Studies in Social Science Research
ISSN 2690-0793 (Print) ISSN 2690-0785 (Online)
Vol. 2, No. 2, 2021
www.scholink.org/ojs/index.php/sssr

Original Paper

The Importance of Understanding False Discoveries and the Accuracy Paradox When Evaluating Quantitative Studies

Kirk Davis[1] & Rodney Maiden[2]

[1] Department of Psychology, Troy University, USA

[2] Department of Counseling and Psychology, Troy University, USA

Received: March 8, 2021 Accepted: March 19, 2021 Online Published: March 29, 2021

doi:10.22158/sssr.v2n2p1 URL: http://dx.doi.org/10.22158/sssr.v2n2p1

Abstract

Although the limitations of null hypothesis significance testing (NHST) are well documented in the psychology literature, the accuracy paradox, which concisely states an important limitation of published research, is never mentioned. The accuracy paradox appears when a test with higher accuracy does a poorer job of correctly classifying a particular outcome than a test with lower accuracy, which suggests that a reliance on accuracy as a metric for a test's usefulness is not always the best metric. Since accuracy is a function of type I and II error rates, it can be misleading to interpret a study's results as accurate simply because these errors are minimized. Once a decision has been made regarding statistical significance, type I and II error rates are not directly informative to the reader. Instead, false discovery and false omission rates are more informative when evaluating the results of a study. Given the prevalence of publication bias and small effect sizes in the literature, the possibility of a false discovery is especially important to consider. When false discovery rates are estimated, it is easy to understand why many studies in psychology cannot be replicated.

Keywords

accuracy paradox, false discovery, false omission, statistical significance, NHST

1. Introduction

1.1 Understanding the Limitations of NHST

The goal of this paper is not to rehash the limitations associated with null hypothesis significance testing (NHST). The limitations of NHST are well known across diverse fields such as psychology (Cohen, 1990, 1994; Schmidt, 1996), education (Robinson & Levin, 1997; Thompson, 1996), and medicine (Ioannidis, 2008). Instead, the goal of this paper is to use the accuracy paradox to demonstrate why false discoveries, not false negatives, should be considered when evaluating published research, especially if a claim of statistical significance has been made. Specifically, the accuracy paradox reveals that it can be misleading to interpret a study's results as accurate simply because type I and II errors have been minimized.

Studies that use null hypothesis significance testing (NHST) rely on *p*-values, allowing the researcher to determine if their results are statistically significant. In essence, whenever *p*-values are reported, a classification system is being used. Since this classification system, statistically significant or not, remains popular with quantitative studies in psychology (Counsell & Harlow, 2017), it is crucial to understand its limitations. One limitation, highlighted by the accuracy paradox, is that a test's accuracy is not always relevant after the data have been classified as statistically significant or not. Small *p*-values imply that a study's results are accurate, but and understanding of the accuracy paradox will reveal that this is not always the case. The accuracy paradox reveals that factors important before classification, such as limiting type I and type II errors, are less important than factors, specifically false discoveries and false omissions, after classification. As such, understanding the accuracy paradox, along with false discovery and false omission rates, is vital to understanding published quantitative research. Given the prevalence of publication bias and small effect sizes in the psychology literature, the possibility of a false discovery is especially important to consider.

1.2 The Accuracy Paradox

A key limitation of NHST can be easily understood by examining the so-called accuracy paradox. The accuracy paradox, a term well known in the field of machine learning, states that high accuracy is not necessarily an indicator of high classifier performance (Valverde-Albacete & Peláez-Moreno, 2014; Afonja, 2017). As such, a test or model with higher accuracy does not always do a better job of correctly classifying a particular outcome than a test or model with lower accuracy. Since NHST is essentially a classification system where results are classified as statistically significant or not, this means a test's accuracy is not always informative after the data have been classified as statistically significant or not. This suggests that a reliance on accuracy as a metric for a test's usefulness is not always the best metric. Instead, false discoveries and false omissions should be considered.

Since a test's accuracy is a function of the number of false positives (type I error rate) and false negatives (type II error rate), it can be misleading to interpret a study's results as accurate simply because these errors are minimized. The implication is that the wrong information, such as an overreliance on small *p*-values, may be used as the most important criterion when evaluating empirical studies if the reader does not understand the difference between a false positive and a false discovery.

