Research Report

Busy Making Other Plans: A Simulation Study on the Effects of Deviations from Preregistrations

A. J. Vijlbrief

December 18, 2025

Supervisors

A. M. Scheel
H. I. Oberman

Word Count

2,480

Introduction

Preregistrations were popularized during the 2010s in psychology as a way to increase reproducibility (Lindsay et al., 2018). Preregistrations are documents that are published at the start of the research process, in which decisions concerning data collection and analysis are recorded. This is done with the aim of limiting the extent to which such decisions can be influenced by results.

This practice increases transparency about research choices and has become more popular, with the number of preregistrations increasing annually (Ferguson et al., 2023; Lindsay et al., 2018). However, whether the desired effect of reduced flexibility and increased reproducibility has been achieved is debated. Preregistration should lead to more credible and reproducible results (Lakens, 2019; Nosek et al., 2018; Wagenmakers et al., 2012). However, no difference has been found between preregistered and non-preregistered papers in the number of significant findings (van den Akker, van Assen, et al., 2024), and no reduction in *p*-hacking (Brodeur et al., 2024).

There are many reasons for the uncertainty of the effect of preregistration. Preregistrations are often incomplete or overly vague about their decisions (Claesen et al., 2021; Heirene et al., 2024). Researchers also deviate from their predetermined choices (Claesen et al., 2021; van den Akker, Bakker, et al., 2024). Deviations have been found in as many as 93% of preregistered papers, with up to 89% being incompletely reported (Claesen et al., 2021; Willroth & Atherton, 2024). These high deviation rates could be the reason why we see no reduction in significant findings and *p*-hacking.

The aim of this project is to identify potentially problematic deviations and investigate their effects. I first discuss the goals and problems of preregistration and the potential impact of deviations in more detail. Then, through a simulation study, I examine the effect of said deviations on type I and type II error rates.

The Problem and the Solution

The inability to reliably reproduce findings in the field of psychology is a known problem. The Open Science Collaboration (2015) found that only 36% of significant results were reproducible. One reason for this “replication crisis” is “researcher degrees of freedom” (Simmons et al., 2011). This term describes the flexibility that researchers have in decisions made during their work. This includes decisions such as how many observations to collect, which statistical model to use and which covariates to add. When these decisions are made during data collection or analysis, researchers can make opportunistic choices which steer their results towards desired outcomes. Such opportunistic choices are called *p*-hacking and can dramatically increase false-positive rates (Stefan & Schönbrodt, 2023).

The preregistration is proposed as a way to limit these types of practices. Templates are available to researchers, to inform them about what information is important to preregister. The preregistration is then uploaded to a public repository, with a time stamp of when it was published. Peer reviewers, and others, have access to the preregistration and can evaluate it before the start of the study or compare it to the final research paper.

In theory preregistration works well, if researchers are specific in their preregistration and follow it closely, this should lead to better reproducible findings. In practice, preregistrations are often imprecise and incomplete, making them less effective at restricting researcher degrees of freedom. Another common problem are deviations from preregistration (Claesen et al., 2021; van den Akker, Bakker, et al., 2024). Deviating means that the preregistered plan is not being followed during the execution of the study. Deviations occur most commonly in data collection procedures, statistical models and exclusion criteria (van den Akker, Bakker, et al., 2024). This includes differences in the total sample size, how outliers are selected and which covariates are used. These deviations could serve to further diminish the effectiveness of preregistrations, as it reintroduces researcher degrees of freedom and leaves room to revert back to *p*-hacking strategies.

Are all deviations bad?

Deviations from preregistrations occur in all different aspects of the research process, but what is not known is the effect of these deviations. Do changes from original research plans reduce research quality? Lakens (2024) theorized what the possible consequences could be. If deviations occur due to assumptions being violated or data no longer being suitable for the planned analysis, then deviating from the original research plan can have positive effects on the validity of a test. Similarly, adding additional analyses can increase the robustness. On the other hand, changes in testing also mean that a hypothesis is generally less "severely" tested. This means that a theory has not been given the proper opportunity to be falsified which can result in higher type II errors. Consequently, little can be said about whether deviations are bad, especially when the reason for the deviation is unknown.

