What should be considered as belonging to a body of evidence? Where do we draw the line between "good evidence" and "bad evidence"?

In short, think that's a decision that should be taken out of a researchers' hands. There is no clear line between pilot, exploratory, and confirmatory—even if there were, people would find a way to exploit it. Further, without *knowing all of the evidence* used to make a claim, there is no way to evaluate the quality of the claim.

So, I think the only solution is to require that everything be registered in a permanent and public repository like the Open Science Framework. This framework is ideal because studies are organized according to projects and registrations automatically become publicly available after a period of time, thus mitigating some of the file-drawer issues. Further, the hope is that researchers would run fewer higher-quality studies so as to avoid yielding evidence that is uninterpretable or flawed.

"Preregistration doesn't apply to my kind of research." One might make this argument because it seems that there is only one way to analyse the data, because one might think it would be obvious why one did what they did, or because they do the same thing every time. However, it's important to consider that what may be obvious to you may not be obvious to others. In fact, it might not even be obvious to you at a later date!

Research has shown that different people can make completely different decisions on how to test a single hypothesis (Landy et al., 2020; Silberzahn et al., 2015); the more variables and the more degrees of freedom in a study, the more options there are. Even study with the simplest of designs could be re-imagined in a more complicated fashion by another researcher. A careful, time-stamped plan of why and when decisions were made is often the only way to backup what may be later seen as arbitrary decisions.

"Registered Reports already solve these problems." Perhaps. Registered reports offer the benefit of having others evaluate and provide

feedback on your study plans prior to them being executed and also help reduce file drawer problems by guaranteeing publication of null results (Scheel, Schijen, & Lakens, 2018). However, the process depends on editors and reviewers pointing out flexibility during the review process—something that is difficult without standards in place. In contrast, a study plan preregistered on one's own will not be reviewed until after the results are obtained, making critical aspects of the study unchangeable.

However, many researchers may see registered reports as daunting and intimidating, or as suitable only for replications, risky research questions, or very large projects. While these beliefs need not be true, the concern is understandable as it is still quite a new practice in psychology. Further, those who argue that preregistering everything is already too slow are probably likely to argue that registered reports are even slower. Thus, the effective version of preregistration outlined in this article offers the best of both worlds. Although, for it to be effective, it must be implemented widely and consistently.

"A bad preregistration is better than no preregistration." It could be argued that a sloppy preregistration is better than no preregistration because it at least gives us *some* information. It allows us to see what was not planned, what wasn't considered, and it gives us something to point to in a critique. Put differently, there is also value in seeing what *wasn't* constrained prior to conducting a study.

However, we need a clear standard of what a good preregistration should contain if we are going to label preregistrations as *good* or *bad*. Further, from the perspective of one who does a "bad" preregistration, they think they've created a "good" one. One goal, I think, should be to help them improve their next preregistration instead of simply criticising a bad preregistration. The form I've provided and the guidelines laid out in the current manuscript should make it easy for a beginner to create an effective preregistration from the start.

Conclusions

If psychologists are serious about psychology as a scientific endeavour and not simply as a qualitative field with numbers (Yarkoni, 2019), then psychologists need to hold their research and their claims to a higher standard. Such changes can only come with field-wide implementation and with requirements at the journal-level. Thus, it remains the responsibility of journal editors and steering committees of professional societies—the European Association for Social Psychology, the Association for Psychological Science, and the Society for Personality and Social Psychology, for example—to implement these changes.

It may seem reasonable and fair to favour a short and simple version of preregistration. However, lukewarm suggestions such as "The pre-registration of methods, data collection plans, and data analysis plans can play a role in increasing transparent scientific practice; research that involves pre-registration is welcome… …it is not required." (Crandall, Leach, Robinson, & West, 2018) do little to increase the quality and credibility of results. Such statements may falsely signal credibility and support for transparency while avoiding any accountability and while also not contributing to field-wide changes.

To be sure, preregistration, itself, does not solve all of the problems facing psychological science. Preregistration can work as a bandage, though a bandage is no replacement for a surgeon. What the field really needs is better methods and field-wide standardised practices in order to improve psychological science. Along the way, the least we can do is carefully plan out our studies.

Declaration of Competing Interests

Jonathon McPhetres is an ambassador for the Center for Open Science and the Open Science Framework.

Acknowledgements

Thanks to Will Gervais, Roger Giner-Sorolla, Daniel Lakens, David Mellor, Thuy-vy Nguyen, Gordon Pennycook, and Simine Vazire for their helpful thoughts and comments on various versions of this manuscript.

