# Why Most Published Research Conclusions in Psychology May Not Be False

David M. Lane
Rice University
lane@rice.edu

Running head: Conclusions may not be false

Abstract

It has recently been argued that published research in fields such as epidemiology and neuroscience contain mostly false conclusions. The main causes of these false conclusions are (1) editorial practices that rarely publish nonsignificant effects, (2) the relatively small proportion of studies conducted in which the null hypothesis is false, and (3) the relatively low power of most experiments. This paper explores the false conclusion rate in subfields of psychology in which the null hypothesis is rarely true. Simulations showed that the false conclusion rate for null hypotheses testing is not necessarily high even with low power. The rate at which the lower bound of a confidence interval was higher than the true effect was considerably higher than the nominal level of 0.025. As long as significant differences are given priority for publication, errors in effect size estimation will persist. Online publication has the potential to ameliorate this bias.

# Why Most Published Research Conclusions in Psychology Do Not Have To Be False

Recently Ioannidis and his colleagues (Button *et al*., 2013; Ioannidis, 2005) have shown that low power can cause the published literature to contain mostly false conclusions. A false conclusion is said to occur when a researcher makes a claim about an effect that is contrary to fact. Since no conclusions are drawn when a null hypothesis is not rejected, the proportion of false conclusions is computed based on research for which the null hypothesis is rejected.

The proportion of false conclusions (the false conclusion rate or FCR) can be computed from the proportion of studies for which the null hypothesis is true and the Type II error rate (β). Ioannidis (2005) showed that

$$FCR = 1 - (1 - \beta)R/(R - \beta R + \alpha)$$

where FCR is the probability that a conclusion is false, *R* is the ratio of the number of "true relationships" to "no relationships" in the field in question, β is the Type II error rate, and α is the Type I error rate.

Taking the example presented by Button *et al*. (2013), assume a scientific field with the following characteristics: (a) for each tested effect that is truly non-null there are four tested effects that are truly null so that R

is 1/4 = 0.25, (b) a conclusion is drawn if $p < 0.05$, and (c) $\beta$ is 0.80. The FCR can then be computed as

$$FCR = 1 - (1 - 0.80) \times 0.25 / (0.25 - 0.80 \times 0.25 + 0.05) = 0.50.$$

Therefore, only half the conclusions would be correct.

In some fields, the proportion of null effects is very high. For example, Ioannidis (2005) estimated that in fields for which the study of the relationship between gene polymorphisms and susceptibility to schizophrenia is typical, the pre-study probability that a null hypothesis is true is 0.9999. Assuming power of 0.60, the probability that a conclusion is true is only 0.0012.

Ioannidis (2005) and Button *et al*. (2013) made compelling arguments that the null hypothesis is very often true in epidemiology and some branches of neuroscience. In contrast, many authors including Bakan (1966), Berkson (1938), Cohen (1990), Edwards, Lindman, & Savage (1963), Friston (2012), Greenwald (1975) and Tukey (1991) have argued that in many fields, the null hypothesis is almost never true[1]. For example, Tukey (1991) stated "All we know about the world teaches us that the effects of A and B are always different-in some decimal place-for any A and B. Thus asking 'Are the effects different?' is foolish." (p. 100). Similarly, Cohen (1990) stated "A little thought reveals a fact widely understood among statisticians: The null hypothesis, taken literally (and that's the only way you

can take it in formal hypothesis testing), is always false in the real world." (p. 1308).

Clearly *R,* the ratio of the number of "true relationships" to "no relationships" differs greatly by field. In most subfields of psychology it can be expected that R is very high. For example, it would appear unlikely that any of the following is true: (a) visual and auditory presentation lead to exactly the same level of memory, (b) a person's implicit self-esteem is completely unaffected by how well their partner does on a social intelligence task, and (c) user satisfaction with two website designs is exactly equal. Of course, the null hypothesis is likely to be true for research in fields such as astrology and parapsychology.

This article addresses the problem of false conclusions in subfields in which the null hypothesis is very rarely true. False conclusions in both null hypothesis testing and effect size estimation are considered.

*Null Hypothesis Testing When All Null hypotheses are False*

If two-tailed tests are performed (as will be assumed here) then it is correct to reject every null hypothesis and, at first blush, it would appear that the probability of a false conclusion is 0. This would be true if significant two-tailed tests were not followed by conclusions about the directions of the effects. However, as pointed out by several authors including Bock (1975) and Tukey (1991), a significant result allows the researcher to draw a

confident conclusion about the direction of the effect. This is true for two-tailed as well as one-tailed tests since conducting a two-tailed test at a given $\alpha$ level is equivalent to conducting two one-tailed tests each at the $\alpha/2$ level (Harris, 1997; Kaiser, 1960; Lane, 1993). This procedure allows a conclusion about the direction of the effect while maintaining the Type I error rate at $\alpha$. Although many researchers may not be aware of this logical justification, it is standard practice to infer the direction of an effect from a two-tailed test (Harris, 1997).