There is no clear consensus on whether inconsistencies between preregistrations and final publications are a problem. When asked, researchers deemed deviations to be problematic to varying degrees (Willroth & Atherton, 2024). Changes to analyses and hypotheses were considered the least acceptable, whereas changing research platforms or software were deemed relatively justifiable. Whether the deviation was reported transparently and the estimated impact on the results of the study were also taken into consideration. To summarize, even though deviations from preregistration are common, a clear conclusion on their effects is still missing.

The present project

With this simulation study, I intend to close this gap in the literature and explore the effects of typical preregistration deviations on research outcomes and the quality of published results. Specifically, I aim to answer the question: 'How do deviations from preregistrations affect type I and type II error rates?' The effect of deviations is studied in three domains: sample size, outlier exclusion criteria, and the statistical model. Different deviations are simulated in each domain and compared to a baseline condition without deviations. This research question is investigated in the context of behavioral psychology and uses a linear regression as the model of interest. The effects are examined across two scenarios: a no-effect scenario, and an effect scenario, in which the main effect parameter is changed. The aim of this study is descriptive and therefore no specific hypotheses are formulated.

As a potential secondary goal, I hope to address the issue of reasons for deviations being unknown. The reason behind a deviation is important for assessing its effect. A researcher being forced to deviate due to circumstance is very different from a researcher *choosing* to deviate. When a researcher chooses to deviate, this could be because they are picking the results they prefer. Initially, the deviation conditions are simulated and the outcomes are used to assess type I and type II error as compared to the baseline condition. In the second phase, deviations are simulated but the outcomes are selectively chosen from either the baseline or the deviation condition, based on which condition has the better *p*-value. This is done in order to proxy the opportunistic selection of results by the researcher.

With these simulations I hope to shed light on the effect of deviations from preregistration on type I and type II errors, as well as the effect of picking deviations opportunistically.

Methods

Conditions

I reviewed which deviations to investigate by (1) how common they are, (2) their potential impact and (3) their justifiability. Based on this I chose deviations in the following domains: sample size, outlier exclusion criteria, and statistical model. Within the three chosen domains, multiple conditions are simulated and compared to a baseline condition without any deviations.

Sample size

Deviations in the sample size are some of the most common deviations, with consistency between preregistrations and published papers only being 28% for the exact sample size (van den Akker, Bakker, et al., 2024). Changing the sample size can inflate type II error (by decreasing power) as well as the type I error (Lakens, 2024; Simmons et al., 2011). Small deviations in sample size are common and often due to factors outside of the researcher's agency. Researchers also admit to stopping collection earlier or continuing collection longer after finding disappointing results (John et al., 2012). However, it is unknown how often this is also used as a reason to deviate from preregistrations. Based on this, the conditions in Table 1 are simulated within the sample size domain.

Outlier exclusion criteria

Literature shows that over half of published papers do not adhere to their pre-specified outlier exclusion criteria (Claesen et al., 2021; van den Akker, Bakker, et al., 2024). Vague or missing criteria in preregistrations make it harder to assess deviations and leave a lot of room for interpretation and ad hoc decisions (Heirene et al., 2024). This could potentially lead researchers to choose how to exclude outliers based on which method provides better results. Researchers themselves, however, deem outlier deviations to be relatively acceptable (Willroth & Atherton, 2024).

Lakens (2024) argues that adding additional exclusion criteria can undermine the severity of a test. This idea is corroborated by Stefan & Schönbrodt (2023), who showed that the type I error increases linearly with the number of outlier detection methods used. Furthermore, 38% of researchers admit to excluding outliers only after examining their impact on the data (John et al., 2012). If this is also the reason why many people deviate from their preregistered outlier criteria, then these deviations are important.