If you are a counselor or a psychologist, chances are you have never heard of the accuracy paradox since the term originates from the field of machine learning. There is no mention of the accuracy paradox in the literature, no statistics textbooks in the social sciences reference the term, and an informal survey of 32 counselors and psychologists with a Ph.D. stated that they were not familiar with the paradox. This is not unexpected since the so-called paradox is more readily seen when the data collected can easily be classified as a correct or incorrect decision, which is more common in fields such as medicine and machine learning. For example, the results of a pregnancy test are easily classified as either correct or incorrect given time or additional tests. Data collected in the social sciences, on the other hand, are usually analyzed at the group-level rather than the individual-level, and the results of these studies are usually expressed as probabilities (*p*-values) rather than certainties. As such, false discovery rates are not as readily available.

1.3 False Positives and False Discoveries

When NHST is used, a false positive or type I error is a classification error that occurs when the null hypothesis is incorrectly rejected. In other words, the data are classified as statistically significant but no effect exists. In contrast, a false discovery occurs when an outcome that has been classified as statistically significant does not exist, and can only occur after the data have been classified as statistically significant, which is the classification most readers encounter when reviewing the literature.

When NHST is used, a false negative or type II error is a classification error that occurs when the null hypothesis is incorrectly retained. In other words, the data are classified as not statistically significant but an effect exists. In contrast, a false omission occurs when an outcome that has not been classified as statistically significant does exist, and can only occur after the data have been classified as not being statistically significant. Thus, a false omission occurs when an outcome that has been determined not to be statistically significant does exist.

Although the differences between false positives and false discoveries, along with false negatives and false omissions, are subtle and often difficult to discern, the differences are critical. Unfortunately, no mention of the accuracy paradox and its relevance to false discoveries and false omissions is discussed in the psychology literature. As such the goal of this paper is to emphasize the importance of these two errors, especially false discoveries, when evaluating published findings from an empirical study by demonstrating that a test’s accuracy is not always relevant after the data have been classified.

1.4 The Relevance of False Discoveries

Two factors, publication bias and small effect sizes, make false discoveries of particular interest when considering statistically significant results. According to Dickersin (1990) Publication bias occurs when decisions to submit or publish research is based on the direction or strength of a study’s results. In other words, research that reports results that are statistically significant are more likely to be published or submitted, especially if the study is original. In a study conducted by Kühberger, Fritz, and Scherndl (2014), 1,000 psychological articles from all areas of psychological research were examined. The authors concluded the negative correlation between effect size and sample size and the biased distribution of *p*-values were “indicate pervasive publication bias in the entire field of psychology.” The second element of publication bias arises due to a reluctance to report findings from replicated studies. A study conducted by Martin and Clarke (2017) found that only 3% (33/1151) of the journals explicitly stated that they accepted replicated studies. However, 33% (380/1151) *33%* emphasized the need for original studies. This means that studies that do not support an original study’s results are not likely to get published, and authors know this (Rosenthal, 1979; Franco et al., 2014).

In addition to publication bias, small effect sizes make false discoveries more interesting than false omissions since small effect sizes are more likely to lead to false discoveries than false omissions. Although effect sizes can be difficult to objectively quantify as small, medium or large, a study conducted by Shafer and Schwartz (2019) reviewed over 900 psychology articles and found median effects sizes for studies that preregistered and non-preregistered to be to be r = 0.16 and r = 0.36 respectively. This means 13% of the variance was accounted for in the studies that did not preregister and only 2.6% of the variance was accounted for in the preregistered studies. Two independent studies, one that reviewed the social psychology literature (Richard et al., 2003) and one that reviewed the personality literature (Fraley & Marks, 2007), both found an average effect size of only r = .21 or r^2 = .044. Whether the effect size is determined to small or large, most of the variance is unaccounted for in the psychological literature.

