Minimizing Mistakes In Psychological Science

Jeffrey N. Rouder[1], Julia M. Haaf[2], & Hope K. Snyder[2]

[1] University of California, Irvine

[2] University of Missouri

Author Note

This paper was written in R-Markdown with code for demonstration integrated into the text. The Markdown script is open and freely available at https://github.com/PerceptionAndCognitionLab/lab-transparent. J.R. adapted the database and computer automation solutions reported herein. J.H. adapted the Rmarkdown expanded document solutions reported herein. H.S. has served as a critical user of these approaches and has suggested several improvements. The three authors jointly wrote the manuscript.

Correspondence concerning this article should be addressed to Jeffrey N. Rouder. E-mail: jrouder@uci.edu

Abstract

Developing and implementing best practices in organizing a lab is challenging, especially in the face of new cultural norms such as the open-science movement. Part of this challenge in today's landscape is using new technologies such as cloud storage and computer automation. Here we discuss a few practices designed to increase the reliability of scientific labs by focusing on what technologies and elements minimize common, ordinary mistakes. We borrow principles from the Theory of High-Reliability Organizations which has been used to characterize operational practices in high-risk environments such as aviation and healthcare. From these principles, we focus on five elements: 1. implementing a lab culture focused on learning from mistakes; 2. using computer automation in data and meta-data collection wherever possible; 3. standardizing organization strategies; 4. using coded rather than menu-driven analyses; 5. developing expanded documents that record how analyses were performed.

Keywords: Reliable Science, Open Science, High Reliability Organizations, Data Management

Minimizing Mistakes In Psychological Science

If you have been a member of a psychology lab, then perhaps you are familiar with things not going as well as planned. You may have experienced a programming error, equipment failure, or, more likely, some rather mundane human mistake. Our mistakes include failing to properly randomize an experiment, overwriting a file by typing in the wrong name, forgetting to notate an important code, putting relevant information in the wrong directory, analyzing the wrong data set, mislabeling figures, and mistyping test statistic values when transcribing from output to manuscripts. Finding these mistakes and preventing them from affecting publications is frustrating and time consuming.

Lab practices, especially statistical practices, have come under scrutiny in the last several years. We perhaps sit at the confluence of three troubling trends: First, some findings that were once thought to be rock solid have failed to replicate in registered reports (Ebersole et al., 2016; Hagger et al., 2016; Harris, Coburn, Rohrer, & Pashler, 2013; Open Science Collaboration, 2015; Wagenmakers et al., 2016). Second, the field has been beset by a number of high-profile fraud cases where researchers made up their data (Bhattacharjee, 2013). Third, seemingly improbable findings have been published in top tier journals. The most famous of these is Bem (2011), but several other claimed phenomena seem implausible as well (Primestein, n.d.).

In response to this confluence, there have been many diagnoses and proposed solutions. Some range from the global where the problem is that incentives reward superficial success at the expense of knowledge accumulation (Finkel, Eastwick, & Reis, 2015; Nosek & Bar–Anan, 2012; Nosek, Spies, & Motyl, 2012). Researchers operating under such incentives may cut corners especially in statistical testing (John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011; Yu, Sprenger, Thomas, & Dougherty, 2014). Among the specific recommendations are to value replication experiments (Nosek et al., 2015; Roediger, 2012; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012), to be open with data and methods so others may check your work (Rouder, 2016; Vanpaemel, Vermorgen,

Deriemaecker, & Storms, 2015; Wicherts, Borsboom, Kats, & Molenaar, 2006), and to adopt statistical approaches that require more thought and care (Benjamin et al., 2018; Erdfelder, 2010; Gigerenzer, 1998; Rouder, Morey, Verhagen, Province, & Wagenmakers, 2016).

This paper is about none of these issues. It is about mundane, commonly-made, obvious mistakes—the type that we can all agree are detrimental. They include bone-headed moves such as reporting statistics on incomplete data, using the wrong version of a figure, failing to properly randomize an experiment, and misplacing important codes. Everyone has seemingly made them, and nobody appears to be immune. Obviously, nobody wants to make these mistakes, but is it worthwhile to address them? Here is why we think ordinary mistakes should be taken seriously:

First, we think these mistakes are common in the literature. Perhaps the easiest mistake to detect is to search for malformed statements of statistical tests. A statistical test is malformed if the combination of test statistic and degrees of freedom do not match the corresponding p-value. Nuijten, Hartgerink, Assen, Epskamp, & Wicherts (2016) set out to document the frequency of this one type of error by web scraping statistical tests from published papers. Worryingly, Nuijten et al. (2016) found that about half the papers in 30 years of literature contained at least one malformed statistical test. Although this result has been controversial (Schmidt, 2016), it shows that this type of preventable mistakes enter the literature frequently. What about other types of mistakes? It is hard to know how often people make mistakes that get codified in the literature because outside of the authors, they are difficult if not impossible to catch. If we use Nuijten et al. (2016) as a proxy for these other types of mistakes, then there is reason to suspect they too are common.