The most common method for identifying outliers in the social sciences is the z-score (Bakker & Wicherts, 2014; Leys et al., 2013). A minimum of three standard deviations from the mean is often maintained as the rule of thumb for excluding cases. Within linear regression, outliers are also often identified based on their influence, commonly using Cook's distance. Cook's distance excludes outliers based on how much their exclusion would influence the regression coefficients. In this study, the baseline condition will be to not exclude any datapoints, with deviation conditions based on z-score and Cook's distance (Table 1).

Statistical model

Amongst all of the aforementioned domains, the least is known about why people deviate from their preregistered statistical model. "Statistical model" is a very broad domain and includes many types of deviations, such as changes in the dependent variable, independent variable, statistical inference criteria and the actual statistical test. In this paper, “statistical model” refers to the above-mentioned types of changes, not only the statistical test and its specifications.

The broadness of the domain also means that deviations within the domain have been investigated from multiple angles. Claesen et al. (2021) found a deviation rate of 70% within the “analysis” domain, including deviations like examining additional effects, changing which model is tested and performing unregistered robustness checks. The majority of these deviations went unreported. Another study reported inconsistencies in 40% of the statistical models, deviations included the specifications of the variables, which model was tested and how variables were used (van den Akker, Bakker, et al., 2024).

As Lakens (2024) argues, at times it might be necessary to alter the statistical model. This could happen when a variable has been measured at a different level than expected or a test assumption has been violated. However, changing the statistical model can also be done opportunistically. Choosing to add an extra dependent variable because results are not yet satisfying can increase the type I error rate to 9.5%, and the addition of a covariate can increase it to 11.7% (Simmons et al., 2011). The reasoning behind the change thus becomes very important. Therefore, deviations due to failing to properly specify how variables would be operationalized in the preregistration are also seen as a larger shortcoming than model deviations due to unforeseen consequences (Willroth & Atherton, 2024). Within the statistical model, three deviation conditions are simulated: outcome switching, adding a continuous covariate, and adding a dichotomous covariate (Table 1).

Table 1: Parameter Values for Baseline Condition and Deviation Conditions per Domain

Domain	Baseline condition	Deviations conditions
Sample size	200	+5 +30 -5 -30
Outlier exclusion criteria	No outliers excluded	Exclusion based on *z*-scores Exclusion based on Cook’s distance
Statistical methods: *covariates*	No covariates	Continuous covariate Dichotomous covariate
Statistical methods: *outcome switching*	Y	Ya

Data-generating mechanism

Data is generated under two scenarios: “X has an effect on Y” and “X has no effect on Y”. The only difference in data generation between the two scenarios is in the main effect parameter, β_1. In the effect scenario, β_1 is 0.3, and in the no effect scenario, β_1 is 0. In the baseline condition, the statistical model is defined as

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where Y is the dependent variable, β_0 the intercept, β_1 the main effect, and ϵ the random error.

The population parameter values are presented in Table 2, together with the parameter values for the deviation conditions. The complete data-generating model is defined as

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 D + \epsilon,$$
$$Y_a = Y + \epsilon_a,$$

where Z and D represent a continuous and categorical covariate (respectively) and Y_a represents an alternative outcome variable based on Y with additional error ϵ_a.

As mentioned, data is generated under two scenarios: an effect scenario and a no-effect scenario. The only difference in data generation between the two scenarios is in the main effect parameter, β_1. In the effect scenario, β_1 is 0.3, and in the no-effect scenario, β_1 is 0.

In order to assess differences between outlier exclusion criteria, the data needs to include outliers. These are simulated by generating 5% of the data with increased or decreased values for ϵ.