References

Anvari, F., & Lakens, D. (2019). Using Anchor-Based Methods to Determine the Smallest Effect Size of Interest. *PsyArxiv Preprint*.

https://doi.org/https://doi.org/10.31234/osf.io/syp5a
Bem, D. J. (2011). Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect. *Journal of Personality and Social Psychology*, *100*(3), 407–425. https://doi.org/10.1037/a0021524
Bliese, P. (2016). *multilevel: Multilevel Functions*. Retrieved from https://cran.r-project.org/web/packages/multilevel/index.html
Buchanan, E., & Hvizdak, E. (2009). Online Survey Tools: Ethical and Methodological Concerns of Human Research Ethics Committees. *Journal of Empirical Research on Human Research Ethics*, *4*(2), 37–48. https://doi.org/10.1525/jer.2009.4.2.37
Buchanan, E. M., & Pavlacic, J. (2019). Hypothesize once, plan twice. *OSF PReprints*. Retrieved from https://osf.io/86vms
Campbell, L. (2018). Week 4: All about preregistration. Retrieved from Science and Psychology website: https://www.lornecampbell.org/?p=181
Chambers, C. (2020). CALLING ALL SCIENTISTS: Rapid evaluation of COVID19-related Registered Reports at Royal Society Open Science. Retrieved from NeuroChambers website: https://neurochambers.blogspot.com/2020/03/calling-all-scientists-rapid-evaluation.html
Claesen, A., Gomes, S., Tuerlinckx, F., & Wolf Vanpaemel. (2019). Preregistration: Comparing Dream to Reality. *PsyArXiv*. https://doi.org/10.31234/osf.io/d8wex
Crandall, C. S., Leach, C. W., Robinson, M., & West, T. (2018). PSPB Editorial Philosophy. *Personality and Social Psychology Bulletin*, *44*(3), 287–289. https://doi.org/10.1177/0146167217752103
DeHaven, A. C. (2017). Preregistration: A Plan, Not a Prison. Retrieved January 20, 2020, from Center for Open Science website: https://cos.io/blog/preregistration-plan-not-prison/
Fiedler, K. (2018). The Creative Cycle and the Growth of Psychological Science. *Perspectives on Psychological Science*, *13*(4), 433–438. https://doi.org/10.1177/1745691617745651
Forstmeier, W., Wagenmakers, E. J., & Parker, T. H. (2017). Detecting and avoiding likely false-positive findings – a practical guide. *Biological Reviews*, *92*(4), 1941–1968. https://doi.org/10.1111/brv.12315
Gelman, A. (2017). How is preregistration like random sampling and controlled experimentation. Retrieved from Statistical modeling, causal inference, and social science website: https://statmodeling.stat.columbia.edu/2017/03/09/preregistration-like-random-sampling-controlled-experimentation/
Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no fi shing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, COlumbia University*. Retrieved from http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf
Gervais, W. M. (2020). Practical Methodological Reform Needs Good Theory. *PsyArxiv Preprint*.
Gigerenzer, G. (1998). Surrogates for theories. *Theory & Psychology*, *8*(2), 195–204.
Gigerenzer, G. (2018). Statistical Rituals: The Replication Delusion and How We Got There. *Advances in Methods and Practices in Psychological Science*, *1*(2), 198–218. https://doi.org/10.1177/2515245918771329
Goldin-Meadow, S. (2016). Why Preregistration Makes Me Nervous. *APS Observer*, *29*(7). Retrieved from https://www.psychologicalscience.org/observer/why-preregistration-makes-me-nervous
Guest, O., & Martin, A. E. (2020). *How computational modeling can force theory building in psychological science*. 1–13.
Kellen, D. (2019). A Model Hierarchy for Psychological Science. *Computational Brain & Behavior*, *2*(3–4), 160–165. https://doi.org/10.1007/s42113-019-00037-y
Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217.
L. Haven, T., & Van Grootel, D. L. (2019). Preregistering qualitative research. *Accountability in Research*, *26*(3), 229–244. https://doi.org/10.1080/08989621.2019.1580147
Lakens, D. (2014). *Special issue article : Methods and statistics in social psychology : Re fi nements and new developments*. *700*(October 2013), 691–700.
Lakens, D. (2017). Equivalence Tests. *Social Psychological and Personality Science*, *8*(4), 355–362. https://doi.org/10.1177/1948550617697177
Lakens, D. (2019). *The Value of Preregistration for Psychological Science: A Conceptual Analysis*. 1–14.
Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/2515245918770963
Landy, J. F., Jia, M., Ding, I. L., Viganola, D., Tierney, W., Dreber, A., … Wagenmakers, Er, E. L. (2020). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. Title. *Psychological Bulletin*. Retrieved from https://doi.org/10.1037/bul0000220
Ledgerwood, A. (2018). The preregistration revolution needs to distinguish between predictions and analyses. *Proceedings of the National Academy of Sciences*, *115*(45), E10516–E10517. https://doi.org/10.1073/pnas.1812592115
Mayo, D. G. (1991). Novel evidence and severe tests. *Philosophy of Science*, *58*(4), 523–552.
Mayrhofer, M. (2017). Harking. *The International Encyclopedia of Communication Research Methods*, pp. 1–3. https://doi.org/10.1002/9781118901731.iecrm0112
Meehl, P. E. (1967). Theory-Testing in Psychology and Physics: A Methodological Paradox. *Philosophy of Science*, *34*(2), 103–115.
Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(1), 806–834. https://doi.org/10.1016/j.appsy.2004.02.001
Meehl, P. E. (1990). Appraising and Amending Theories:

The Strategy of Lakatosian Defense and Two Principles That Warrant It. *Psychological Inquiry*, *1*(2), 108–141. https://doi.org/10.1207/s15327965pli0102_1

Moore, D. A. (2016). Pre-register if you want to. *American Psychologist*, *71*(3), 238–239.