A second reason to take these mistakes seriously is that they may act as a difficult-to-detect, field-wide bias for certain results. The argument comes from Gould (1996), who notes that simple mistakes tend to go in researchers' preferred direction. In his famed monograph, *The Mismeasure of Man*, Gould traces how scientists concluded that women were less intelligent than men and that colonized people were less intelligent than Europeans.

Gould documents how questionable research practices were used in reinforcing preconceived notions about race, gender, and intelligence. In this context, he discusses the role of simple errors. One might think if simple errors just occur randomly, they should be as likely to go against the researchers' preferred direction as for it. Gould argues that simple errors, however, go in the preferred direction more often than against it. One mechanism is selective checking. Researchers tend to check their work vigorously for mistakes when results are against their stated hypotheses and beliefs. They do not check as vigorously when the results are in the anticipated direction. Therefore, uncaught mistakes tend to reinforce confirmation bias.

Gould's hypothesis strikes us as reasonable. Imagine a researcher who fails to properly randomize an experiment. Say the researcher is studying Stroop context effects, and unbeknownst to the researcher, there is an overrepresentation of incongruent trials. When there are many incongruent trials, Stroop effects tend to be attenuated (Logan & Zbrodoff, 1979). Upon failing to find the anticipated priming effect the researcher rechecks the code and finds the randomization mistake. Suppose instead, however, that there is an overrepresentation of congruent trials. In this case, Stroop effects are exaggerated (Logan & Zbrodoff, 1979). The researcher, having established the Stroop effect, is less likely to recheck the code.

With these two reasons, that ordinary mistakes may be common and that they may serve as a source of confirmation bias, we next provide brief coverage of how mistakes are avoided in high-risk environments.

High Reliability Organizations

A starting point for us in improving lab practices is to consider practices in high risk fields where mistakes, failures, and accidents can have devastating consequences, say in aviation, the military, the nuclear power industry, and healthcare. Fortunately, there is a sub-discipline of management devoted to studying and improving organizations that serve in

high-risk environments where accidents may be catastrophic. Organizations that mitigate risks through ongoing processes are sometimes known as high reliability organizations, and the principles they follow are known as high reliability organization (HRO) principles (Weick, Sutcliffe, & Obstfeld, 2008).

Should your lab be a high reliability organization? Fortunately, mistakes in our labs do not have life-or-death consequences. Nonetheless, errors in how we produce knowledge waste our time when caught and threaten our reputations when not. The good news is that the principles of a high reliable organization transfer well to the academic lab setting. In the following sections we review the five principles. We describe how they lead to the construction of a better lab.

Principle I: Sensitivity to Operations: Those of us in experimental psychology are in the knowledge-production business. We often focus on the *what* of this business. What are our experiments? What are the data? What are the theories? What do the data allow us to infer about the theories? Our attention is on outcomes rather than processes. Sensitivity to operations means focusing on the processes underlying the *how* of knowledge production. How do we insure experiments are properly randomized? How do we document who ran what where? How do we insure the integrity of the knowledge we produce? In practice, sensitivity to operations means studying the more mechanistic processes by which a lab produces knowledge.

Principle II: Preoccupation With Failure: High reliability organizations are preoccupied with failures. They not only scrutinize their operations, they scrutinize them for points of failure. They are constantly trying to envision how things could go wrong and to take safeguards before they do. One element of this preoccupation is taking near-miss events as seriously as consequential mistakes. In aviation, for example, runway incursions that have no effect on operations are scrutinized much like runway incursions that materially threaten safety. In a lab setting, preoccupation means looking for ways to proactively anticipate and avoid mistakes, and taking small mistakes seriously.