Table 2: Parameter Values for the Data generating mechanism

Parameter	Value
Intercept	$\beta_0 = 0$
Regression coefficient	$\beta_1 = 0$ or $\beta_1 = 0.30$
Independent variable	$X \sim \mathcal{N}(\mu, \sigma^2)$
Random error	$95\% = \epsilon \sim \mathcal{N}(0, 1)$
	$2.5\% = \epsilon \sim \mathcal{N}(-2, 1)$*
	$2.5\% = \epsilon \sim \mathcal{N}(2, 1)$*
Continuous covariate regression coefficient	$\beta_2 = 0.10$
Continuous demographic variable	$Z \sim \mathcal{N}(\mu, \sigma^2)$
Dichotomous covariate regression coefficient	$\beta_3 = 0.10$
Dichotomous demographic variable	$D \sim$ Bernoulli(0.5)
Alternative outcome error	$\epsilon_a \sim \mathcal{N}(\mu, \sigma^2)$
Sample size	200

*Specific value of epsilon to be determined

Estimands

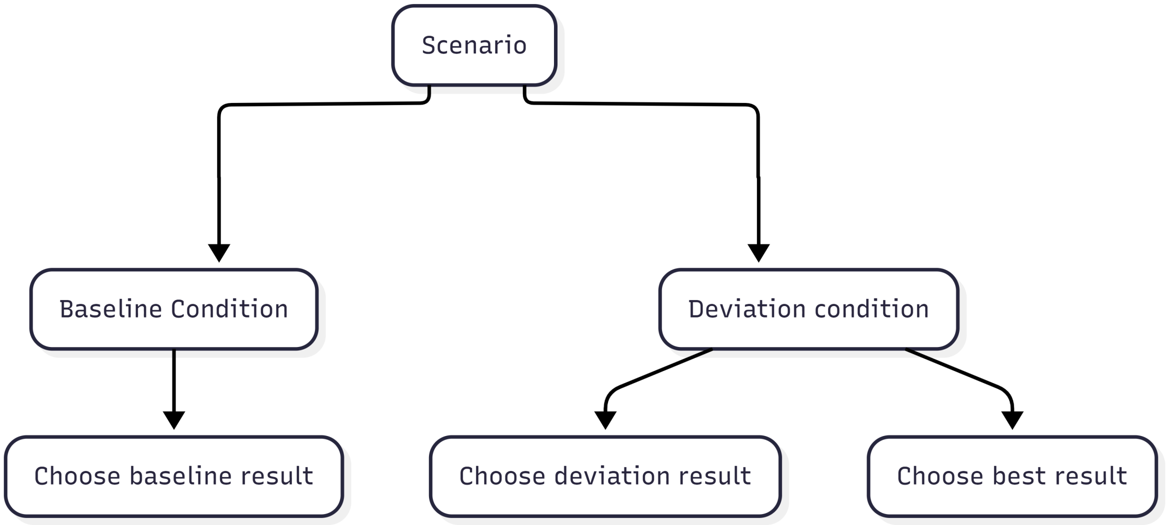
The estimand of this study is β_1, which represents the effect of the independent variable X on dependent variable Y. The coefficient is estimated using the `stats` package in `R` (R Core Team, 2024), under the baseline and deviation conditions.

Methods

As previously mentioned, there are two scenarios under which data is generated. A no-effect scenario and an effect scenario. Within these scenarios each simulated data set is examined under different conditions. The conditions include a baseline condition, with no deviations, and one condition for each possible deviation (Table 1).

In phase 1, the type I and type II error rates are compared between the baseline result and each deviation result. In phase 2, the effect of choosing whether or not to deviate is examined. This is done by simulating each deviation again, but instead of automatically reporting the deviation result, a choice is made. In each iteration the p-values of the deviation result and baseline result are compared, the most desired p-value is reported as the outcome for the deviation condition. The error rates will then be compared between the baseline results and the best values result. An overview can be found in Figure 1.

Figure 1: Condition Comparisons



Note. This figure illustrates the study phases. In phase 1, the baseline result is compared to the deviation result. In phase 2, the baseline result is compared to the best result.

Performance measures

Performance is assessed through the type I and type II error for the estimand β_1. Specifically, the regression coefficient itself, the *p*-value (for decision making in phase 2) and the confidence interval (to assess coverage) are collected for each condition in each iteration.

The type I error refers to a false-positive, or rejecting the null-hypothesis when it is true. In this study that would mean detecting an effect in the no-effect scenario. Type I error is generally acceptable at a rate of 5% or below, based on an alpha level of .05. In this case, a rate of higher than 5% is considered an inflated type I error rate.