Not all journals accept *p*-value reporting from HNST. According to Trafimow and Marks (2015), the journal of *Basic and Applied Social Psychology* banned the use of *p*-values and null hypothesis significance testing, and many prominent researchers have recommended the removal of statistical significance (Amrhein & Greenland, 2018; Wasserstein, Schirm, & Lazar, 2019). However, most journals still report empirical results that rely on *p*-values. In a review of the literature conducted by Bakker and Wicherts (2011), the majority of empirical articles in psychology report statistical results that are the outcome of NHST, which ultimately report *p*-values. Unfortunately, studies reporting small *p*-values, typically less than .05, are more likely to be published than studies reporting p-values greater than .05. Small p-values suggest that a study's results are accurate, but this is not always the case which can be seen when the accuracy paradox is considered.

1.5 Estimating False Discoveries

Consider the following medical analogy where a physician informs you that you have tested positive for a disease. Do you actually need to know the probability of making a type I error/false positive of the test or the probability of making a type II error/false negative? Probably not. How about the accuracy (the overall probability that the test provided a correct diagnosis) of the test? You may think you do, but the accuracy of the test won't provide the information you want either. Given that you have tested positive, you want to know the probability that you actually have the disease. In other words, you want to know how often the test provides a positive result when you don't actually have the disease, which is its false discovery rate (Trevethan, 2017). Conversely, if you tested negative, you want to know the probability that you actually don't have the disease. To understand why false discovery rates, not false positive rates, provide the information you need, consider the following example.

Consider the confusion matrix in Table 1 inspired by a paper published in the journal *Psychological Science in the Public Interest* (Gigerenzer et al., 2007) which displays the hypothetical results of a test administered to 1,000 people. The data are consistent with a false positive rate of 5%, a false negative rate of 20%, and a prevalence rate of 1%, which are analogous to a study with an alpha = .05, power

= .8, and an effect size of .01. The overall accuracy of the test based on these data is 948/1000 or 94.8%. Although the test provides the correct classification 94.8% of the time, the accuracy of the test is misleading for someone classified as having the disease. Given someone has been classified as having the disease, the probability that the person actually has the disease is only 8/58 or 14%, which makes the false discovery rate 50/58 or 86%. Despite the high accuracy, the false discovery rate is also very high due to the low prevalence/effect size, which reveals the accuracy paradox and illustrates the importance of considering the possibility of a false discovery once a statistically significant classification has been made. Given that you have tested positive, you want to know the probability that you actually have the disease. In other words, you want to know how often the test provides a positive result when you don't actually have the disease, which is its false discovery rate (Trevethan, 2017).

Table 1. Confusion Matrix Displaying Disease State as a Function of Test Classification with Small Effect Size and High Power

	Disease State		
Classification	Disease Present	Disease Absent	Overall
Positive	8 True Positives	50 False Positives	50
Negative	2 False Negatives	940 True Negatives	940
Overall	10	990	1000

Admittedly, the prevalence/effect size (.01) in Table 1 is smaller than the typical effect size of r^2=.044 in psychology, but the sensitivity/power (8.0), is much greater than typically reported in the psychology literature. According to (Sedlmeier & Gigerenzer, 1989), the average power of a psychological study to detect a medium effect (r = .3 or r^2 = .09) is only .5. Using these parameters (effect size = .09 and power = .5), the data in Table 2 reveal an overall accuracy of 91% ((45 + 865)/1000)) but a false discovery rate of 50% (45/90). Again, the overall accuracy of the test is high, but the false discovery rate, although lower than in Table 1, is also high despite the effect size and power parameters being larger than typically found in the literature.

Table 2. Confusion Matrix Displaying Disease State as a Function of Test Classification with Average Effect Size and Power

	Disease State		
Classification	Disease Present	Disease Absent	Overall
Positive	45 True Positives	45 False Positives	90
Negative	45 False Negatives	865 True Negatives	910
Overall	90	910	1000

Consider Table 3, which illustrates the results of a test with high power (.80) and a large effect size (.25). In this example, which is rare in the psychological research, accuracy is again high at 91.2% ((200 + 712/1000)), and the false discovery rate of 16% (38/238) is much smaller than seen in Tables 1 and 2. However, despite having parameters that exceed those typically found in the psychological literature, the false discovery rate is over three times larger than the false positive rate of 5%, highlighting the importance of considering false discoveries when evaluating published research with statistically significant findings.

