



UvA-DARE (Digital Academic Repository)

Flexible Yet Fair: Blinding Analyses in Experimental Psychology

Dutilh, G.; Sarafoglou, A.; Wagenmakers, E.-J.

DOI

[10.31234/osf.io/d79r8](https://doi.org/10.31234/osf.io/d79r8)

[10.1007/s11229-019-02456-7](https://doi.org/10.1007/s11229-019-02456-7)

Publication date

2021

Document Version

Final published version

Published in

Synthese

License

CC BY

[Link to publication](#)

Citation for published version (APA):

Dutilh, G., Sarafoglou, A., & Wagenmakers, E.-J. (2021). Flexible Yet Fair: Blinding Analyses in Experimental Psychology. *Synthese*, 198(Supplement issue 23), 5745-5772.

<https://doi.org/10.31234/osf.io/d79r8>, <https://doi.org/10.1007/s11229-019-02456-7>

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (<https://dare.uva.nl>)



Flexible yet fair: blinding analyses in experimental psychology

Gilles Dutilh¹ · Alexandra Sarafoglou² · Eric-Jan Wagenmakers²

Received: 10 December 2018 / Accepted: 29 October 2019 / Published online: 19 November 2019
© The Author(s) 2019

Abstract

The replicability of findings in experimental psychology can be improved by distinguishing sharply between hypothesis-generating research and hypothesis-testing research. This distinction can be achieved by preregistration, a method that has recently attracted widespread attention. Although preregistration is fair in the sense that it inoculates researchers against hindsight bias and confirmation bias, preregistration does not allow researchers to analyze the data flexibly without the analysis being demoted to exploratory. To alleviate this concern we discuss how researchers may conduct blinded analyses (MacCoun and Perlmutter in *Nature* 526:187–189, 2015). As with preregistration, blinded analyses break the feedback loop between the analysis plan and analysis outcome, thereby preventing cherry-picking and significance seeking. However, blinded analyses retain the flexibility to account for unexpected peculiarities in the data. We discuss different methods of blinding, offer recommendations for blinding of popular experimental designs, and introduce the design for an online blinding protocol.

Keywords Replication crisis · Scientific learning · Preregistration

When extensive series of observations have to be made, as in astronomical, meteorological, or magnetical observatories, trigonometrical surveys, and extensive chemical or physical researches, it is an advantage that the numerical work should be executed by assistants who are not interested in, and are perhaps unaware of, the expected results. The record is thus rendered perfectly impartial. It may even be desirable that those who perform the purely routine work of measurement and computation should be unacquainted with the principles of the subject.

W. Stanley Jevons, 1874

This research was supported by a talent Grant from the Netherlands Organisation for Scientific Research (NWO) to AS (406-17-568).

✉ Eric-Jan Wagenmakers
EJ.Wagenmakers@gmail.com

¹ University of Basel Hospital, Basel, Switzerland

² Department of Psychology, Psychological Methods Group, University of Amsterdam, Nieuwe Achtergracht 129B, 1018XE Amsterdam, The Netherlands

In recent years, large-scale replication studies revealed what some had foreseen (e.g., Ioannidis 2005): psychological science appears to suffer from a replication rate that is alarmingly low. For instance, the Open Science Collaboration (2015) showed that out of 100 replication studies, only 39 supported the conclusions that were drawn in the original article (but see Etz and Vandekerckhove 2016; see also Camerer et al. 2018). Similarly disappointing results were obtained for specific subfields (e.g., Marsman et al. 2017; Klein et al. 2018; Nosek and Lakens 2014), and for particular effects (e.g., Eerland et al. 2016; Harris et al. 2013; Matzke et al. 2015; Meyer et al. 2015; de Molière and Harris 2016; Shanks et al. 2013; Unsworth et al. 2015; Wagenmakers et al. 2016, among many others).

One of the contributing factors to the low replication rate is that researchers generally do not have to state their data analysis plans beforehand. Consequently, the reported hypothesis tests run the risk of being misleading and unfair: the tests can be informed by the data, and when this happens the tests lose their predictive interpretation, and, with it, their statistical validity (e.g., Chambers 2017; De Groot 1956/2014; Feynman 1998; Gelman and Loken 2014; Goldacre 2009; Munafò et al. 2017; Peirce 1883, 1878; Wagenmakers et al. 2012).¹ In other words, researchers may implicitly or explicitly engage in cherry-picking and significance seeking. To break the feedback loop between analysis plan and analysis outcome, and thus prevent hindsight bias from contaminating the conclusions, it has been suggested that researchers should tie their hands and preregister their studies by providing a detailed analysis plan in advance of data collection (e.g., De Groot 1969). In this article, we argue that in addition to preregistration, *blinding* of analyses can play a crucial role in improving the replicability and productivity of psychological science (e.g., Heinrich 2003; MacCoun and Perlmutter 2015, 2017).

1 Preregistration

The preregistration of analysis plans ensures that statistical analyses are designed independently of specific data outcomes. Preregistration is an important component of the Transparency and Openness Promotion guidelines (Nosek et al. 2015; see also Munafò et al. 2017) that over 800 journals and societies currently have under consideration. Moreover, the journal *Psychological Science* has adopted a preregistration badge, and in a recent editorial, Steve Lindsay stated that “Personally, I aim never again to submit for publication a report of a study that was not preregistered.” (Lindsay 2015, p. 1827).

Preregistration comes in different forms, ranging from unreviewed preregistration, where researchers upload their plans to an online archive with time stamp, to Chris Chambers’ “Registered Report” format (Chambers 2013, 2015; Lindsay et al. 2016; van’t Veer and Giner-Sorolla 2016) which over 195 journals have now adopted.²

¹ In the frequentist paradigm, exploratory analyses introduce a multiple comparisons problem with the number of comparisons unknown; in the Bayesian paradigm, exploratory analyses are vulnerable to a double use of the data, where an informal initial update is used to select a relevant hypothesis, and a second, formal update is used to evaluate the selected hypothesis.

² For an overview see <https://osf.io/rr/>; for a discussion of various forms of preregistration see <https://osf.io/crg29/>.

This especially attractive form of preregistration allows authors to initially submit for publication their introduction and method section together with a detailed data analysis plan. After review and successful revision the authors obtain “In Principle Acceptance”, which ensures that the eventual publication of the paper does not depend on the critical outcome (but it does depend on the data being of sufficiently high quality, where “quality” is unambiguously defined up front). This way the Registered Report prevents both data-dependent analyses as well as publication bias while rewarding researchers for ideas and execution rather than outcome.

1.1 Lack of flexibility in preregistration

Preregistration is a powerful and increasingly popular method to raise the reliability of empirical results. Nevertheless, the appeal of preregistration is lessened by its lack of flexibility: once an analysis plan has been preregistered, that plan needs to be executed mechanically in order for the test to retain its confirmatory status. The advantage of such a mechanical execution is that it prevents significance seeking; the disadvantage is that it also prevents the selection of statistical models that are appropriate in light of the data. Consider the scenario where the data violate the statistical assumptions underlying the planned analyses in one or more unexpected ways. For instance, sequences of response times may show a pronounced and unanticipated fatigue effect. It is evident that modeling the fatigue effect is good statistical practice, but the presence of a fatigue effect was not foreseen in the preregistered analysis plan. This deviation from the analysis plan means that the appropriate statistical analysis has to be downgraded from confirmatory to exploratory.

Two recent high-profile examples of studies where the preregistered analyses had to be discarded due to unexpected peculiarities of the data come from the Reproducibility Project: Cancer Biology, organized by the Open Science Framework. First, Horrigan et al. (2017) investigated if a specific antibody treatment reduces the growth of tumors in mice. The results of their replication attempt could not be interpreted because the authors had to deviate from the preplanned analyses since several mice showed spontaneous tumor regressions and therefore had to be excluded from the analysis. Second, Aird et al. (2017) attempted to replicate the clinical benefit of a specific cancer treatment. The authors planned to compare medical parameters from mice that received the treatment with untreated controls by means of a paired analysis. However, due to unexpected early deaths in the control group the authors had to move from the preregistered paired analysis to an unpaired analysis.

The core problem is that preregistration does not discriminate between “significance seeking” (which is bad) and “using appropriate statistical models to account for unanticipated peculiarities in the data” (which is good). Preregistration paints these adjustments with the same brush, considering both to be data-dependent and hence exploratory.

