COMMENT

ILLUSTRATION BY DALE EDWIN MURRAY

Hide results to seek the truth

More fields should, like particle physics, adopt blind analysis to thwart bias, urge **Robert MacCoun** and **Saul Perlmutter**.

Decades ago, physicists including Richard Feynman noticed something worrying. New estimates of basic physical constants were often closer to published values than would be expected given standard errors of measurement[1]. They realized that researchers were more likely to 'confirm' past results than refute them — results that did not conform to their expectation were more often systematically discarded or revised.

To minimize this problem, teams of particle physicists and cosmologists developed methods of blind analysis: temporarily and judiciously removing data labels and altering data values to fight bias and error[2]. By the early 2000s, the technique had become widespread in areas of particle and nuclear physics. Since 2003, one of us (S.P.) has, with colleagues, been using blind analysis for measurements of supernovae that serve as a 'cosmic yardstick' in studies of the unexpected acceleration of the Universe's expansion[3].

In several subfields of particle physics and cosmology, a new sort of analytical culture is forming: blind analysis is often considered the only way to trust many results. It is also being used in some clinical-trial protocols (the term 'triple-blinding' sometimes refers to this[4]), and is increasingly used in forensic laboratories as well.

But the concept is hardly known in the biological, psychological and social sciences. One of us (R.M.) has considerable experience conducting empirical research on legal and public-policy controversies in which concerns about bias are rampant (for example, drug legalization), but first encountered the concept when the two of us co-taught a transdisciplinary course at the University of California, Berkeley, on critical thinking and the role of science in democratic group decision-making. We came to recognize that the methods that physicists were using might improve trust and integrity in many sciences, including those with high-stakes analyses that are easily plagued by bias.

Many motivations distort what inferences we draw from data. These include the desire to support one's theory, to refute one's competitors, to be first to report a phenomenon, or simply to avoid publishing 'odd' ▶

NATURE.COM
For *Nature*'s special collection on reproducibility, see:
go.nature.com/huhbyr

results. Such biases can be conscious or unconscious. They can occur irrespective of whether choices are motivated by the search for truth, by the good mentor's desire to help their student write a strong PhD thesis, or just by naked self-interest[5].

We argue that blind analysis should be used more broadly in empirical research. Working blind while selecting data and developing and debugging analyses offers an important way to keep scientists from fooling themselves.

WHO KNOWS WHAT

Some forms of blinding are well known: for example, shielding both patients and clinicians from knowing who receives an experimental drug or a placebo (double-blinding), or removing names and affiliations from scientific manuscripts to keep peer reviewers from being swayed by authors' identities. But these practices apply to the collection and source of data, rather than the analysis.

Blind analysis ensures that all analytical decisions have been completed, and all programmes and procedures debugged, before relevant results are revealed to the experimenter. One investigator — or, more typically, a suitable computer program — methodically perturbs data values, data labels or both, often with several alternative versions of perturbation. The rest of the team then conducts as much analysis as possible 'in the dark'. Before unblinding, investigators should agree that they are sufficiently confident of their analysis to publish whatever the result turns out to be, without further rounds of debugging or rethinking. (There is no barrier to conducting extra analyses once data are unblinded, but doing so risks bias, so researchers should label such further analyses as 'post-blind'.)

There are many ways to do blind analysis. The computer need not (and probably will not) be blinded to data values; it is the display of results that masks information. Techniques must obscure meaningful results while showing enough of the data's structure to allow researchers to find and debug measurement artefacts, irrelevant variables, spurious correlates and other problems. For example, researchers who analyse clinical-trial results without knowing which patients received a placebo should still be able to identify implausible values.

"Blinding analyses could be as simple as asking a colleague to scramble labels."

The best methods for blinding depend on the properties of the data (for example, the type of statistical distribution, lower and upper bounds, whether values are discrete or continuous and whether cases were randomly assigned to experimental conditions or passively observed). Both data values and labels can be manipulated to develop a suitable strategy (see 'Blinding strategies').

A fertile approach is to present panels of possible results, in which the real results may or may not be interspersed among various decoys. Such a blinded presentation of possibilities typically triggers useful questions. For example, a plausible, although still blinded, graph may lead the researcher to ask whether a sample explores the full range of an independent variable, or it might trigger a revisiting, before unblinding, of the scaling of one of the variables. Another graph might suggest that the whole effect is driven by a single outlier point, and suggest that the researcher needs more data, again before unblinding. Often, a panel can seem implausible until the investigator recognizes an assumption that, if wrong, would produce such a pattern.

BLINDING STRATEGIES

Technique examples	Perturbation	Potential application
Noising $\theta_{ij}=y_{ij}+n_{ij}$ or $\theta_{ij}=\beta_k+n_{ij}$	Add a random number (from an appropriate statistical distribution) to data points or model parameters.	Testing which of several prevention messages is most effective in reducing smoking.
Biasing $\theta_{ij}=y_{ij}+b_j$	Obscure differences in experimental conditions by adding a hidden value that is biased in a particular direction.	Estimating whether the costs of a controversial safety regulation exceed its benefits.
Cell scrambling $\theta_{ij}=y_{\#}$	Shuffle labels for experimental conditions, so that it is unclear which set of results matches which conditions.	Testing a prediction that hard-copy books are better comprehended than audiobooks.
Item scrambling $\theta_{ij}=y_{\#\#}$	Randomly relabel each data point to de-identify experimental conditions.	Analysing group differences that might be easy to recognize even with noise and bias (for example, effects of neighbourhood and school on crime victimization).
Various combinations	Row scrambling: keep pairs of variables together to preserve correlation. Variable blinding: swap labels of various variables.	

y_{ij} is the ith observation in the jth condition ('cell') of the study; β_k is the kth parameter of a model; θ_{ij} is y_{ij} or β_k after blinding; n_{ij} is random error, b_j is a bias term, and # denotes a randomly swapped subscript.