Principles III & IV: Resiliency in the Face of Failure and Reluctance to Simplify: Principles III and IV both apply to failures either small or catastrophic. Resiliency refers to a maturity about failures—that, although they are to be minimized, they will occur from time to time. This maturity means that the organization has the processes in place to learn from failures so that they will not be repeated. Reluctance to simplify means that in diagnosing the cause of failures, simple answers, such as operator error, are not considered satisfying. The goal here is to go to the root of the problem with the acceptance that the organization is responsible for anticipating routine human and machine failures. Resiliency and reluctance to simplify may be implemented in an experimental psychology lab setting as well. The key is to avoid considering failures as a failure of meticulousness. In a resilient lab, when things go wrong, and they will, it is critical to talk about them, document them, and learn from them.

Principle V: Deference to Expertise: Deference to expertise is a principle designed to address hierarchies in organizations. Whereas administrators may be higher in the organizational structure, decisions about operations need to reflect deference to people who execute these operations on a daily basis. In healthcare, hospital administrators must defer to the expertise of nurses and doctors who execute the daily operations. In aviation, the mechanics who work on planes each day have a unique vantage about safety in the maintenance of planes. Labs too have a hierarchy. Deference to expertise means that each lab member, be it an undergraduate research assistant, a lab manager, a graduate student, a post-doctoral fellow, or a PI, has certain expertise. Undergraduates are helpful at understanding where human mistakes can happen in executing the experiments; graduate students can comment on errors when programming experiments and performing analyses. If a mistake is made in executing an experiment, then given their expertise, undergraduates may have the best insight into why the mistake occurred and what may be done to correct it. Listening to undergraduates in this regard is a form of deference to expertise.

From HRO Principles to Practices

We adopted HRO principles in 2014 and have been following them since. The five principles do not lead to any specific set of practices *per se*. Instead, they serve as guiding principles for how actions may be formulated in response to real-world circumstances. The result reflects the pressures faced by the organization, the operations in place, the foresight of the principles, etc.

In our case, given the nature of the mistakes we were making, following HRO principles has led to the following five practices: 1. Adopting a lab culture focused on learning from mistakes; 2. Implementing radical computer automation; 3. Standardizing organizational strategies across lab members; 4. Insuring statistical analyses are coded; and 5. Adopting expanded manuscripts where documentation of analyses are woven into the manuscript files. These five reflect the problems we faced and our overall comfort with computers when devising solutions.

We report our journey from the HRO principles to these five practices because we think it can help others minimize and mitigate mistakes. We provide no evidence of their effectiveness other than our experiences. Moreover, others who follow the HRO principles may come up with additional or alternative behavioral recommendations. We also describe our practices at a fairly general level without specific recommendations on implementation if only because no one set of specific recommendations is best for all labs. Klein et al. ({in press}) describe in far more detail the cyber-landscape for sharing data.

A Lab Culture Focused on Learning From Mistakes

In our current lab culture, we discuss problems and mistakes readily and often. Mistakes are socialized; that is, they reflect a failure of systems rather than a failure of people. This was not always the case, and the following story helps set up the contrast between a lab that learns from mistakes and one that does not.

Michael Pratte, a former graduate student and current assistant professor, tells the

following story: Back when our experiments were programmed in C and executed in DOS, he mistakenly typecast a variable as an integer rather than as a float. As a result, the code was not warning participants when they responded too quickly. We routinely provide this warning to discourage participants from responding very quickly as they may do to shorten the duration of the experimental session. When this mistake was discovered, Pratte recounts feeling sick to his stomach because the mistake may have affected several months worth of data collection. He believed at the time that this mistake was his alone.

Why did this mistake happen? It would be easy enough to blame Pratte for miscasting the variable. That blame does nothing to improve the reliability of the lab. Instead, errors, mistakes, and failures need to be brought out into the open where they may be examined. Otherwise, it is difficult to learn from them.

Let's see these recommendations in action for Pratte's case. The PI, Rouder, set up the lab so that experiments were programmed in C. Although the PI knows C well, it is a notoriously difficult programming language for newcomers, and newcomers tend to make this type of mistake. The core problem is the choice of C as opposed to a more user-friendly language. In response to Pratte's mistake, the lab moved on from C. Our current experiments are programmed in the more forgiving Psychophysical Toolbox.

One way of learning from mistakes is to record and log all mistakes. When we make a mistake, we open an adverse-event record, collaboratively, at a lab meeting. Our adverse events form is simple: One box is for a statement of the problem and the mistake it led to, another is for a set of possible solutions, a third is for the resolution (which solution was chosen and why), and a fourth is whether the resolution results in formal policy changes in the lab. We fill these boxes out together in a lab meeting, and they are logged within our database.