1.2 An alternative to preregistration: blinding analyses

To address the lack of flexibility inherent to preregistration, we follow MacCoun and Perlmutter (2015) and argue for the broader adoption of a technique known as *analysis blinding*. Similar to preregistration, analysis blinding serves to prevent significance seeking and to inoculate researchers against hindsight bias and confirmation bias (e.g., Conley et al. 2006; Robertson and Kesselheim 2016). But in contrast to preregistration, for confirmatory tests analysis blinding does not prevent the selection of statistical models that are appropriate in light of the observed data. We believe that blinding, in particular when combined with preregistration, makes for an ideal procedure that allows for flexibility in analyses while retaining the virtue of truly confirmatory hypothesis testing. The remainder of this article is organized as follows. We first comment on how biases can enter the research process, and then list the different types of blinding that have been proposed to prevent this from happening. Next we propose specific implementations of blinding for popular experimental designs, and illustrate the use of blinding with a hypothetical example study that showcases how preregistration can be fruitfully combined with blinding. Finally, we propose an online registration protocol for blinding procedures.

2 How biases enter the research process

In the classic text “Experimenter Effects in Behavioral Research”, Rosenthal (1966) discusses how a researcher’s biases can influence the result of a study. On the one hand, such biases can influence the behavior of participants: through study design and task instructions, researchers may transmit and impose their own biases onto their participants. On the other hand, a researcher’s biases can also influence the conclusions that are drawn from the observations: researchers project their biases while observing and coding behavior, analyzing and interpreting data. These biases might exert their effects outside of researchers’ awareness, which makes them particularly insidious.

Researchers’ biases can harm the reliability of results at different stages of a study (Schulz and Grimes 2002; Miller and Stewart 2011). During the *data production stage*, participants can be influenced, intentionally or not, to behave according to expectations (Orne 1962). During the *measurement stage*, for example when the observed data are submitted to a coding scheme, expectations of the coders can influence the results (Hróbjartsson et al. 2012). For both stages at which biases lurk, blinding procedures have been proposed as a remedy (Barber 1976): If neither the participant, nor the experimenter knows which experimental condition is administered, biases at the data production and measurement stage can be prevented. Such *double blind* designs are the gold standard in medical trials and are recommended in the widely adopted guidelines provided by the International Council for Harmonisation (ICH).

A third stage at which biases can influence a study’s result is the *analysis stage*. Researchers may unwittingly compromise the interpretation of their results by cherry-picking among conditions, variables (Bakker et al. 2012; Bohannon 2015), variable transformations, analyses, statistical thresholds, outlier criteria, and samples sizes (i.e., optional stopping; but see Rouder 2014). Such cherry-picking can turn statistical

analysis into “a projective technique, such as the Rorschach, because the investigator can project on the data his own expectancies, desires, or biases and can pull out of the data almost any ‘finding’ he may desire” (Barber 1976, p. 20).

For the purpose of this article, it is irrelevant whether cherry-picking is performed intentionally, as we believe may be suggested by the term “p-hacking” (Bakker and Wicherts 2011; Head et al. 2015; Simmons et al. 2011), or whether researchers have the best intention but nevertheless get lost in what Gelman and Loken (2014) termed “the garden of forking paths”, where the data take control over the analysis and steer it in a direction that appears worth pursuing. Regardless of whether or not bias was introduced intentionally, its end effect is the same: an overly optimistic impression of a study’s result.

3 Preventing bias by analysis blinding

All methods of analysis blinding aim to hide the analysis outcome from the analyst. Only after the analyst has settled upon a definitive analysis plan is the outcome revealed. A blinding procedure thus requires at least two parties: a *data manager* who blinds the data and an *analyst* who designs the analyses. By interrupting the feedback loop between results and outcomes, blinding eliminates an important source of researcher bias. The unbiased nature of the blinding procedure is symbolized by Lady Justice in Fig. 1 (see Resnik and Curtis 2016 for a historical overview and a critique of the blindfold symbolism in law). A detailed cartoon of a blinding procedure is presented in the “Appendix”.

The idea of blinding analyses goes back several decades. Indeed, as early as 1957, Sainz, Bigelow, and Barwise introduced to medicine the term “triple blind” design, referring to the procedure in which the participants, the experimenter, as well as the analyst are blind to the experimental manipulations. Blinding analyses has not become as widespread as single and double blind designs, but it is commonly advocated as a tool in medical research, for example in the CONSORT reporting guidelines (Moher et al. 2010; see also Gøtzsche 1996). In nuclear and particle physics, blinding of data in the analysis stage is common practice (e.g., Dunnington 1937, p. 501; Akerib et al. 2017; Heinrich 2003; Klein and Roodman 2005). In other fields, including psychology, blinding of analyses is exceedingly rare. In experimental psychology, our literature search revealed only a handful of studies reporting blinded analyses (van Dongen-Boomsma et al. 2013; Dutilh et al. 2017; Moher et al. 2014).

4 Methods of analysis blinding

Analyses may be blinded in various ways and the selection of the appropriate blinding methodology requires careful thought. As Conley et al. (2006, p. 10) write: “[...] the goal is to hide as little information as possible while still acting against experimenter bias.” Thus, a first consideration is how much to distort a variable of interest. The distortion should be strong enough to hide any existing effects of key interest, yet small enough to still allow for sensible selection of an appropriate statistical model

Fig. 1 Lady Justice weighs the evidence in favor of each of two competing hypotheses. The blindfold symbolizes the unbiased nature of the evaluation. See Resnik and Curtis (2016) for a historical overview and a critique



(Klein and Roodman 2005). A second consideration is that some relationships in the data may need to be left entirely untouched. For example, in a regression design it is important to know about the extent to which predictors correlate. If this collinearity is not part of the hypothesis, the applied blinding procedure should leave it intact.

Below we outline different blinding procedures and discuss their advantages and limitations. We then propose the blinding methods that are most suitable for popular designs in experimental psychology. For a similar attempt in physics see Klein and Roodman (2005).

4.1 Method 1: Masking a subset

Each blinding method is illustrated using a fictitious example featuring data of students from different schools who performed a math test. For each of the students, an estimate of their IQ is also available (Fig. 2). Blinding can be achieved when the data manager

Fig. 2 Blinding method 1: masking a subset. The analysis plan is designed using a subset of the data while the remainder of the data has been masked by the data manager

ID	School	IQ	Math Score
1	A	98	78
2	B	105	101
3	C	114	81
4	A	101	139
5	B	124	149
6	C		
7	A	101	122
8	B	102	124

splits the data into a calibration set and a test set (Fig. 2), similar to what happens in the model-selection technique known as cross-validation (e.g., Browne 2000; Yarkoni and Westfall 2017). For the design of an appropriate analysis plan, the analyst is given access only to the calibration set. Once the analyst has committed to a specific analysis plan, the data manager provides the data from the test set, and the proposed analysis plan is then applied mechanically, without any adjustment, on the test set. Importantly, the final conclusions depend exclusively on the analysis outcome for the test set. A special case of this procedure is an exact replication study: the complete analysis procedure is defined by the original experiment and is applied unchanged to data from the replication attempt. The main benefit of masking a subset is that it can be executed with relatively little effort, and that it is certain to prevent any feedback from results to analysis.

One drawback of this procedure, which is shared by cross-validation as a model-selection technique, is that it is not clear how to determine the relative size of the calibration and test set. A second drawback is that, once the relative size has been decided upon, the data manager should resist the temptation to examine the two sets and ‘correct’ perceived imbalances in particular characteristics. The third and main drawback of this procedure is that the construction of the calibration set costs data. For example, a study initially thought to have sufficient power (or sufficiently high probability of finding compelling evidence, see Stefan et al. 2019), may become under-powered in light of a split-half cross-validation technique in which 50% of the data have to be sacrificed to construct the calibration set.

One way to alleviate this problem is to simulate data based on a small calibration set and design the analyses on the basis of these simulated data (Heinrich 2003). This simulation procedure, however, involves a range of non-trivial choices that might lower the representativeness of the simulated calibration set.

In conclusion, the advantage of masking a subset is particularly pronounced for the analysis of large data sets with ample opportunity for significance seeking. In such cases, the costs of subset-masking are low (as the test set will still be sufficiently large so as to draw confident conclusions) and the benefits are substantial.