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*(3), 221–229. https://doi.org/10.1038/s41562-018-0522-1

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). *Psychology ' s Renaissance*.

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., … Vazire, S. (2019). Preregistration Is Hard, And Worthwhile. *Trends in Cognitive Sciences*, *23*(10), 815–818. https://doi.org/10.1016/j.tics.2019.07.009

Nosek, B. A., Ebersole, C. R., Dehaven, A. C., & Mellor, D. T. (2018). *The preregistration revolution*. *2017*. https://doi.org/10.1073/pnas.1708274114

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

Nuzzo, R. (2015). How scientists fool themselves – and how they can stop. *Nature*, *526*(7572). Retrieved from https://www.nature.com/news/how-scientists-fool-themselves-and-how-they-can-stop-1.18517

Revelle, W. (2019). *psych: Procedures for Psychological, Psychometric, and Personality Research.* Retrieved from https://cran.r-project.org/package=psych

Rozin, P. (2001). Social psychology and science: Some lessons from solomon asch. *Personality and Social Psychology Review*, *5*(1), 2–14. https://doi.org/10.1207/S15327957PSPR0501_1

Rubin, M. (2017). *When Does HARKing Hurt ? Identifying When Different Types of Undisclosed Post Hoc Hypothesizing Harm Scientific Progress*. *21*(4), 308–320. https://doi.org/10.1037/gpr0000128

Scheel, A. M. (2020). Crisis research, fast and slow. Retrieved April 17, 2020, from THe 100% CI website: http://www.the100.ci/2020/03/26/crisis-research-fast-and-slow/

Scheel, A. M., Schijen, M., & Lakens, D. (2018). *An excess of positive results : Comparing the standard Psychology literature with Registered Reports*. 1–14.

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis. *Annual Review of Psychology*, *69*(1), 487–510. https://doi.org/10.1146/annurev-psych-122216-011845

Silberzahn, R., Luis Uhlmann, E., Martin, D., Anselmi, P., Aust, F., C. Awtrey, E., … A. Nosek, B. (2015). Many analysts, one dataset: Making transparent how variations in analytical choices affect results Authors. *Open Science Framework*, (April), 1–56. https://doi.org/10.17605/OSF.IO/QKWST

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., … 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*. https://doi.org/10.1177/0956797611417632

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2018). *False-Positive Citations*. https://doi.org/10.1177/1745691617698146

Smaldino, P. (2019). Better methods can't make up for mediocre theory. *Nature*, *575*(7781), 9. https://doi.org/10.1038/d41586-019-03350-5

Szollosi, A., & Donkin, C. (2019). Arrested theory development: The misguided distinction between exploratory and confirmatory research. *PsyArxiv Preprint*. Retrieved from https://psyarxiv.com/suzej/

Szollosi, A., Kellen, D., Navarro, D. J., Shiffrin, R., van Rooij, I., Van Zandt, T., & Donkin, C. (2020). Is preregistration worthwhile? *PsyArxiv Preprint*. Retrieved from https://psyarxiv.com/x36pz/

van 't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology*, *67*(march 2016), 2–12. https://doi.org/10.1016/j.jesp.2016.03.004

van Lange, P. A. M. (2013). What We Should Expect From Theories in Social Psychology: Truth, Abstraction, Progress, and Applicability As Standards (TAPAS). *Personality and Social Psychology Review*, *17*(1), 40–55. https://doi.org/10.1177/1088868312453088

van Rooij, I., & Baggio, G. (2020). The theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *PsyArxiv Preprint*. https://doi.org/10.31234/osf.io/7qbpr

Veldkamp, C. L. S., Bakker, M., van Assen, M. A. L. M., Crompvoets, E. A. V, Ong, H. H., Soderberg, C. K., … Wicherts, J. M. (2017). *Restriction of opportunistic use of researcher degrees of freedom in pre-registrations on the Open Science Framework*. 1–25. Retrieved from https://psyarxiv.com/g8cjq

Wagenmakers, E., & Dutilh, G. (2016). Seven Selfish Reasons for Preregistration. *APS Observer*, *29*(9), 13–14.

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

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

Yarkoni, T. (2019). The generalizability crisis. *PsyArXiv*, 1–26.