Several years ago, Ioannidis (2005) famously argued that "most published research findings are false." And indeed, recent results from biomedical and cancer research suggest that replication rates are lower than 50%, with some as low as 11% (Begley & Ellis, 2012; Osherovich, 2011; Prinz, Schlange, & Asadullah, 2011). If the above results carry over to psychology, our discipline is in serious trouble (Carpenter, 2012; Roediger, 2012; Yong, 2012). Research findings that do not replicate are worse than fairy tales; with fairy tales the reader is at least aware that the work is fictional.

In this article, we focus on what we believe to be the main "fairy-tale factor" in psychology today (and indeed in all of the empirical sciences): the fact that researchers do not commit themselves to a plan of analysis before they see the data. Consequently, researchers can fine tune their analyses to the data, a procedure that make the data appear to be more compelling than they really are. This fairy-tale factor increases the probability that a presented finding is fictional and hence non-replicable. We propose a radical remedy—preregistration—to ensure scientific integrity and inoculate the research process against the inalienable biases of human reasoning. We conclude by illustrating the remedy of preregistration using a replication attempt of an extrasensory-perception (ESP) experiment reported by Bem (2011).

Bad Science: Exploratory Findings, Confirmatory Conclusions

Science can be bad in many ways. Flawed design, faulty logic, and limited scholarship engender no confidence or enthusiasm whatsoever.[2] In this section, we discuss another important factor that reduces confidence and enthusiasm for a scientific finding: the fact that almost no psychological research is conducted in a purely confirmatory fashion[3] (e.g., Kerr, 1998; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011; for a similar discussion in biology, see Anderson, Burnham, Gould, & Cherry, 2001). Only rarely do psychologists indicate, in advance of data collection, the specific analyses they intend to carry out. In the face of human biases and the vested interest of the experimenter, such freedom of analysis provides access to a Pandora's box of tricks that can be used to achieve any desired result (e.g., John et al., 2012; Simmons, Nelson, & Simonsohn, 2011; for what may happen to psychologists in the afterlife, see Neuroskeptic, 2012). For instance, researchers can engage in cherry picking: They can measure many variables (gender, personality characteristics, age, etc.) and only report those that yield the desired result, and they can include in their papers only those experiments that produced the desired outcome, even though these experiments were designed as pilot experiments that could be easily discarded had the results turned out less favorably. Researchers can also explore various transformations of the data, rely on one-sided *p* values, and construct post-hoc hypotheses that have been tailored to fit the observed data (MacCallum, Roznowski, & Necowitz, 1992). In the past decades, the development of statistical software has resulted in a situation in which the number of opportunities for massaging the data is virtually infinite.

True, researchers may not use these tricks with the explicit purpose to deceive—for instance, hindsight bias often makes exploratory findings appear perfectly sensible. Even researchers who advise their students to "torture the data until they confess"[4] are hardly evil geniuses out to deceive the public or their peers. Instead, these researchers may genuinely believe that they are giving valuable advice that leads the student to analyze the data more thoroughly and increases the odds of publication along the way. How could such advice be wrong?

In fact, the advice to torture the data until they confess is not wrong—just as long as this torture is clearly acknowledged in the research report. Academic deceit sets in when this does not happen and partly exploratory research is analyzed as if it had been completely confirmatory. At the heart of the problem lies the statistical law that, for the purpose of hypothesis testing, the data may be used only once. So when you turn your data set inside and out, looking for interesting patterns, you have used the data to help you formulate a specific hypothesis. Although the data may still serve many purposes after such fishing expeditions, there is one purpose for which the data are no longer appropriate—namely, for testing the hypothesis that they helped to suggest. Just as conspiracy theories are never falsified by the facts that they were designed to explain, a hypothesis that is developed on the basis of exploration of a data set is unlikely to be refuted by that same data. Thus, one always needs a fresh data set for testing one's hypothesis. This also means that the interpretation of common statistical tests in terms of Type I and Type II error rates is valid only if the data were used only once and if the statistical test was not chosen on the basis of suggestive patterns in the data. If you carry out a hypothesis test on the very data that inspired that test in the first place then the statistics are invalid (or "wonky", as Ben Goldacre put it). In neuroimaging, this has been referred to as "double dipping" (Kriegeskorte, Simmons, Bellgowan, & Baker, 2009; Vul, Harris, Winkielman, & Pashler, 2009). Whenever a researcher uses double-dipping strategies, Type I error rates will be inflated and *p* values can no longer be trusted.

As illustrated in Figure 1, psychological studies can be placed on a continuum from purely exploratory, where the hypothesis is found in the data, to purely confirmatory, where the entire analysis plan has been explicated before the first participant is tested. Every study in psychology falls somewhere along this continuum; the exact location may differ depending on the initial outcome (i.e., poor initial results may encourage exploration), the clarity of the research question (i.e., vague questions allow more exploration), the amount of data collected (i.e., more dependent variables encourage more exploration), the a priori beliefs of the researcher (i.e., strong belief in the presence of an effect encourages exploration when the initial result is ambiguous), and so on. Hence, the amount of exploration, data dredging, or data torture may differ widely from one study to the next; consequently, so does