# Minimizing Mistakes In Psychological Science

Jeffrey N. Rouder[1,2], 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. adopted the expanded documented reported herein as well as assisted in writing the manuscript. 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. The challenge is compounded by the availability of new technologies such as cloud storage and by new cultural norms such as the open-science movement. Here we discuss a few practices designed to increase the reliability of scientific labs by focusing on what technologies and elements minimize mistakes in the collection, curation, and sharing of data and materials. 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. creating audit trails; 2. using computer automation wherever possible; 3. coding analyses; 4. developing expanded documents that include analyses; and 5. implementing a lab culture focused on learning from mistakes.

*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. Mistakes like these are unfortunately common, and they are frustrating and time consuming to correct. And that is for the mistakes we collectively catch. How much carnage there is in the literature from mistakes that have not been caught is anyone's guess.

One concept related to mistakes is transparency. Most of us have projects that have taken months if not years to complete. The question is whether steps we are taking today are transparent a few years from now. The problem is that we all tend to rely too much on our memory, and as a consequence, the location, organization, purpose, and meaning of some of our work is lost as our memories fade. Therefore, making work transparent into the future is key in avoiding mistakes.

When we ask researchers about how they avoid mistakes and insure their work is transparent and replicable into the future, many reference their well-honed sense of meticulousness in curating and documenting their lab products and activities. Our view is that relying on personal meticulousness is neither satisfying nor effective. It certainly doesn't work in our lab because we are not meticulous. Even those who rely on meticulousness know that meticulousness itself is not fail safe. It works fine until it doesn't. And when it doesn't, the mistakes become a matter of personal failure with an obvious locus of blame, usually on some poor graduate student. And what can be said other than be more careful?

In psychology, we as a community have devoted much time and thought on how to improve the trustworthiness of the literature. There have been a number of proposals about replicating experiments (Nosek et al., 2015; Roediger, 2012; E.-J. Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012), being open with data and methods (Rouder, 2016; Vanpaemel, Vermorgen, Deriemaecker, & Storms, 2015; J. M. Wicherts, Borsboom, Kats, & Molenaar, 2006), and revising our statistical approaches (Benjamin et al., 2017; Erdfelder, 2010; Rouder, Morey, Verhagen, Province, & Wagenmakers, 2016). In this climate, the time is ripe to push for lab practices and lab cultures that minimize mistakes.

## 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 up experiments and doing analysis, and PIs are well suited to shape the culture of the lab and keep long-term goals in mind. By explicitly deferring to expertise, more reliable pipelines may be established.

These five principles have led in concert to a number of innovations in our lab: a thorough audit trail, a large degree of computer automation, a standardized organization, coded analysis, an expanded view of a manuscript, and a lab culture focused on learning from mistakes.

## Audit Trails

One of the largest recent changes in our lab is the adoption of audit trails. An audit trail refers to the set of information about a study that allows one to reconstruct the work. Audit trails are necessary, and in their absence, the integrity of the work cannot be insured.

One way of knowing if an audit trail is sufficient is to perform *stress tests.* Here are a few examples:

- The IRB has just audited you. You have several protocols, and they want all the following broken out by protocol over the last three years. Can you tell them how many participants were run on each and how many withdrew consent before the experiment began and how many withdrew consent during the experiment? Can you show the signatures for each participant for each protocol?

- 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)

- 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? Can you find the raw data?

- A post doc who has long since departed made a very impressive figure for a talk. You would like it as a model for a current project. Can you find it?

We started designing our audit trails to address these and other stress tests. The main issue was not what to store in these audit trails. Instead, it was how to insure the relevant information is recorded.

## Computer Automation

Most labs have audit trails. The usual problem is incompleteness. People simply forget to record all the relevant information. We believe that one remedy for incomplete audit trails is computer automation. We have programmed the data-collection computers to enter information whenever a participant is run. 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, etc. No fuss; no mess; perfect every time!

We cannot stress the importance of computer automation in recording an audit trail. It is the single, most important step labs can take to avoid and mitigate mistakes. Computers are seemingly perfect at remembering to record information; people, unfortunately, are not.

In service of the communal goal of improving the trustworthiness of the literature, it should become a field-wide imperative to adopt computer automation. Labs that feel they do not have the expertise may be pleasantly surprised to learn that such expertise surely exists in every university. We think it is incumbent that PIs push their departments and universities to work with the labs in automating their processes.

## Standardization

The work of a lab may be organized in a many ways. It is our experience that if each lab member is free to chose her or his own organizational strategy, each will chose 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.

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 common set of components in projects. In our case, it is developmental branches, outputs, and shared libraries, and the key is not these particulars, but that everyone in the lab names components the same way.