Labs do not need to wait for mistakes to have these discussions. In the HRO setting, there are *after-event reviews,* where processes are reviewed on a routine basis. For example, after a paper is submitted, the lab may engage in a review of the process without a

precipitating mistake. Our lab, however, has not taken this course and find the adverse-event approach sufficient.

Our recommendation is that all adverse events be socialized rather than privatized. Explicit statements of lab values should include some sense that mistakes, when they occur, are as much a failure of foresight of the lab as they are a failure of any individual.

Implementing Radical Computer Automation

All labs keeps records about their experiments. The question is whether these records are sufficiently detailed to minimize mistakes and mitigate them when they occur. One way of knowing if the records are sufficient is to perform *stress tests*. Consider the following:

- A graduate student has just discovered that the keyboard in Room 3 is sticky, and it must be hit multiple times to record a single keystroke. You have no idea how long this condition has been in play, but are sure the keyboard was fine last year. Can you identify all the data that has been obtained in Room 3 this year for inspection? (This is a true story from our lab. We had about a three week period with a bad keyboard. We were able to identify all data that were affected.)

- You have returned to a project after a long hiatus. You notice that the data have been previously cleaned by a graduate student who, unfortunately, dropped out in his first year. Do you have a system for recording these cleaning decisions or are those decisions gone with the student? If the latter, can you find the raw data? (This is a true story as well.)

- The new graduate student just changed the refresh rates on just one of the computers to run her psychophysics experiment. She did so through the control panel and outside the experimental software. This is both possible and common in Windows and Mac OS. Unfortunately, the change affects the timing of the other experiments run on the same computer. Do you have a record of which sessions were run at which timings?

(This is a true story too.)

The problem we faced was that of incomplete records. People simply forgot to record all the information they should have. To address these mistakes, we undertook a fairly large-scale effort to radically automate meta-data collection. The key for us was to adopt two new technologies: scripting and database management. As part of the experimental session, the computer launches a simple script asking the participant and experimenter to log in, and then collects demographics on the participant. The computer records a session entry with all desired information: who ran it, what room was it run in, who was the participant, what was IRB protocol, what were the screen resolutions, random-number generator seed, etc. These recordings are made into a relational database where they may be queried with readily-available tools.

In service of the communal goal of improving the trustworthiness of the literature, we think it should become a field-wide imperative to adopt greater computer automation. Some labs may be able to implement computer automation on their own, as was the case with our lab. Our main tool is a relational database, mySQL, which is run on a lab server. We have prewritten little scripts that insert the metadata into the database, and these are called by the experiments written in Psychophysical Toolbox. Of course, different labs may adopt different solutions in search of radical automation. Those that choose to do it on their own will find much help on the web for learning database management, shell scripting, and programming. Software Carpentry, a nonprofit organization for improving research computing (https://software-carpentry.org), may be quite helpful; they provide a large collection of web-based lessons in several useful technologies. The other approach is to use outside-the-lab expertise. The good news is that the needed expertise surely exists at your university—it might be in your department or college and available for free or at reasonable rates.

Standardization

The work of a lab may be organized in many ways. It is our experience that if each lab member is free to choose her or his own organizational strategy, each will choose a different approach. These differences are fine so long as each lab member tends to her or his work. The differing strategies, however, become fodder for mistakes as soon as work is shared. From our experience, it is best if all lab members used the same organization.

Perhaps the best example of standardization is that promoted by the Open Science Framework (OSF) storage system. The basic organization unit in OSF is a *project*. Underneath projects are data, manuscripts, and other components. Although projects differ somewhat, the basic structure helps researchers find elements with little if any documentation.

A well-organized lab should have a specified organizational structure. Particular attention should be given to the following: standardization of experimental meta-data, standardization of folder-naming conventions, and standardization in versioning. Standardization in meta-data means that each experiment should look similar. The lab should have a standard format for elements such as participants, sessions, IRBs, etc. Of course, variables in experiments differ, but standardization of the naming conventions across experiments is always helpful. Likewise, we find it helpful to have a standardized naming convention across directories and files so that future understanding of projects is seamless.