Drawing a conclusion about the direction of an effect makes it possible for the inferred direction of the effect to be incorrect. This error has been called an "error of the third kind" (Mosteller, 1948). Therefore false conclusions are possible even if all the null hypotheses are false.

A simulation was performed to explore the prevalence of errors of the third kind. The simulation was based on the following:

1. The experimental design was a two-group randomized design.
2. Effect sizes were defined as the difference between population means divide by the population standard deviation (d).
3. The populations had equal standard deviations and were normally distributed.
4. Effect sizes across experiments were normally distributed with a mean of 0.5 and a standard deviation of 0.143 (so that 0 would be

3.5 standard deviations below the mean) with the constraint that the lowest possible effect size was 0.001.

5. Simulated experiments were conducted for even numbers of subjects per group ranging from 2 to 50, inclusive. For each simulation, an effect size was sampled randomly from the distribution described above. There were 500,000 simulated experiments for each sample size.
6. The 0.05 level of significance was used.

Power, defined here as the proportion of times the null hypothesis was rejected across simulations, was tallied for each sample size. These rejections were categorized as leading to a correct conclusion or an incorrect conclusion.

Figure 1 shows the proportion of significant differences for which a false conclusion was drawn as a function of power. Naturally, the conditional probability of an incorrect conclusion drops very sharply with power. For power of 0.20, which corresponds to an n of 10, the proportion of significant differences in the wrong direction was only 0.01. For the entire set of simulations which used even-numbered n’s from 2 to 50, the proportion of wrong conclusions was 0.013 while the power was 0.41. Note that the power

across the set of experiments is approximately the same as the power with the mean effect size (0.5) and the mean sample size (26) of 0.42.