One source of mistakes for all researchers is the clutter presented by retaining multiple versions of work products. We think each lab should have a standardized approach to versioning. There are several ways to skin this cat. Some are simple, such as putting markers directly into file names. This may include appending dates or version numbers. Our standard is to rely on a dedicated versioning system rather than filaments. Most cloud storage systems such as Google Drive, Box, and Drop Box have automatic versioning. Box, for example, automatically assigns version numbers to documents. Changing the file name, in fact, defeats this feature. We use a versioning system called Git, and highly recommend it. A tutorial may be found at Vuorre and Curley (submitted). The key point here is that a lab should have a versioning standard that is known and followed throughout.

## Coded Analysis

To provided 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 choice 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, curate, and organize their code.

There are code-based systems without menus including R, Matlab, and SAS. These systems run off a simplified computer language that is 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 codes may be shared. In many cases, the code itself is so transparent that no further documentation is needed for understanding and replicating the analyses. In our view, trustworthy labs retain codes of their analyses so that they are replicable and transparent.

## 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. We suspect that mistakes like these are common in the literature. One source of concern are the results of Nuijten, Hartgerink, Assen, Epskamp, and Wicherts (2016) who have mined the literature to assess the frequency of one type of error, malformed statistical tests. A statistical test is malformed if the combination of test statistic and degrees-of-freedom do not match the corresponding *p*-value. Worryingly, Nuijten et al. (2016) found that half the papers in 30 years of literature contained at least one malformed statistical test. In our view, this is one small member of the class of errors where what is reported in the paper does not match what was actually done.

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 paper 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. A version with comments may be audited to check for errors.

We take a more integrated approach to expanding documents that we think is so beneficial that we describe it in depth. We use Rmarkdown a new composite of two very powerful platforms. One is R (R Core Team, 2016), 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 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)

tResults=t.test(dat)

tOut=apa_print(tResults)$statistic
```

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 transcription errors are avoided.

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.

## A Lab Culture Focused on Learning From Mistakes

One of our former graduate students tells the following story: Back when our experiments were programmed in C and executed in DOS, he mistakenly used a count from 1 rather than from 0. As a result, he was timing response times from the warning signal rather than from the stimulus presentation. When this mistake was discovered, this student recounts feeling sick to his stomach. It took him a number of days to gather the courage to confess to the mistake.

We think there are two critical errors in the above true account. The first is that experiments were programmed in a language that was ill-suited for graduate students in psychology. C is notoriously powerful but notoriously difficult. It gives the programmer just enough rope to hang him or herself. The PI is responsible for this choice, and what worked for him was not good for his students. Second, the lab culture was such that mistakes were individualized. The former student was thinking the mistake was his alone.

It is critical that errors, mistakes, and failures are brought out into the open without shame. Otherwise, it is difficult to learn from them. Our recommendation is that 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 it is a failure of any individual.

One of our current practices is to record and log all mistakes. When we make a mistake, we open an adverse-event record, collaboratively, at lab meeting. We hash out what happened with full deference to expertise. And we discuss how our operations can be improved to prevent the mistake in the future.

## Conclusions: Minimizing Mistakes and Moving Toward a More Open Science

The above processes are designed to make a lab more reliable by instituting an audit trail; by using computer automation, standardization, and coded analyses, and by expanding the document to include analyses. Making a lab more reliable does not *per se* commit one to open science. All of these steps minimize mistakes, but none of them need be done in public.

We define open science as working to preserve the ability of others to reach their own opinions of our data and analyses. Open science is a scary proposition that involves some intellectual risk and professional vulnerability. Having a reliable pipeline gives us the confidence to be public and open. And being open and public reinforces the need to be reliable.

We have been practicing open science for two years. It is our view that there are some not-so-obvious benefits. Here is how it has worked: 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.

Practicing open science has also allowed us to be more honest and authentic in practicing our craft. Open science requires one to have vulnerability, and dealing with this vulnerability promotes needed self-compassion. The argument goes as follows: Although we practice open science and attempt to make our pipelines transparent and reliable, we still make mistakes. Fortunately, we tend to catch them fairly early. But there may come a time where we miss something important. Perhaps we may have to retract or correct a publication. If we have to do so, we will certainly feel shame. But, we will also take solace knowing that we did our best. And perhaps it is this self-compassion and peace-of-mind that has been the best benefit of all.

## References

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

Benjamin, D. J., Berger, J., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., … Johnson, V. (2017). *Redefine statistical significance.* Retrieved from https://osf.io/preprints/psyarxiv/mky9j/

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

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.

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.

R Core Team. (2016). *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.

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., 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