4.1.1 Method 2: Adding noise

A straightforward method of blinding is where the data manager adds random noise to all values of the dependent variable. For example, in Fig. 3, noise is added to the dependent variable ‘Score’. The proper amount of noise will mask any real effects in the data, such that, when executed on the contaminated data, the blinded analysis is unlikely to show the predicted effect. After the analyst has settled on a data analysis plan, the data manager supplies the original data and the analysis plan is mechanically executed on the original data.

For this blinding method, the precise amount of noise is of critical importance. For example, consider a researcher who compares test scores of children on different schools (dummy data presented in Fig. 3). When the added noise is drawn from a uniform distribution between, say, -0.1 and 0.1 , this would not hide existing effects. On the other extreme, when the added noise is drawn from a uniform distribution between, say, -1000 and 1000 , not only would all effects be hidden, but the noise would also dramatically alter the distributional properties of the score variable. As a result, the analyst is no longer able to define a sensible outlier removal protocol or choose an appropriate transformation for the test score variable. In physics, a related blinding method is known as ‘salting’, a process that adds fake events to the data set (Akerib et al. 2017).

ID	School	IQ	Math Score	Math Score (Added Noise)
1	A	98	78	78 + 21
2	B	105	101	101 - 14
3	C	114	81	81 - 19
4	A	101	139	139 + 9
5	B	124	149	149 - 31
6	C	122	98	98 + 18
7	A	101	122	122 + 26
8	B	102	124	124 - 4

Fig. 3 Blinding method 2: adding noise. The analysis plan is designed using ‘blinded score’, a version of the dependent variable to which the data manager has added noise

ID	School	School (Masked Labels)	IQ	Math Score
1	A	Z	98	78
2	B	Y	105	101
3	C	X	114	81
4	A	Z	101	139
5	B	Y	124	149
6	C	X	122	98
7	A	Z	101	122
8	B	Y	102	124

Fig. 4 Blinding method 3: masking labels. The analysis plan is designed on a data set for which the data manager has masked the labels of the factor levels. In this example, school labels A, B, and C are replaced by Z, Y, and X, respectively

4.1.2 Method 3: Masking labels

Another straightforward method of blinding is achieved by masking or shuffling the level labels of an experimental factor. By masking the condition labels, the data stay entirely intact but the analyst does not know whether the effects she finds are in the expected direction. Figure 4 illustrates how the labels of the different schools have been masked by the data manager before the analyst is allowed to access the data.

One drawback of this method is that the analyst is still able to see whether or not there are significant effects between the cells. If the researcher prefers to find any effect over no effect, this method will not stop this bias from influencing the results. Practically, this drawback disappears when factors with many levels are studied. For example, an anthropologist who studies in which countries people are the most generous might use this blinding method and mask the country indicators.

4.1.3 Method 4: Adding cell bias and equalizing means

A relatively subtle method of blinding is to add the same random number to all observations within the same cell of an experimental design in order to “shift the answer” (Heinrich 2003). Consider the researcher who studies the difference between test scores from students on three different schools (Fig. 5). To blind the data analysis, the data manager could change the means in each condition by adding a random number to each of the observations, for instance +10 to all observations from condition A, −15 to all observations from condition B, and +3 to all observations from condition C. This implementation of cell bias leaves the distribution of test scores for the three schools intact. At the same time, the right amount of bias will obscure the differences between the groups. The analyst is unable to search for significance, because the avail-

ID	School	IQ	Math Score	Math Score (Added Bias)
1	A	98	78	78 + 10
2	B	105	101	101 - 15
3	C	114	81	81 + 3
4	A	101	139	139 + 10
5	B	124	149	149 - 15
6	C	122	98	98 + 3
7	A	101	122	122 + 10
8	B	102	124	124 - 15

Fig. 5 Blinding method 4: adding cell bias. The analysis plan is designed using 'blinded score', a version of the dependent variable to which the data manager has added cell-specific bias

able group differences may just reflect cell bias that was injected by the data manager. As for the method of adding noise, the distribution from which these cell biases are chosen is crucial. Specifically, too little noise will not blind the analyses. In addition, adding cell bias obscures only the location of the dependent variable. If the mean of the dependent variable is correlated with its spread (e.g., for a response time distribution, the standard deviation generally increases with the mean; Wagenmakers and Brown 2007), then the data analyst can use the spread to discover the hidden information about the mean.

A special way to shift the answer is by removing all effects that are present in the data set such that the mean is equal for all cells. This *equalizing of means* results in a blinded data set in which the null-hypothesis is true by construction. One advantage of the equalizing of means method is that a biased analyst cannot *p*-hack data that were blinded in this way. On the contrary, the imposed truth of the null hypothesis can serve as a sanity check: any analysis that supports an effect of the experimental manipulation here must be reconsidered.

Note that when the means of the criterion variable have been equalized, an absolute outlier exclusion rule can no longer be used for this variable. For instance, consider a situation in which response time is the criterion variable. The blinding prevents the analyst from knowing the absolute response times. This lack of knowledge makes it impossible to argue from the data that, say, 1200 ms is a good cutoff for outlier removal. Instead, the analyst needs to either formulate a relative outlier criterion, for instance removing the 1% slowest responses, or to formulate an absolute criterion based on theoretical grounds.

Still, no form of adding cell bias provides a bullet-proof solution, as the true ordering of means may sometimes be reconstructed from aspects of the data that the means are correlated with (e.g., the ordering of the standard deviations).

4.1.4 Method 5: Shuffling of variables

A versatile and effective method of blinding is to shuffle the key variables, while leaving the remaining variables untouched (Fig. 6). This procedure can be applied to both correlational and factorial designs.

Shuffling variables in a correlational design For correlation or regression analyses, both predictor or criterion variables can be shuffled. Any correlation with a shuffled variable is based on chance, which breaks the results—analysis feedback loop. When the design is bivariate, it does not matter which variable is shuffled. However, in the case of a multiple regression, it is preferable to shuffle the criterion variable, so that eventual collinearity of the predictors stays intact and can be accounted for. An example of such blinding is performed in Dutilh et al. (2017), who studied whether people’s scores on a working memory task can be predicted by response time and accuracy on a simple two-choice task. Only the criterion variable (i.e., score on a working memory test) was shuffled, whereas the collinearity of the predictor variables (i.e., response time and accuracy) could be explicitly accounted for by a cognitive model. Relative to blinding methods that add noise to observations, shuffling of a variable has the advantage that the distributional properties of the variable stay intact.

Shuffling variables in a factorial design For a factorial design, one can shuffle the predictor variable(s) across observations, that is, randomly permute the condition labels for all participants. The result is a blinded data set in which all differences between cells of the factorial design are based on chance. Shuffling factorial predictors, however, might lead to a misrepresentation of the distributional properties of the original data set. Consider the analyst who plans to perform an ANOVA comparing the

ID	School	IQ	Math Score	Math Score (Shuffled Rows)
1	A	98	78	81
2	B	105	101	139
3	C	114	81	122
4	A	101	139	78
5	B	124	149	149
6	C	122	98	101
7	A	101	122	124
8	B	102	124	98

Fig. 6 Blinding method 5: shuffle variables. The analysis plan is designed using ‘shuffled rows’, a version of the dependent variable that was shuffled by the data manager

test scores on three different schools. Assume that the test scores from the three schools are very different, but within each school, there is a highly skewed distribution of test scores, violating the normality assumptions of an ANOVA. When the analyst would model the data without blinding, he or she would rightfully decide to transform the test scores before executing the ANOVA. With the condition indicator shuffled, however, the true skew of the test scores for the individual schools might be warped by mixing the three differently skewed distributions. As a result, the analyst may recommend an inappropriate transformation.

4.1.5 Method 6: Decoy data analyses

This overarching method addresses a potential problem with many of the blinding methods described above, namely that particular aspects of the blinded data are defined by chance. Depending on chance, the blinding procedure might eliminate existing effects, induce effects where none exist, or change the direction of effects. Consequently, some of the blinded data sets may not be representative of the original data and hence provoke the stipulation of an analysis plan that is inappropriate. For example, the blinding method might have distorted the distributional shape of the data.

To address this issue, MacCoun and Perlmutter (2017) proposed a procedure that we call *decoy data analysis*.³ Here the analyst works with multiple data sets (e.g., six), one of which is the original data set. A suitable blinding method makes it impossible to reliably identify the original data set from among the blinded decoys. By working with multiple blinded data sets as well as the original data set, the analysis that is planned is certain to be appropriate for the original data set.⁴

5 Application to standard designs in experimental psychology

The various blinding procedures come with advantages and disadvantages, the relative importance of which depends on the experimental design. Below we recommend specific blinding methods for the three standard inferential situations in psychology and the social sciences more generally: regressions, contingency tables, and ANOVA designs. Our recommendations are meant as a starting point, as the specifics of the research design sometimes require tailor-made methods. Only when more studies apply blinded analysis techniques will we learn what methods (or what combination of methods) are most appropriate.