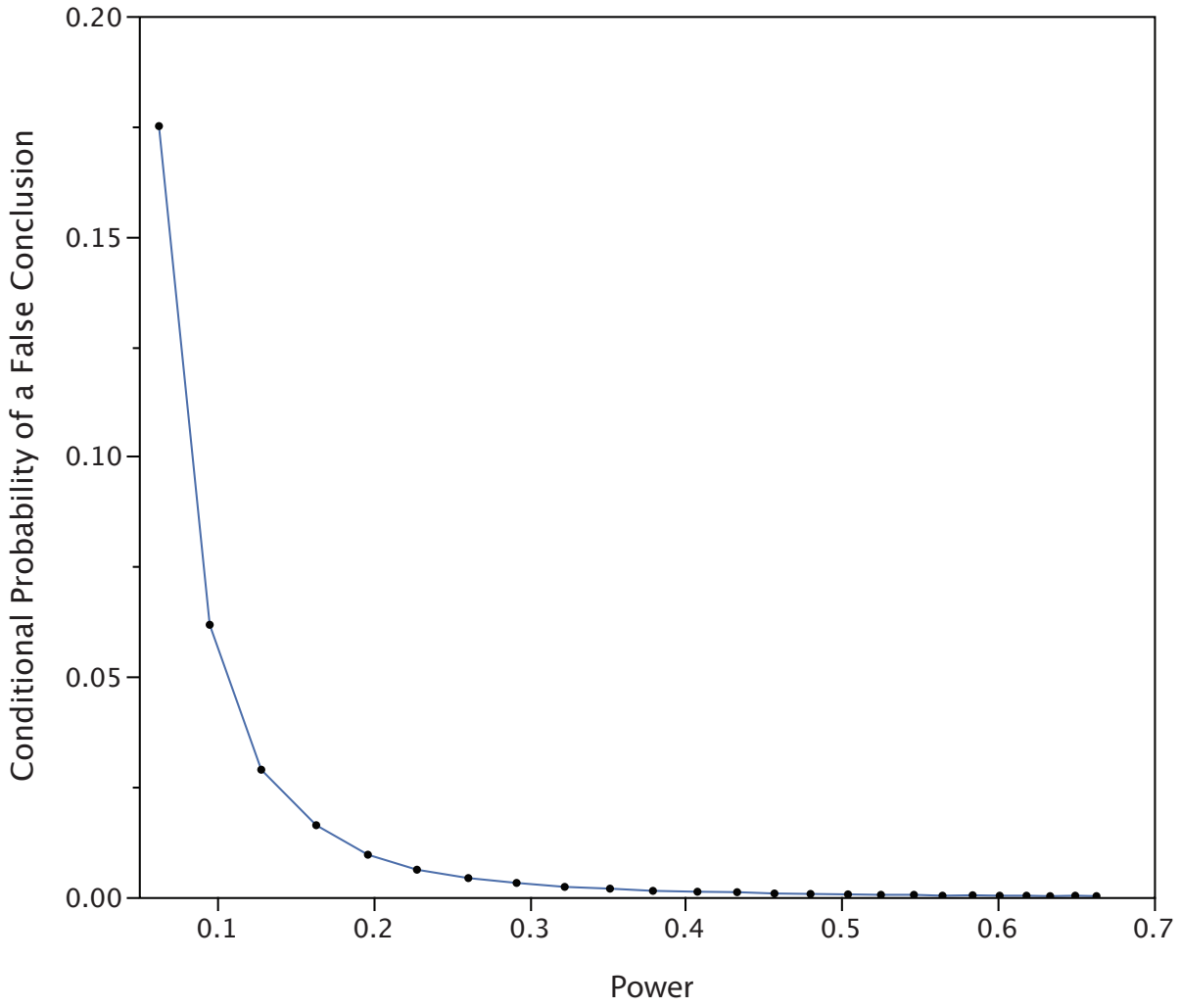


Figure 1. Probability of a false conclusion given that the null hypothesis was rejected.

A second set of simulations equivalent to the first except that the mean effect size was 0.2 and the standard deviation was 0.057 was conducted. The function relating power to the conditional probability of a

false conclusion was essentially the same as the first set of simulations. For this set of simulations, the proportion of false conclusions was 0.07, and the power was 0.11.

These simulations show that the proportion of false conclusions can be relatively small even if power is low. Naturally this result is dependent on the distribution of effect sizes. In his simulation of the publication system, Greenwald (1975) assumed that the effect size was in a null range 20% of the time and greater the other 80% of the time. Applying these assumptions to the present situation, so that 20% of the time the effect size is essentially 0 and the other 80% of the time it is distributed as in the second simulation (mean d = 0.20), the conditional probability of a false conclusion would be (0.8)(.074) + (0.2)(0.50) = 0.16.

Ioannidis (2005) noted that in addition to power and R, bias contributes to the proportion of false conclusions. Ioannidis defined bias as "the combination of various design, data, analysis, and presentation factors that tend to produce research findings when they should not be produced." (p. 697). Simmons, Nelson and Simonsohn (2011) simulated the effects of four biases that are particularly relevant to psychological research: flexibility in (a) choosing among dependent variables, (b) choosing sample size, (c) using covariates, and (d) reporting subsets of experimental conditions. The pessimistic conclusion from their simulation is evident in

their title "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." Fortunately, Simmons *et al*. outlined steps for authors and reviewers that could alleviate most if not all of the effects of these biases. Although these steps involve large changes to standard practice, their implementation is not impractical.

False conclusions resulting from high R and low power are very difficult to eliminate. Unlike the steps outlined by Simmons *et al*. (2011) for reducing bias, it is typically not practical to increase the proportion of studies for which the null hypothesis is false or increase power sufficiently to achieve an acceptable proportion of correct conclusions. Fortunately, for those areas of psychology with very few true null hypotheses, decreasing bias following the steps provided by Simmons *et al*. (2011) has the potential to ensure that the vast majority of conclusions about the direction of effects in psychological research are correct.

Initial findings from the "Many Labs" replication project also suggest that the published literature in at least some fields of psychology need not contain a large proportion of false conclusions. Klein *et al*. (in press) successfully replicated 10 of 13 classic and contemporary effects using 36 independent samples and a total of 6,344 subjects.

*Effect Size Estimation*

The failure of a confidence interval to contain the estimated parameter is an error in effect size estimation[2]. If valid methods are used, these errors will be in line with statistical expectations (e.g., 95% of 95% confidence intervals will contain the estimated parameter).

Most published research contains significant effects and thus is a highly selective subset of conducted research (Greenwald, 1975; Vasilev, 2013). One consequence of this preference for publishing significant effects is that the published literature contains overestimates of effect sizes (Greenwald, 1975; Harris, 1997; Hedges, 1984; Lane & Dunlap, 1978; Schmidt, 1992). For example, one of the simulations conducted by Lane & Dunlap (1978) found that given a population difference between means of 8, a sample size per group of 10, and population standard deviations of 16, the mean difference for results significant at the 0.05 level was more than twice the size of the population difference.