COMMON OBJECTIONS

Blind analysis is not a panacea, but it is much more feasible than many think. Here we address common objections.

Won't people just peek at the raw data? Blind analysis is not immune to fraud. But in ordinary research, teams of investigators can help to enforce compliance. Where blinding is part of the culture, graduate students and postdocs often become its most effective guardians, for example, flagging the risk if their adviser asks for a plot that might accidently unblind the result.

Can't we avoid bias another way? Other solutions have been proposed, including preregistered analysis plans, cross-lab replication, the *p*-curve, adversarial collaboration, Bayesian analytical methods and sensitivity analysis[6]. These techniques all have their place, but they do not fully address the specific problem. For example, preregistration requires that data-crunching plans are determined before analysis, and offers some of the same benefits as blind analysis. But it also limits the scope of analysis. Because many analytical decisions (and computer programming bugs) cannot be anticipated, investigators will be forced to make some decisions knowing (consciously or unconsciously) how their choices affect the results. Blind analysis enables the investigator to engage in analysis, exploration and finalization without worrying about such bias.

Isn't blind analysis too much hassle? There is extra effort involved. Often the analyses that at first seem most worth the trouble are those that involve expensive data, high-stakes decisions or topics especially prone to bias. However, blinding analyses could be as simple as asking a colleague down the hall to scramble labels. And when safety is at stake, such as in some clinical trials, it often makes sense to set up an unblinded safety monitor while the rest of the analytical team is in the dark[7]. Technology could help here: an important advance would be the introduction of off-the-shelf algorithms in standard analysis software to maintain the blinding until the group is ready to reveal the results. A less obvious benefit is the sheer fun of the dramatic moment when the results are revealed.

Won't blinding lose outcomes that depend on analyses done once the result is seen? Ideally, among the panels of blinded results there is also the set of actual results, so the researchers could use this to consider further implications (along with the implications of the other hypothetical results). Of course, there will still be post-hoc (post-unblinding) discussion; it will simply be possible to distinguish work investigators performed while still unaware of the results.

MAKING IT HAPPEN

We see two challenges for the widespread dissemination of blind analysis. The first is technical: learning to blind what should be

ILLUSTRATION BY DAVID PARKINS

blinded while preserving features needed to permit appropriate analysis. The second is motivational: creating incentives for investigators to adopt a method that might make it harder for them to come up with desirable (although possibly false) results.

Supplementary research grants that encourage testing blind-analysis methods across multiple fields could help to tackle both challenges. The efficacy of various approaches — methods of blinding, pre-registration and other measures against confirmation bias — should be treated as empirical questions to be answered by future research, as demonstrated by a 2015 study of the effects of preregistration[8]. Many blinding techniques have already been developed[2], and hopefully, a meta-science of best practices will emerge.

Wider use of blinded analysis could be a boon to the scientific community. The main use is to filter out biased inferences, but there are other benefits, too. First, blind analysis can help investigators to consider the opposite of their expectations, a proven strategy for sound reasoning[9]. Second, blinding exposes the investigator to unexpected patterns that fuel both creativity and scrutiny of the theory and methodology[10].

Finally, blind analysis helps to socialize students into what sociologist Robert Merton called science's culture of 'organized scepticism'. As Feynman put it: "This long history of learning how to not fool ourselves — of having utter scientific integrity — is, I'm sorry to say, something that we haven't specifically included in any particular course that I know of. We just hope you've caught on by osmosis. The first principle [of science] is that you must not fool yourself — and you are the easiest person to fool." ■

Robert MacCoun *is a psychologist and a professor of law at Stanford University in California, USA.* **Saul Perlmutter** *is a professor of physics at the University of California, Berkeley, USA. He shared the 2011 Nobel Prize in Physics.*
e-mails: rmaccoun@stanford.edu; saul@lbl.gov

1. Feynman, R. P. *Surely You're Joking, Mr. Feynman!* (W. W. Norton, 1985).
2. Klein, R. J. & Roodman, A. *Annu. Rev. Nucl. Part. Sci.* **55,** 141–163 (2005).
3. Conley, A. *et al. Astrophys. J.* **644,** 1–20 (2006).
4. Miller, L. E. & Stewart, M. E. *Contem. Clin. Trials* **32,** 240–243 (2011).
5. MacCoun, R. J. *Annu. Rev. Psychol.* **49,** 259–287 (1998).
6. Miguel, E. *et al. Science* **343,** 30–31 (2014).
7. Meinert, C. L. *N. Engl. J. Med.* **338,** 1381–1382 (1998).
8. Kaplan, R. M. & Irvin, V. L. *PLoS ONE* **10,** e0132382 (2015).
9. Lord, C. G., Lepper, M. R. & Preston, E. *J. Pers. Soc. Psychol.* **47,** 1231–1243 (1984).
10. Simonton, D. K. *Rev. Gen. Psychol.* **15,** 158–174 (2012).

Many hands make tight work

Crowdsourcing research can balance discussions, validate findings and better inform policy, say **Raphael Silberzahn** and **Eric L. Uhlmann**.

Our experience with crowdsourced analysis began in 2013, shortly after we published research[1] suggesting that noble-sounding German surnames, such as König (king) and Fürst (prince), could boost careers. Another psychologist, Uri Simonsohn at the University of Pennsylvania in Philadelphia, asked for our data set. He was sceptical that the meaning of a person's name could affect life outcomes. While our results were featured in newspapers around the world, we ▶