5.1 Regression designs

When one or more continuous variables are assumed to predict a criterion variable we argue that the best method of blinding is to shuffle the criterion variable Y and leave the predictors intact (i.e., method 5 ‘shuffling of variables’ above).

³ We thank an anonymous reviewer for suggesting this term.

⁴ This is why we believe it is important that the original data set is always among the set of decoys; instead, MacCoun and Perlmutter (2017) propose to let chance determine whether or not the original data set is included.

Consider again the study by Dutilh et al. (2017), who set out to test whether elementary processing speed (as measured by performance on a perceptual two-choice task) predicts working memory capacity. Performance on the perceptual task was measured by response time and accuracy, whereas working memory capacity was measured by a single composite score obtained from a battery of memory tests. The data manager blinded the analysis by shuffling the working memory capacity variable and then sent the shuffled data set to the analyst. The analyst was free to explore different ways to model the relation between response time and accuracy. The analyst was also free to account for peculiarities in the distribution of working memory capacity (e.g., eliminate outliers). Once the analyst was satisfied with the statistical procedure, he shared the intended analysis plan with the co-authors by publishing it online. The blind was then lifted and the planned procedure was applied without any changes to the original version of the working memory variable.

A similar situation often occurs in neuroscience, when researchers seek to study the correlation between behavior and particular measures of brain activation. Much like in the example above, there is a need for flexibility of analysis on one side of the regression equation. For example, when functional MRI signals are to be correlated with behavioral measures, the rich fMRI data first need to be preprocessed and compressed, and this can be done in many different plausible ways (Carp 2012; Poldrack et al. 2017). When behavior is measured with one variable such as a test score, the easiest way to blind the analyses is to shuffle this test score variable. The analyst is then free to explore different ways to process the fMRI data without unduly and unwittingly influencing the results concerning the correlation with the criterion variable.

The situation becomes only slightly more complicated when there is a need for flexibility at both sides of the regression equation. Consider for instance a study that aimed to relate activity in certain brain regions to behavior as reflected in the parameters of a cognitive model (e.g., Forstmann et al. 2008). Here, both the neural analyses and the cognitive modeling require flexibility of analysis. In this case, the solution is to keep the variables of interest intact for each participant at each side of the equation. Blinding is achieved by shuffling the case identifier, thereby destroying the connection between the brain activity and behavior. Note that another effective way to destroy this connection is to involve two analysts, one who receives the neural data, and one who receives the behavioral data; only after both analysts have completed their work are the results brought together.

5.2 Contingency tables

For fully categorical data, we again recommend to shuffle the dependent variable. For example, when studying whether class attendance ('always', 'sometimes', 'never') predicts whether students 'pass' or 'fail' a course, it is convenient to shuffle the pass/fail variable. Table 1 shows fictitious original data. Table 2 shows how this table could look after shuffling the pass/fail variable.

Note that this shuffling ensures that the margin counts are kept intact. This way the analyst is given access to the total number of students who pass and fail, and the total number of students who reported to attend class never, sometimes, and always.

Table 1 Fictitious data for students who pass or fail a course depending on their class attendance

Class attendance	Outcome		<i>Total</i>
	Pass	Fail	
'Always'	30	3	33
'Sometimes'	10	20	30
'Never'	3	6	9
Total	43	29	72

Total sample size in bold

Table 2 Fictitious data for students who pass or fail a course depending on their class attendance, with shuffled outcome variable

Class attendance	Outcome		<i>Total</i>
	Pass	Fail	
'Always'	17	16	33
'Sometimes'	19	11	30
'Never'	7	2	9
Total	43	29	72

Total sample size in bold

The analyst is then free to decide on sensible outlier removal criteria and variable transformations without having to fear that unconscious biases unduly influence the analysis outcome. For example, the analyst might want to merge two very similar categories since one of these has very few counts.

5.3 ANOVA designs

The ANOVA design is ubiquitous in experimental psychology. In the simplest scenario, the ANOVA concerns a comparison between the means of two groups (i.e., a *t*-test). For example, a researcher may seek to study whether people who hold a pen between their teeth perceive cartoons to be funnier than do people who hold the pen between their lips (Wagenmakers et al. 2016). In this design, it is essential to use a blinding technique that distorts the mean perceived funniness in each group. Distortion of the cell means could be achieved by shuffling the condition indicators (holding a pen or not).

However, because this shuffling of condition indicators will also distort the form of the within-cell distributions of the dependent variable, we propose for ANOVA designs to *equalize the cell means*. This way, the effects of interest are masked while the distribution of the data is left intact. The easiest way to equalize cell means is by setting them all to zero, i.e., subtracting the cell mean from each observation. Importantly, the coordinator does not adjust variables that are not the focus of the hypothesis, thereby allowing the analyst the freedom to use these extra variables sensibly. For example, the answers to an exit interview can be used to exclude participants.

Still, even after equalizing the means a particularly determined analyst may still try and learn the identity of the conditions (e.g., by considering the spread of the distributions), after which the resulting analyses are again susceptible to bias. Therefore, in addition to equalizing means, we recommend to shuffle the labels of the factor levels. For instance, in the pen study the analyst would not know whether a particular participant was in the 'teeth' condition or in the 'lips' condition.

6 Blinding as integral part of preregistration

Below we analyze fictitious data to illustrate the strength of blinding when combined with preregistration. The example shows how blinding can prevent a real and substantial effect from being downgraded from a confirmatory to an exploratory finding.⁵ A comprehensive assessment of the strengths and weaknesses of the different blinding methods demands a series of Monte Carlo simulations similar to those presented in MacCoun and Perlmutter (2017). In such simulations one can vary effect size, sample size, direction of effect size, and particular aspects of the design or the data. Such an in-depth assessment is beyond the scope of the present work.

6.1 Blinding and preregistration in a hypothetical research project

Consider the following hypothetical research project, preregistered but without a blinding procedure: An experimental psychologist aims to test the hypothesis that priming participants with the concept of god makes them more willing to help others (Shariff and Norenzayan 2007). More interestingly, the psychologist hypothesizes that this positive effect is attenuated by paying participants for their participation, a speculation motivated by the work of Deci et al. (1999) who suggested that monetary incentives decrease participants' intrinsic motivation. To test this hypothesis, the psychologist measures helpfulness by the amount of time that a participant voluntarily commits to perform extra tasks for the experimenter. The design has two factors with two levels each: god prime vs. no god prime and monetary reward vs. no reward. The hypothesis is defined as an *interaction*, such that the size of the boost in helpfulness due to the god prime depends on whether or not participants get paid for their participation.

In order to protect herself from publication bias in case the results turn out in favor of the null-hypothesis, the psychologist submits her proposal as a Registered Report (<https://osf.io/rr/>). The preregistration protocol includes a sampling plan (i.e., testing 50 participants in each cell of the design for a total of 200 participants) and an analysis plan (i.e., a two-factor ANOVA, where the dependent variable is the time that participants voluntarily commit). The protocol states that the hypothesis is said to be supported when the ANOVA interaction of god prime and payment condition shows a p value lower than .05. Similar studies in the past had performed the exact same analysis. The preregistered proposal is reviewed and eventually accepted by the editor, who rewards the psychologist with “In Principle Acceptance” conditional on a data quality check involving a significant ($p < .05$) main effect of the god-prime manipulation.⁶

The collected data are depicted in the four left-most histograms of Fig. 7 (simulated data). The researcher notices that the data are heavily skewed but is required to execute the preregistered analysis plan, which produces the ANOVA output shown in Table 3.

The psychologist is disappointed to find that none of the effects are significant at the .05 level. She decides to deviate from the preregistered analysis plan and log-transform

⁵ The .jasp file for this example is accessible via: <https://osf.io/p7nxx/>.

⁶ Both the psychologist and the editor are unaware that the ANOVA harbors a hidden multiplicity problem, as explained in Cramer et al. (2016).