Type II error refers to a false negative, or failing to reject the null-hypothesis when it is false. In this study that would mean failing to detect an effect in the effect scenario. Type II error is commonly assessed in the form of power. Power is $1 - \beta$, where β is the type II error. The nominal type II error rate of .20 results in a generally accepted power level of 80%. Consequently, in this study, power rates of below 80% are considered an inflated type II error rate.

References

Bakker, M., & Wicherts, J. M. (2014). Outlier Removal and the Relation with Reporting Errors and Quality of Psychological Research. *PLOS ONE*, *9*(7), e103360. https://doi.org/10.1371/journal.pone.0103360

Brodeur, A., Cook, N. M., Hartley, J. S., & Heyes, A. (2024). Do Preregistration and Preanalysis Plans Reduce *p* -Hacking and Publication Bias? Evidence from 15,992 Test Statistics and Suggestions for Improvement. *Journal of Political Economy Microeconomics*, *2*(3), 527–561. https://doi.org/10.1086/730455

Claesen, A., Gomes, S., Tuerlinckx, F., & Vanpaemel, W. (2021). Comparing dream to reality: An assessment of adherence of the first generation of preregistered studies. *Royal Society Open Science*, *8*(10), 211037. https://doi.org/10.1098/rsos.211037

Ferguson, J., Littman, R., Christensen, G., Paluck, E. L., Swanson, N., Wang, Z., Miguel, E., Birke, D., & Pezzuto, J.-H. (2023). Survey of open science practices and attitudes in the social sciences. *Nature Communications*, *14*(1), 5401. https://doi.org/10.1038/s41467-023-41111-1

Heirene, R., LaPlante, D., Louderback, E., Keen, B., Bakker, M., Serafimovska, A., & Gainsbury, S. (2024). Preregistration specificity and adherence: A review of preregistered gambling studies and cross-disciplinary comparison. *Meta-Psychology*, *8*. https://doi.org/10.15626/MP.2021.2909

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Lakens, D. (2019). *The Value of Preregistration for Psychological Science: A Conceptual Analysis*. PsyArXiv. https://doi.org/10.31234/osf.io/jbh4w

Lakens, D. (2024). When and How to Deviate From a Preregistration. *Collabra: Psychology*, *10*(1), 117094. https://doi.org/10.1525/collabra.117094

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 764–766. https://doi.org/10.1016/j.jesp.2013.03.013

Lindsay, S. D., Nosek, B. A., & Stephen, D. (2018). Preregistration Becoming the Norm in Psychological Science. *APS Observer*, *31*.

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

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

R Core Team. (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

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

Stefan, A. M., & Schönbrodt, F. D. (2023). *Big little lies: A compendium and simulation of p-hacking strategies*.

van den Akker, O. R., Bakker, M., Van Assen, M. A. L. M., Pennington, C. R., Verweij, L., Elsherif, M. M., Claesen, A., Gaillard, S. D. M., Yeung, S. K., Frankenberger, J.-L., Krautter, K., Cockcroft, J. P., Kreuer, K. S., Evans, T. R., Heppel, F. M., Schoch, S. F., Korbmacher, M., Yamada, Y., Albayrak-Aydemir, N., … Wicherts, J. M. (2024). The potential of preregistration in psychology: Assessing preregistration producibility and preregistration-study consistency. *Psychological*

Methods. https://doi.org/10.1037/met0000687

van den Akker, O. R., van Assen, M. A. L. M., Bakker, M., Elsherif, M., Wong, T. K., & Wicherts, J. M. (2024). Preregistration in practice: A comparison of preregistered and non-preregistered studies in psychology. *Behavior Research Methods*, *56*(6), 5424–5433. https://doi.org/10.3758/s13428-023-02277-0

Wagenmakers, E.-J., 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

Willroth, E. C., & Atherton, O. E. (2024). *Best Laid Plans: A Guide to Reporting Preregistration Deviations*.