One source of mistakes is the clutter presented by retaining multiple versions of work products. Labs should have a common approach to versioning. Approaches to versioning may be as simple as putting set strings directly into file names. This may include appending dates or version numbers. However, this approach is not ideal in many ways, and we find that people tend to make mistakes. Moreover, the file-name approach defeats versioning on most cloud storage systems such as Google Drive, Box, and Dropbox. These systems have automatic versioning. Box, for example, automatically assigns version numbers to documents. Changing the file name defeats this feature. We use Git for versioning because it

gracefully deals with all our versioning needs. A tutorial may be found at Vuorre & Curley (submitted) and lessons and books are readily available online (see https://swcarpentry.github.io/git-novice/ and www.git-scm.com/book). Mistakes from the clutter of multiple versions are easily avoided by standardizing the versioning strategy ahead of time.

In our lab, we organize our research output by projects. A project is conceptually related research, and we tend to use a rather small scale to define a project. A project lives in two places. One is in the file system and the other is in the database.

In the file system, we use the following conventions. Nested in projects are the same five folders: *dev*, *share*, *papers*, *presentations*, and *grants*. In *papers* there is are subdirectories for each submission, and for this paper there are currently three directories, *sub*, *sub2*, and *rev1*, for the three main versions of this paper. There is also a directory *private* for all communication with editors and reviewers. The *private*, *dev*, and *grants* directories are not included in our public branch of the project. All projects follow this form, and with it, it is easy to find things.

In the database, we record the title of the project, its description, who is the lead, when it was last modified. We also have log entries for each project and as people work on the project, they can write what they did in such an entry. Projects also have output recorded—what are are the publications and talks associated with the project. Finally, projects in the database are associated with one or more repositories—those places on a file system where the files are.

Experiments are separate from projects in our system. They have similar conventions about where the IRBs are stored, how columns are named, etc. The link between experiments and projects is made by integrating analysis into papers as discussed subsequently.

Coded Analysis

We found in practice that researchers who use Excel occasionally cannot recreate a graph. To provide for the greatest reliability, data analysis should be coded. The alternative to coding is to use menu-driven systems. The problem with menu-driven systems is that there are choices that need to be made while navigating the menus. These menu choices may be made quickly, and often without any record of doing so. Excel, an example of a menu-driven system, is unreliable because while the outputs and formula may be saved, there are many steps, say the copying of cells, that are not documented. Some analysis programs have both a menu-driven interface and a code-based representation. An example is SPSS. These programs are reliable to the degree researchers remember to save their code.

There are code-based systems without menus including R, Matlab, and SAS. These systems run simplified computer languages that are tailored for data analysis. The inputs are the code, which are usually stored as a matter of routine. One of the nicest features of coded analyses is that the code may be shared. In many cases, the code itself is so transparent that no further documentation is needed for understanding and replicating the analyses.

Expanded Manuscripts

One common source of errors is accurately reporting results of analyses. Over the years, we have made such tragic mistakes as including the wrong figure in a paper and failing to analyze the cleaned data. One way of minimizing these errors is to expand the notion of a manuscript to include the provenance of analyses. A trustworthy manuscript includes a healthy trail indicating what code produced the analysis, what version of the code was used, what version of the data were used, when the analysis was conducted, and by whom. One simple approach could be to use comment functions available in most word processor and typesetting systems. All analyses can be extensively documented in the comments, and the comments, though not published, should remain with the document.

We take a more integrated and reliable approach to expanding documents. We use

Rmarkdown, a new composite of two very powerful platforms. One is R (R Core Team, 2017), which was discussed previously. The other is Markdown, which is a simple typesetting system for creating outputs in pdf, Word, or html. The Markdown environment is used to typeset the text and equations. Markdown is one of the simplest, easiest-to-use programs. It is not as powerful as Word, but it has all the features researchers need to do reliable science. Markdown documents are styled, and one of the developed styles is an APA-formatted document (Aust & Barth, 2017).

The key feature of a Markdown document is that it may contain special boxes that are executed when the document is formatted. We use this feature to place R-code chunks into the markdown document. These chunks are executed in R when the document is formatted. The process of formatting the text and executing the R code at the same time is called *knitting*.