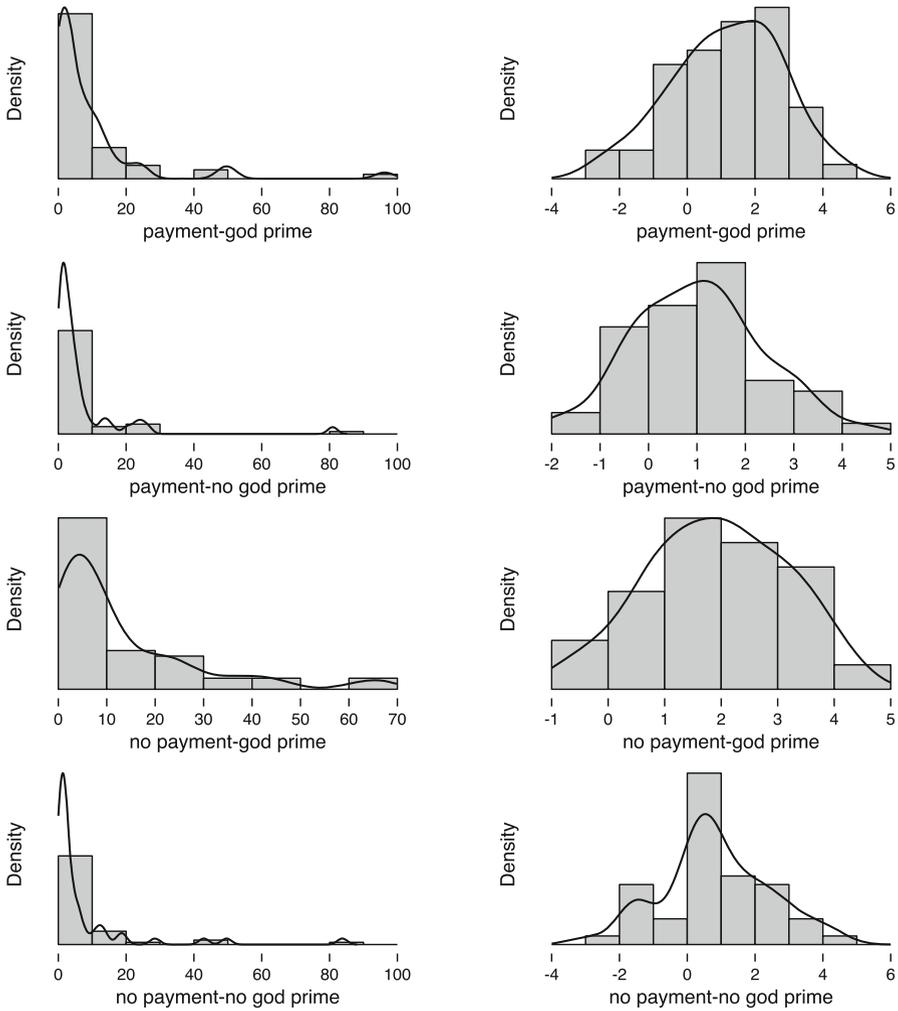


Fig. 7 Raw data (left hand panels) and log-transformed data (right hand panels). For each payment × god-prime condition, the raw data are heavily skewed. After a log-transformation, the distributions appear approximately normal

Table 3 ANOVA results from a hypothetical registered report investigating the interaction between god-priming and monetary reward on willingness to help

Cases	Sum of squares	df	Mean square	F	p
Payment	270.657	1	270.657	1.214	0.272
Godprime	847.498	1	847.498	3.801	0.053
Payment × godprime	106.296	1	106.296	0.477	0.491
Residual	43700.868	196	222.964		

Note. ANOVA table based on the raw data from Fig. 7

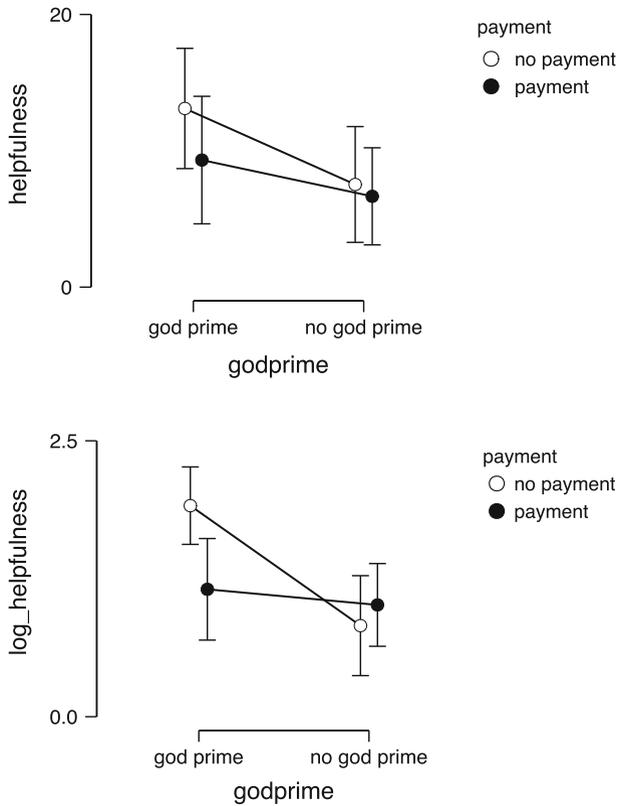


Fig. 8 Means and 95% confidence intervals based on the raw data (upper panel) and log-transformed data (lower panel)

Table 4 ANOVA results based on the log-transformed data

Cases	Sum of squares	df	Mean square	F	p
Payment	4.065	1	4.065	1.931	0.166
Godprime	18.865	1	18.865	8.959	0.003
Payment × godprime	11.183	1	11.183	5.311	0.022
Residual	412.709	196	2.106		

ANOVA table based on data from Fig. 7 after a log-transformation

the dependent variable (i.e., the number of minutes volunteered) to account for the obvious skew. The right four histograms of Fig. 7 show that the transformation indeed removed the skew of the dependent variable. The means and their 95% confidence intervals for both the raw and the log-transformed data are shown in Fig. 8. The ANOVA on the log-transformed data leads to the result in the Table 4.

The ANOVA on the transformed data shows the effect precisely as predicted: the data quality check is met (i.e., there is a significant main effect of the god-prime

manipulation with $p = .0031$) and, more importantly, there is a significant interaction between the payment and god-prime manipulations (i.e., $p = .0222$) in the expected direction, supporting the hypothesis that the payment reduces the effect of the god prime on helpfulness.⁷

The psychologist now has a serious problem: the analysis with the log-transformation cannot be reported as a preregistered confirmatory analysis. Instead, in the results section of the preregistered study, the researcher has to conclude that the confirmatory test did not show support for the hypothesis. In the section “exploratory results”, the ANOVA on the log-transformation is reported with the encouragement to carry out a confirmatory test of this hypothesis in future work.

This unsatisfactory course of events could have been prevented by adding a blinding protocol to the preregistered analyses plan. The blinding protocol could outline that in the blinded data set the cell means will be equalized and the condition labels will be shuffled, as recommended above. This blinding procedure would have enabled the analyst to observe the extreme right skew of the data, while offering the flexibility to explore several transformations before settling on a definite analysis plan. Because these transformations were all applied on a properly blinded version of the data, the results could still have been presented as truly confirmatory.

7 Incentivizing blinding

The example above illustrates that, in addition to guarding against bias, blinding can also prevent confirmatory findings from being demoted to exploratory findings. Thus, for preregistered studies there is a strong incentive to include a blinding protocol. In studies that are not preregistered, however, the incentive to apply blinding may not be readily apparent, and the blinding procedure itself may seem relatively involved. We now present the structure of an online blinding protocol that facilitates blinded analyses.

7.1 Online blinding protocol

We propose an online blinding protocol that allows researchers to declare the blinding procedure they follow and receive a certificate. The protocol kills two birds with one stone: The protocol serves science by making explicit the difference between exploration and confirmation, and it serves the scientist by convincing editors, reviewers, and colleagues that the reported results are untainted by bias.

The protocol is easy to follow. The only requirement is that there are two actors involved: (1) the data manager, who has the access to the original data set and (2) the analyst, who designs the analyses and has no direct access to the original data set. The protocol consists of three steps, preferably preceded by a preregistration step (see Fig. 9).

⁷ The psychologist is inconvenienced by recent arguments that p -values higher than .005 provide only suggestive evidence (Benjamin et al. 2018) and therefore decides that these arguments are not compelling and can best be ignored.