Although this selective bias's effect on effect size estimation has been discussed extensively in the context of meta-analysis (Dalton, Aguinis, Dalton, Bosco, & Pierce, 2012; Hedges, 1984), it does not appear to have been discussed with respect to errors of effect size estimates in individual studies.

The two simulations presented in the section on false conclusions in significance testing also tallied the proportion of times the lower bound of the 95% confidence interval was higher than the true difference between means. Figure 2 shows the results from the simulation for which the mean effect size was 0.5.

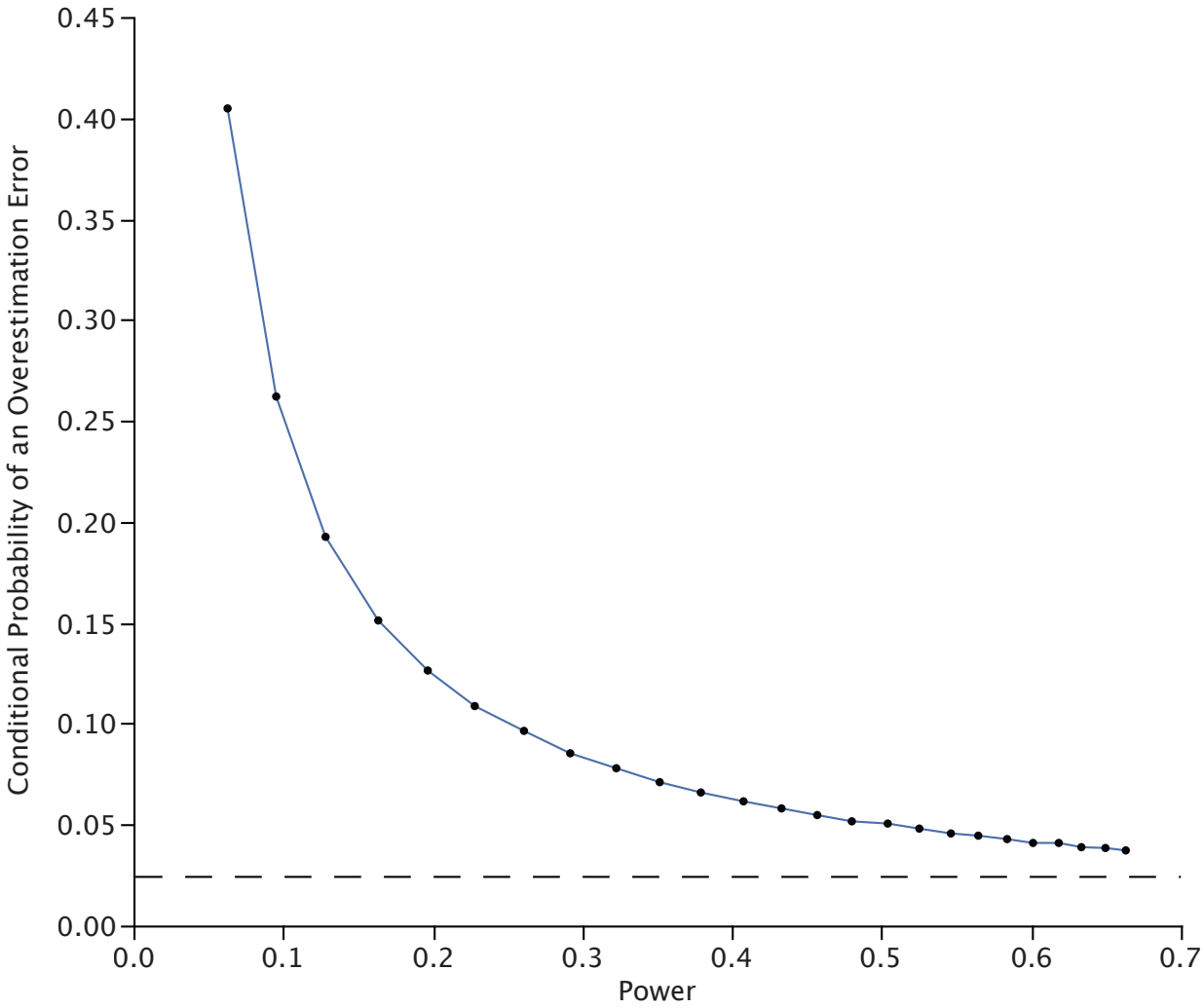


Figure 2. Probability of an overestimation error given that the null hypothesis was rejected. The dotted line represents the nominal level of 0.025.