Here we provide an example of knitting. This manuscript is available at https://github.com/PerceptionAndCognitionLab/lab-transparent. The project file `paper.Rmd` contains numerous R chunks. The following chunk is in the paper, and it assigns values -1, 0, 1, 2 to the variable `dat`, takes its sample mean, and does a one-sample t-test to see if the true mean is different from zero:

```
dat <- c(-1,0,1,2) #the data are -1, 0, 1, 2
sampMean <- mean(dat) # takes the mean of the data
tResults=t.test(dat) #performs t-test
tOut=apa_print(tResults)$statistic #apa-formatted string of t-test
```

The outputs are stored in the variable `sampMean` and `tOut`. We can reference them within the text using, ‘r round(sampMean,2)‘. When this document is knitted, the value of `sampMean` is rounded to two digits and printed; it is 0.50. A similar approach can be taken with the t-test, for example ‘r tOut‘ yields $t(3) = 0.77$, $p = .495$. Note how we never type the actual value of the statistics, and this approach prevents transcription errors. Moreover, if

the data change—say they are updated to include new participants—the code when run again updates the values. And if a researcher chooses different settings, say in cleaning, again, a simple run of the paper updates the values to reflect these new settings. We put our settings in a separate chunk for transparency.

The above chunk is too simple to be of much service. In a real-world application we need to retrieve data from a cloud, clean the data, perform analyses, and draw figures and tables. However, as users improve their R skills, these tasks become routine.

The knitted approach with Rmarkdown is growing. Here we highlight two innovative package that we think are broadly useful. The first, which was mentioned previously, is Frederik Aust's package for writing APA-compliant manuscripts in Rmarkdown. This package, called `papaja`, does most of the formatting work for the author. In the above chunk, the function `apa_print()` comes from the `papaja` package, and it takes common test statistics computed in R and formats them in an APA-compliant style. A review of `papaja` is provided in Aust & Barth (2017). The package also provides APA compliant LaTeX tables. For word processors, `apaTables` is an innovative `R` package provided by David Stanley. This package prints matrices and tables in R in an APA compliant format. It too is not only convenient, but eliminates transcription errors in typesetting tables. See Stanley (2018) for an introduction and guide to the package.

Conclusions: Minimizing Mistakes and Moving Toward a More Open Science

In this paper, we used the high-reliability-organization principles to make recommendations for minimizing mistakes. The recommendations are to adopt a lab culture focused on mistakes; use radical computer automation, standardization, and coded analyses; and expand the document to include documentation of analyses. We have yet to find a researcher who argues against these practices. Instead, the more common response concerns the time commitment. Is it worth the time to implement these recommendations? This question is pertinent in the neo-liberalized university where administrators are stressing

bean-counting of publications, citations, and grant revenues. And it is especially pertinent for younger scholars eyeing their first appointment or a tenure clock.

We think there are a few different questions rolled up here. The first, and the easiest, is whether it is worth the time to read and implement HRO principles. The answer here is assuredly, "yes." The principles are simple, and the reading takes under an hour. Implementing them at the most general level is more a matter of mindset and focus. Shifting towards sensitivity to operations, towards being preoccupied with failure, towards resilience in the face of failure, and towards deference to expertise is always worth the time.

The real time commitment, however, comes in changing how things are done to avoid mistakes. We have listed our five behavioral approaches. Some are not time intensive; for example, standardization comes with little to no time cost. Others, however, such as radical computer automation and expanded documents, may require learning new skills. And that does take time. For us, time spent on making the lab more reliable is an up-front investment. Once we implement a change, it seems that many subsequent activities become easier and more convenient. Over the course of years or decades, it seems that the specific recommendations, especially those about automation, are great time savers. Those new to the technologies need not pick them all up at once. They may be implemented in steps. Perhaps this year you may learn about `Rmarkdown` and next year about relational databases, and so on.

One of the hidden benefits of making the lab more reliable is that it opens the door to open science. We define open science as working to preserve the ability of others to reach their own opinions of our data and analyses. Because others can reach their own opinion, opening up our work to others is a scary proposition that involves some intellectual risk and professional vulnerability. After all, we would be grateful and mortified if someone found a critical error in our work, and definitely not in that order. One way of managing this risk and vulnerability is to do the best we can to avoid mistakes. Having a reliable pipeline gives us the confidence to be public and open. And being open reinforces the need to be reliable.

We have been practicing open science for about two years. It is our view that there are some not-so-obvious benefits that have improved our work as follows: There are many little decisions that people must make in performing research. To the extent that these little decisions tend to go in a preferred direction, they may be thought of as subtle biases. These decisions are often made quickly, sometimes without much thought, and sometimes without awareness that a decision has been made. Being open has changed our awareness of these little decisions. Lab members bring them to the forefront early in the research process where they may be critically examined. One example is that a student brought up outlier detection very early in the process knowing that not only would she have to report her approach, but that others could try different approaches with the same data. Addressing these decisions head on, transparently, and early in the process is an example of how practicing open science improves our own science.