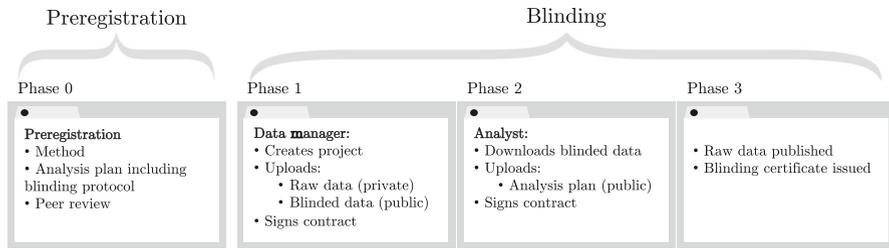


Fig. 9 An online protocol for data blinding. In three steps, the blinding procedure is standardized and registered. The full transparency is awarded with a blinding certificate

Phase 0 In case of a Registered Report, the study’s method and analysis plan are preregistered, including the blinding protocol. All of this is peer-reviewed and revised if required.

Phase 1 The data manager creates a project on the online blinding portal. She uploads the raw data that is kept private until Phase 3. Then, she uploads the blinded data, which become publicly available. By signing an online contract, she declares that she did not share the raw data or sensitive information about the blinding procedure with the analyst.

Phase 2 The analyst downloads the blinded data and designs the analyses based on that data. When she is satisfied with the analysis plan, she uploads it to the online blinding portal. By signing an online contract, she declares that she did not have access to the raw data or sensitive information about the blinding procedure when designing the analyses.

Phase 3 The raw data are revealed to the analyst, so that she can apply her analysis plan. If both data manager and analyst agree, at this point the data are also made publicly available. A blinding certificate is issued in the form of an online document that describes the blinding procedure followed. The data manager and author can include a link to this certificate in the manuscript that reports the results.

Thanks to the blinding certificate, the reader of the eventual report can trust that specific analyses were performed without knowledge of the outcome. This is important, since only when an article precisely reports how analyses are performed (as advocated in Moher et al. 2010), can the results be interpreted appropriately. Sadly, the reader of psychological articles is currently most often blind to which analyses, data transformations, and exclusion criteria have been tried before the authors settled upon the reported analyses. Thus, the online blinding protocol presents a promising opportunity to unblind the reader.

Signing the contract does of course not prevent cheating. In principle, researchers could sign the contract, but still perform all the hazardous post-hoc practices described earlier. Thanks to the signing of the contract, however, these questionable research practices are now clearly classified as scientific misconduct and cannot be engaged in unwittingly.

8 Discussion and conclusion

This article advocates analysis blinding and proposes to add it to the toolbox of improved practices and standards that are currently revolutionizing psychology. We believe that the blinding of analyses is an intuitive procedure that can be achieved in almost any study. The method is of particular interest for preregistered studies, because blinding keeps intact the virtue of true confirmatory hypothesis testing, while offering the flexibility that an analyst needs to account for peculiarities of a data set. Nonetheless, for studies that are not preregistered, blinding analyses can substantially improve the rigor and reliability of experimental findings. Blinding allows a sharp distinction between exploratory and confirmatory analyses, while allowing the analyst almost complete flexibility in selecting appropriate statistical models.

We are aware that, as any other method to improve the reliability of science, blinding is not a silver-bullet solution. One problem that is not overcome by blinding alone is that a single analyst will generally select a single analysis method and report the results from a single model. This selection of analyst, method, and model means that a lot of uncertainty remains hidden below the surface. To reveal this uncertainty one may use a *many-analyst* approach involving a number of independent analysis teams (Silberzahn et al. 2018), a *multiverse analysis* assessing the robustness of the statistical conclusions to data selection choices (Steege et al. 2016), and a *multi-model* approach that applies different statistical models, possibly combined using model-averaging (e.g., Burnham and Anderson 2002; Hoeting et al. 1999). Regardless, whenever blinding is performed, this has to be done honestly and accurately.

8.1 Honesty

First of all, blinding brings the acclaimed virtues only when performed honestly. Claiming that you have performed analyses blinded although you peeked at the data is of course highly questionable. It is, however, not always easy to abstain from doing so. Data is most often collected in the same lab as where it is analyzed. A discussion during lunch between the data manager and the analyst might supply the analyst with information he or she did not want to know. As a partial solution to this problem, the proposed online registration of a blinding protocol increases the awareness of sticking to the rules.

8.2 Errors in analysis discovered after unblinding

Another potential problem that can occur when data analyses are designed blindly, is that they simply turn out not work when the blind is lifted. For example, in spite of a careful choice of the blinding method, the analyses turn out not to be able to account for a crucial property of the data, e.g., bimodality of a variable's distribution. Also, it is possible that simple coding errors are only discovered after the blind is lifted. Such mistakes are frustrating: the analyses cannot be interpreted as purely confirmatory anymore.

When analyses turn out not to work on the unblinded data, there are two possible solutions. First, one can simply describe what went wrong and include a description of both the planned and the corrected analyses in the manuscript. Another solution is to go one step further and get a second analyst involved and repeat the blinding procedure.

We want to stress that without blinding, the chances of ending up with exploratory analyses is much larger. Researchers often try a number of analyses on the real data before settling on the analysis to be described in the eventual article (John et al. 2012). The analyses they eventually present should often be labeled exploratory.

8.3 Can blinding really improve reproducibility?

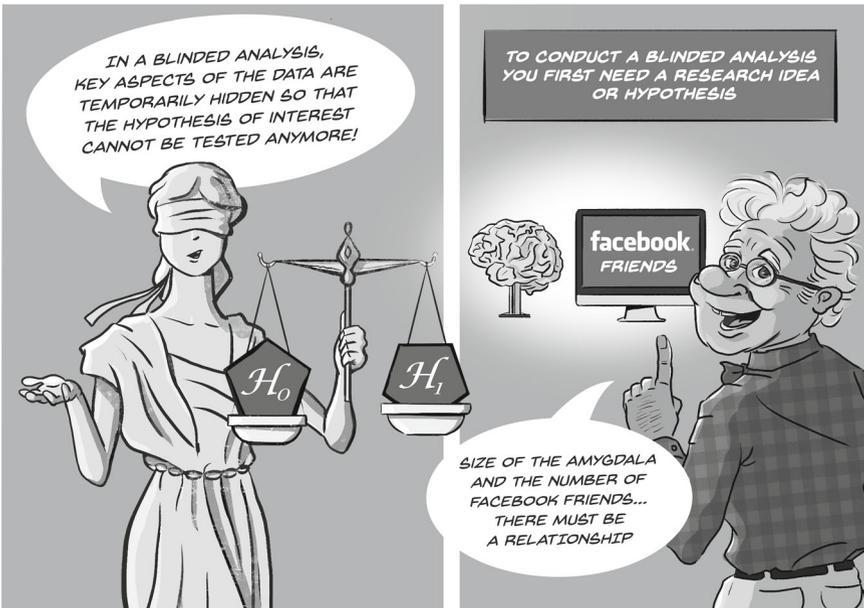
As noted above, analysis blinding is already being employed in other fields such as medicine and physics. Meta studies have revealed that experiments in which a blinding technique is applied show on average fewer positive results (Hróbjartsson et al. 2014) and report smaller effect sizes (Holman et al. 2015; Bello et al. 2014) than experiments without a blinding procedure, although work that focuses specifically on blinded analyses (instead of blinded data collection) is rare. These results should be viewed in relation to the reproducibility project by Open Science Collaboration (2015), who reported that only 39% of the effects found in the original articles were qualitatively replicated, and that the average effect size of the replication studies was about half as large as the average effect sizes reported in the original studies. In addition, a recent study by Allen and Mehler (2019, their Fig. 1) showed that the percentage of null findings was dramatically higher for Registered Reports than for standard empirical research. These and other results suggest—but do not establish—that blinding (and other procedures to tighten the methodological screws) can increase the reproducibility and reliability of results reported in experimental psychology. Empirical research on the effects of blinded analysis would be highly desirable.