An inspection of this figure reveals that for low power experiments, the conditional probability of an overestimation error is very high. For example, with power of 0.20 (n=10), the probability of an overestimation error is 0.13 compared to the nominal value of 0.025. Even when power was 0.50 (n = 32), the error rate of 0.05 was twice the nominal rate. Across all simulated experiments, the error rate was 0.092.

Although the argument has been made that publication decisions should not depend heavily on whether or not the null hypothesis has been rejected (Greenwald, 1975; Levine, 2013), there is at least one good reason to continue to base these decisions in part on whether an effect is significant: significant differences allow a confident conclusion about the direction of an effect whereas nonsignificant differences do not (Tukey 1991). Since publication costs are high and, as a result, only a small proportion of methodologically sound studies can be published, it is natural to favor studies that reach a confident conclusion over studies that are unable to do so.

Given the current publication system, this problem is intractable: Published estimates of effect size are and will continue to be too high and should be interpreted cautiously. However, if the major journals did away with their print versions and became entirely online, the marginal cost of publication would be low enough to justify publishing methodologically-

sound studies without regard to the outcome of significance testing. This would greatly reduce the overestimates of effect size and make false conclusions about effect size much less common. Unfortunately, this is very unlikely to happen in the near future.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*, 423-437.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association*, *33*, 526-536.

Bock, R. D. (1975). *Multivariate Statistical Methods in Behavioral Research,* New York: McGraw Hill.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *1*, 1-12.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*, 1304-1312.

Dalton, D. R., Aguinis, H., Dalton, C. M., Bosco, F. A., & Pierce, C. A. (2012). Revisiting the file drawer problem in Meta–Analysis: An assessment of published and nonpublished correlation matrices. *Personnel Psychology*, *65*, 221-249.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193-242.

Friston, K. (2012). Ten ironic rules for non-statistical reviewers. *NeuroImage*, *61*, 1300-1310.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*, 1-20.

Harris, R. J. (1997). Reforming significance testing via three-valued logic. In L. Harlow & S, Mulaik (Eds,) *What if there were no significance tests*, Mahwah. NJ: Erlbaum, 145-174.

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational and Behavioral Statistics*, 9, 61-85.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, *2*, e124.

Kaiser, H. F. (1960). Directional statistical decisions. *Psychological Review*, *67*, 160-167.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., Hasselman, F., Hicks, J. A., Hovermale, J. F., Hunt, J. S., Huntsinger, J. R., IJzerman, H., John, M., Joy-Gaba, J. A., Kappes, H. B., Krueger, L. E., Kurtz, J., Levitan, C. A., Mallett, R., Morris, W. L., Nelson, A. J., Nier, J. A.,

Packard, G., Pilati, R., Rutchick, A. M., Schmidt, K., Skorinko, J. L., Smith, R., Steiner, T. G., Storbeck, J., Van Swol, L. M., Thompson, D., van't Veer, A., Vaughn, L. A., Vranka, M., Wichman, A., Woodzicka, J. A., & Nosek, B. A. (in press). Investigating variation in replicability: A "many labs" replication project, *Social Psychology*.

Lane, D. M. (1993) HyperStat: Hypermedia for Learning Statistics and Analyzing Data., *Academic Press*., New York. NY.

Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology*, *31*, 107-112.

Levine, T. R. (2013). A defense of publishing nonsignificant (ns) results. *Communication Research Reports*, *30*, 270-274.

Mosteller, F. (1948). A k-sample slippage test for an extreme population, *The Annals of Mathematical Statistics*, *19*, 58–65.

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist*, *47*, 1173-1181.

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.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science*, *6*, 100-116.

Vasilev (2013) Negative results in European psychology journals. *Europe's Journal of Psychology*, 9, 717–730.

Footnotes

Correspondence should be sent to:

David Lane
Department of Psychology MS-25
Rice University
Houston, TX 77005

Lane@Rice.edu

1. The argument that the null hypothesis is almost never true is sometimes used to illustrate the supposed pointlessness of null hypothesis testing. For example Cohen (1990) asks "So if the null hypothesis is always false, what's the big deal about rejecting it?" (p. 1308). In other words, significance testing at best, tells you something you already know. However, this argument lacks force since a significant result allows a confident conclusion about the direction of the effect (Bock, 1975; Tukey, 1991).
2. Confidence intervals based on various measures of effect size including the difference between means, Cohen's d and $\omega^2$ are often omitted in published articles. Fortunately, sufficient information to compute confidence intervals is often presented thus allowing interested readers to compute the intervals for themselves. For present purposes, an error in effect size estimation is said to occur if the confidence interval would not

contain the estimated parameter whether or not the confidence interval is reported by the authors.