References

Aust, F., & Barth, M. (2017). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407–425. Retrieved from http://dx.doi.org/10.1037/a0021524

Benjamin, D. J., Berger, J., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., ... Johnson, V. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*, 6.

Bhattacharjee, Y. (2013). The mind of a con man. New York Times, April 26, 2013. Retrieved from http://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html?pagewanted=all

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., ... Nosek, B. A. (2016). Many labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82. Retrieved from http://ezid.cdlib.org/id/doi:10.17605/OSF.IO/QGJM5

Erdfelder, E. (2010). A note on statistical analysis. *Experimental Psychology*, *57*(1-4). Retrieved from 10.1027/1618-3169/a000001

Finkel, E. J., Eastwick, P. W., & Reis, H. T. (2015). Best research practices in psychology: Illustrating epistemological and pragmatic considerations with the case of relationship science. *Journal of Personality and Social Psychology*, *108*(2), 275.

Gigerenzer, G. (1998). We need statistical thinking, not statistical rituals. *Behavioral and Brain Sciences*, *21*, 199–200. Retrieved from https://doi.org/10.1017/S0140525X98281167

Gould, S. J. (1996). *The mismeasure of man*. New York: WW Norton & Company.

Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., ... others. (2016). A multilab preregistered replication of the ego-depletion effect.

Perspectives on Psychological Science, *11*(4), 546–573.

Harris, C. R., Coburn, N., Rohrer, D., & Pashler, H. (2013). Two failures to replicate high-performance-goal priming effects. *PLoS ONE*, *8*, e72467.

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. Retrieved from http://pss.sagepub.com/content/23/5/524.abstract

Klein, O., Hardwicke, T. E., Aust, F., Breuer, J., Danielsson, H., Hofelich Mohr, A., ... Frank, M. C. ({in press}). A practical guide for transparency in psychological science. *Collabra: Psychology*.

Logan, G. D., & Zbrodoff, N. J. (1979). When it helps to be misled: Facilitative effects of increasing the frequency of conflicting stimuli in a stroop-like task. *Memory & Cognition*, *7*(3), 166–174.

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

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*, 615–631.

Nuijten, M. B., Hartgerink, C. H., Assen, M. A. van, Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*(4), 1205–1226.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6521), 943. Retrieved from dx.doi.org/10.1126/science.aac4716

Primestein, D. J. (n.d.). *Psi-chology*. Retrieved from http://www.psi-chology.com

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna,

Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Roediger, H. L. (2012). Psychology's woes and a partial cure: The value of replication. *APS Observer*, *25*.

Rouder, J. N. (2016). The what, why, and how of born-open data. *Behavioral Research Methods*, *48*, 1062–1069. Retrieved from 10.3758/s13428-015-0630-z

Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., & Wagenmakers, E.-J. (2016). Is there a free lunch in inference? *Topics in Cognitive Science*, *8*, 520–547.

Schmidt, T. (2016). Sources of false positives and false negatives in the statcheck algorithm: Reply to nuijten et al.(2016). *ArXiv Preprint ArXiv:1610.01010*.

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. Retrieved from https://doi.org/10.1177/0956797611417632

Stanley, D. (2018). *ApaTables: Create american psychological association (apa) style tables*. Retrieved from https://CRAN.R-project.org/package=apaTables

Vanpaemel, W., Vermorgen, M., Deriemaecker, L., & Storms, G. (2015). Are we wasting a good crisis? The availability of psychological research data after the storm. *Collabra*, *1*(1:3), 1–5.

Vuorre, M., & Curley, J. P. (submitted). *Curating research assets: A tutorial on the git version*. Retrieved from https://psyarxiv.com/6tzh8

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams Jr, R., ... others. (2016). Registered replication report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, *11*(6), 917–928.

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. Retrieved from https://doi.org/10.1177/1745691612463078

Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (2008). Organizing for high reliability: Processes of collective mindfulness. *Crisis Management*, *3*(1), 81–123.

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*(7), 726–728. Retrieved from http://wicherts.socsci.uva.nl/datasharing.pdf

Yu, E. C., Sprenger, A. M., Thomas, R. P., & Dougherty, M. R. (2014). When decision heuristics and science collide. *Psychonomic Bulletin & Review*.