8.4 Get excited again

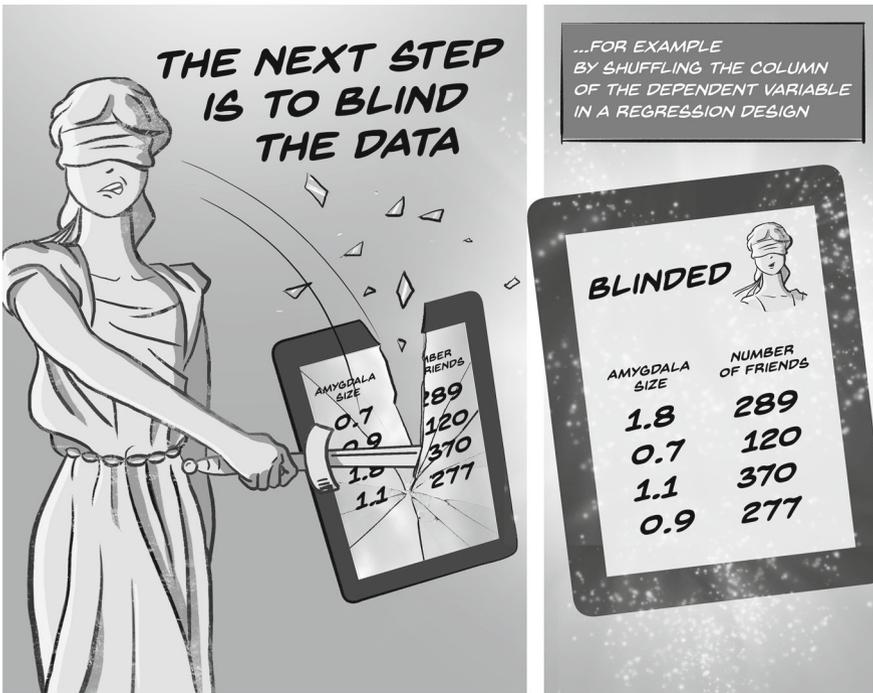
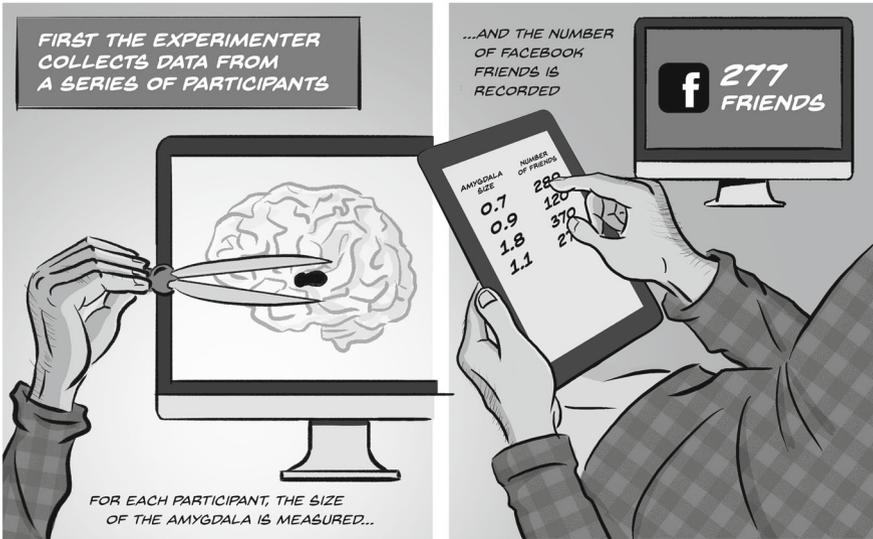
We want to finish our plea for analysis blinding on a personal note. We know many students and colleagues who analyze their data inside and out. So much time is spent on the analysis, iterating between analysis and outcome, that the eventual results can hardly be called exciting. We ourselves have had this experience too. Now, having used a blinding protocol in our own work, we have experienced how blinding can bring back the excitement in research. Once you have settled on a particular set of analyses, lifting the blind is an exciting event—it can reveal the extent to which the data support or undermine your hypotheses, without having to worry about whether the analysis was either biased or inappropriate.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

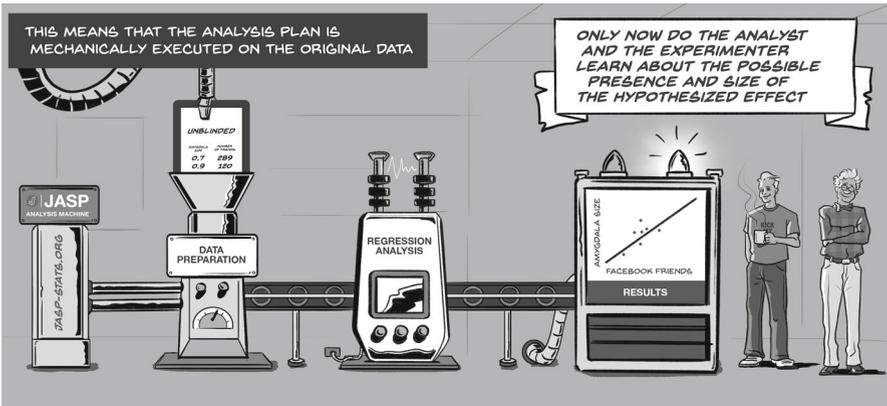
Appendix: A cartoon to explain how blinding works



Artwork by Viktor Beekman - [instagram.com/viktordepictor](https://www.instagram.com/viktordepictor)



Artwork by Viktor Beekman - [instagram.com/viktordepictor](https://www.instagram.com/viktordepictor)



Artwork by Viktor Beekman - [instagram.com/viktordepictor](https://www.instagram.com/viktordepictor)

References

- Aird, F., Kandela, I., Mantis, C., et al. (2017). Replication study: BET bromodomain inhibition as a therapeutic strategy to target c-myc. *Elife*, 6, e21253.
- Akerib, D. S., Alsum, S., Araújo, H. M., Bai, X., Bailey, A. J., Balajthy, J., et al. (2017). Results from a search for dark matter in the complete lux exposure. *Physical Review Letters*, 118, 021303.
- Allen, C., & Mehler, D. M. A. (2019). Open science challenges, benefits and tips in early career and beyond. *PLOS Biology*, 17, e3000246.
- Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, 7, 543–554.
- Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, 43, 666–678.
- Barber, T. X. (1976). *Pitfalls in human research: Ten pivotal points*. New York: Pergamon Press Inc.
- Bello, S., Krogsbøll, L. T., Gruber, J., Zhao, Z. J., Fischer, D., & Hróbjartsson, A. (2014). Lack of blinding of outcome assessors in animal model experiments implies risk of bias. *Journal of Clinical Epidemiology*, 67(9), 973–983.
- Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., et al. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2, 6–10.
- Bohannon, J. (2015). I fooled millions into thinking chocolate helps weight loss. Here's how. (Blog No. May 27). <http://io9.com/i-fooled-millions-into-thinking-chocolate-helps-weight-1707251800>.
- Browne, M. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, 44, 108–132.
- Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York: Springer.
- Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., et al. (2018). Evaluating replicability of social science experiments in nature and science. *Nature Human Behaviour*, 2, 637–644.
- Carp, J. (2012). On the plurality of (methodological) worlds: Estimating the analytic flexibility of fMRI experiments. *Frontiers in Neuroscience*, 6, 149.
- Chambers, C. D. (2013). Registered reports: A new publishing initiative at cortex. *Cortex*, 49, 609–610.
- Chambers, C. D. (2015). Ten reasons why journals must review manuscripts before results are known. *Addiction*, 110, 10–11.
- Chambers, C. D. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton: Princeton University Press.
- Conley, A., Goldhaber, G., Wan, L., Aldering, G., Amanullah, R., & Commins, E. D. (2006). The Supernova Cosmology Project. Measurement of ω_m , ω_λ from a blind analysis of type Ia supernovae with CMAGIC: Using color information to verify the acceleration of the universe. *The Astrophysical Journal*, 644, 1–20.
- Cramer, A. O. J., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R. P. P. P., et al. (2016). Hidden multiplicity in multiway ANOVA: Prevalence, consequences, and remedies. *Psychonomic Bulletin and Review*, 23, 640–647.
- De Groot, A. D. (2014). The meaning of “significance” for different types of research. Translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han L. J. van der Maas. *Acta Psychologica*, 148, 188–194.
- De Groot, A. D. (1969). *Methodology: Foundations of inference and research in the behavioral sciences*. The Hague: Mouton.
- de Molière, L., & Harris, A. J. L. (2016). Conceptual and direct replications fail to support the stake-likelihood hypothesis as an explanation for the interdependence of utility and likelihood judgments. *Journal of Experimental Psychology: General*, 145, e13.
- Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, 125, 627–668.
- Dunnington, F. G. (1937). A determination of e/m for an electron by a new deflection method. II. *Physical Review*, 52, 475–501.
- Dutilh, G., Vandekerckhove, J., Ly, A., Matzke, D., Pedroni, A., Frey, R., et al. (2017). A test of the diffusion model explanation for the worst performance rule using preregistration and blinding. *Attention, Perception, & Psychophysics*, 79, 713–725.

- Eerland, A., Sherrill, A. M., Magliano, J. P., & Zwaan, R. A. (2016). Registered replication report: Hart & Albarraçín (2011). *Perspectives on Psychological Science*, *11*(1), 158–171.
- Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PLoS One*, *11*, e0149794.
- Feynman, R. (1998). *The meaning of it all: Thoughts of a citizen-scientist*. New York: Perseus Books, Reading, MA.
- Forstmann, B. U., Dutilh, G., Brown, S. D., Neumann, J., von Cramon, D. Y., Ridderinkhof, K. R., et al. (2008). Striatum and pre-SMA facilitate decision-making under time pressure. *Proceedings of the National Academy of Sciences*, *105*, 17538–17542.
- Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*, 460–465.
- Goldacre, B. (2009). *Bad science*. London: Fourth Estate.
- Götzsche, P. C. (1996). Blinding during data analysis and writing of manuscripts. *Controlled Clinical Trials*, *17*, 285–290.
- Harris, C. R., Coburn, N., Rohrer, D., & Pashler, H. (2013). Two failures to replicate high-performance-goal priming effects. *PLoS One*, *8*, e72467.
- Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, *13*, e1002106.
- Heinrich, J. G. (2003). Benefits of blind analysis techniques. Unpublished manuscript. Retrieved November 14, 2019 from https://www-cdf.fnal.gov/physics/statistics/notes/cdf6576_blind.pdf.
- Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*, 382–417.
- Holman, L., Head, M. L., Lanfear, R., & Jennions, M. D. (2015). Evidence of experimental bias in the life sciences: We need blind data recording. *PLOS Biology*, *13*, e1002190.
- Horrigan, S. K., Courville, P., Sampey, D., Zhou, F., Cai, S., et al. (2017). Replication study: Melanoma genome sequencing reveals frequent prex2 mutations. *Elife*, *6*, e21634.
- Hróbjartsson, A., Thomsen, A. S. S., Emanuelsson, F., Tendal, B., Hilden, J., Boutron, I., et al. (2012). Observer bias in randomised clinical trials with binary outcomes: Systematic review of trials with both blinded and non-blinded outcome assessors. *BMJ*, *344*, e1119.
- Hróbjartsson, A., Thomsen, A. S. S., Emanuelsson, F., Tendal, B., Rasmussen, J. V., Hilden, J., et al. (2014). Observer bias in randomized clinical trials with time-to-event outcomes: Systematic review of trials with both blinded and non-blinded outcome assessors. *International Journal of Epidemiology*, *43*, 937–948.
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, 696–701.
- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524–532.
- Klein, J. R., & Roodman, A. (2005). Blind analysis in nuclear and particle physics. *Annual Review of Nuclear and Particle Science*, *55*, 141–163.
- Klein, R., Vianello, M., Hasselman, F., Adams, B., Adams, R., Alper, S., et al. (2018). Many labs 2: Investigating variation in replicability across sample and setting. *Advances in Methods and Practices in Psychological Science*, *1*, 443–490.
- Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science*, *26*, 1827–1832.
- Lindsay, D. S., Simons, D. J., & Lilienfeld, S. O. (2016). Research preregistration 101. *APS Observer*, *29*(10), 14–16.
- MacCoun, R., & Perlmutter, S. (2015). Hide results to seek the truth. *Nature*, *526*, 187–189.
- MacCoun, R., & Perlmutter, S. (2017). Blind analysis as a correction for confirmatory bias in physics and in psychology. In S. O. Lilienfeld & I. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions* (pp. 297–321). Hoboken: Wiley.
- Marsman, M., Schönbrodt, F., Morey, R. D., Yao, Y., Gelman, A., & Wagenmakers, E.-J. (2017). A Bayesian bird's eye view of "replications of important results in social psychology". *Royal Society Open Science*, *4*, 160426.
- Matzke, D., Nieuwenhuis, S., van Rijn, H., Slagter, H. A., van der Molen, M. W., & Wagenmakers, E.-J. (2015). The effect of horizontal eye movements on free recall: A preregistered adversarial collaboration. *Journal of Experimental Psychology: General*, *144*, e1–e15.
- Meyer, A., Frederick, S., Burnham, T. C., Guevara Pinto, J. D., Boyer, T. W., Ball, L. J., et al. (2015). Disfluent fonts don't help people solve math problems. *Journal of Experimental Psychology: General*, *144*(2), e16.

- Miller, L. E., & Stewart, M. E. (2011). The blind leading the blind: Use and misuse of blinding in randomized controlled trials. *Contemporary Clinical Trials*, *32*, 240–243.
- Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gtzsche, P. C., Devereaux, P. J., et al. (2010). CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. *Journal of Clinical Epidemiology*, *63*, e1–e37.
- Moher, J., Lakshmanan, B. M., Egeth, H. E., & Ewen, J. B. (2014). Inhibition drives early feature-based attention. *Psychological Science*, *25*, 315–324.
- Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., et al. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*, 0021.
- Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., et al. (2015). Promoting an open research culture. *Science*, *348*, 1422–1425.
- Nosek, B. A., & Lakens, D. (2014). A method to increase the credibility of published results. *Social Psychology*, *45*, 137–141.
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, p. aac4716.
- Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, *17*, 776–783.
- Peirce, C. S. (1878). Deduction, induction, and hypothesis. *Popular Science Monthly*, *13*, 470–482.
- Peirce, C. S. (1883). A theory of probable inference. In C. S. Peirce (Ed.), *Studies in logic* (pp. 126–181). Boston: Little and Brown.
- Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., et al. (2017). Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, *18*, 115–126.
- Resnik, J., & Curtis, D. (2016). Why eyes? Cautionary tales from law’s blindfolded justice. In C. T. Robertson & A. S. Kesselheim (Eds.), *Blinding as a solution to bias: Strengthening biomedical science, forensic science, and law* (pp. 226–247). Amsterdam: Academic Press.
- Robertson, C. T., & Kesselheim, A. S. (2016). *Blinding as a solution to bias: Strengthening biomedical science, forensic science, and law*. Amsterdam: Academic Press.
- Rosenthal, R. (1966). *Experimenter effects in behavioral research* (pp. 7, 62). Appleton-Century-Crofts.
- Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*, 301–308.
- Sainz, A., Bigelow, N., & Barwise, C. (1957). On a methodology for the clinical evaluation of phrenopraxic drugs. *The Psychiatric Quarterly*, *31*, 10–16.
- Schulz, K. F., & Grimes, D. A. (2002). Blinding in randomised trials: Hiding who got what. *The Lancet*, *359*, 696–700.
- Shanks, D. R., Newell, B. R., Lee, E. H., Balakrishnan, D., Ekelund, L., Cenac, Z., et al. (2013). Priming intelligent behavior: An elusive phenomenon. *PLoS One*, *8*, e56515.
- Shariff, A. F., & Norenzayan, A. (2007). God is watching you: Priming God concepts increases prosocial behavior in an anonymous economic game. *Psychological Science*, *18*, 803–809.
- Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtry, E., et al. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*, 337–356.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.
- Steege, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*, 702–712.
- Stefan, A. M., Gronau, Q. F., Schonbrodt, F. D., & Wagenmakers, E.-J. (2019). A tutorial on Bayes factor design analysis using an informed prior. *Behavior Research Methods*, *51*, 1042–1058.
- Unsworth, N., Redick, T. S., McMillan, B. D., Hambrick, D. Z., Kane, M. J., & Engle, R. W. (2015). Is playing video games related to cognitive abilities? *Psychological Science*, *26*, 759–774.
- van Dongen-Boomsma, M., Vollebregt, M. A., Slaats-Willemse, D., & Buitelaar, J. K. (2013). A randomized placebo-controlled trial of electroencephalographic (EEG) neurofeedback in children with attention-deficit/hyperactivity disorder. *The Journal of Clinical Psychiatry*, *74*, 821–827.
- 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*, 2–12.

- Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R., et al. (2016). Registered replication report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, *11*, 917–928.
- Wagenmakers, E.-J., & Brown, S. D. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. *Psychological Review*, *114*, 830–841.
- 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*, 627–633.
- Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*, 1100–1122.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.