Table 3. Confusion Matrix Displaying Disease State as a Function of Test Classification with Large Effect Size and High Power

	Disease State		
Classification	Disease Present	Disease Absent	Overall
Positive	200 True Positives	38 False Positives	238
Negative	50 False Negatives	712 True Negatives	762
Overall	250	750	1000

1.6 Limitations

The data reported in Tables 1-3 reflect a hypothetical test's sensitivity, specificity, and disease prevalence when 1,000 individuals are classified. In this situation, these terms are analogous to power, confidence level, and effect size, respectively, but they cannot be interpreted as directly equivalent. As such, the data in Tables 1-3 can only be used to estimate the probability of the false discovery rate when NHST results are reported. Additionally, not all studies report both measures of effect size and power. If these factors are not reported, false discovery rates cannot be estimated for a specific study. Furthermore, the use of effect sizes can only be used if the effect is expressed a proportion of variance accounted for, which excludes popular measures such as Cohen's *d* and Hedge's *g*.

2. Conclusion

It can be misleading to interpret a study's results as accurate simply because type I and type II errors are minimized. Once a decision has been made regarding statistical significance, false discovery and false omission rates are more informative when evaluating the results of a study. Given the prevalence of publication bias and small effect sizes in the literature, the possibility of a false discovery is especially relevant. When estimated using effect sizes and power levels commonly found, the false discovery rate for a typical psychological study is 50%. Therefore, when evaluating a study that has reported statistically significant findings, the possibility a false discovery should be considered, which can be estimated if power and effect sizes are reported.

References

Afonja, J. (2017, December 10). *Accuracy paradox*. Toward Data Science. Retrieved from https://www.towardsdatascience.com/accuracy-paradox-897a69e2dd9b

Amrhein, V., & Greenland, S. (2018). Remove, rather than redefine, statistical significance. *Nature Human Behavior*, *2*(4). https://doi.org/10.1038/s41562-017-0224-0

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*(12), 1304-1312. https://doi.org/10.1037/0003-066X.45.12.1304

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, *49*(12), 997-1003. https://doi.org/10.1037/0003-066X.49.12.997

Counsell, A., & Harlow, L. L. (2017). Reporting practices and use of quantitative methods in Canadian journal articles in psychology. *Canadian Psychology/Psychologie canadienne*, *58*(2), 140-147. https://doi.org/10.1037/cap0000074

Dickersin K. (1990). The existence of publication bias and risk factors for its occurrence. *JAMA*, *263*(10), 1385-1389. https://doi.org/10.1001/jama.1990.03440100097014

Fraley, R. C., & Marks, M. J. (2007). The null hypothesis significance-testing debate and its implications for personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of Research Methods in Personality Psychology* (pp. 149-169). The Guilford Press.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(19). https://doi.org/10.1126/science.1255484

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, *8*(2), 53-96. https://doi.org/10.1111/j.1539-6053.2008.00033.x

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*(5), 640-648. https://doi.org/10.1097/EDE.0b013e31818131e7

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *PLoS ONE*, *9*(9), Article e105825. https://doi.org/10.1371/journal.pone.0105825

Martin, G. N., & Clarke, R. M. (2017). Are psychology journals anti-replication? A snapshot of editorial practices. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00523

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*(4), 331-363. https://doi.org/10.1037/1089-2680.7.4.331

Robinson, D. H., & Levin, J. R. (1997). Research news and comment: Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher*, *26*(5), 21-26. https://doi.org/10.3102/0013189X026005021

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638-641. https://doi.org/10.1037/0033-2909.86.3.638

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*(2), 115-129. https://doi.org/10.1037/1082-989X.1.2.115

Sedlmeier, P., & Gigerenzer, G. (1992). Do studies of statistical power have an effect on the power of studies? In A. E. Kazdin (Ed.), *Methodological issues & strategies in clinical research* (pp. 389-406). American Psychological Association. https://doi.org/10.1037/10109-032

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, *25*(2), 26-30. Retrieved February 24, 2021, from http://www.jstor.org/stable/1176337

Trafimow, D., & Marks, M. (2015) Editorial, *Basic and Applied Social Psychology*, *37*(1), 1-2. https://doi.org/10.1080/01973533.2015.1012991

Trevethan R. (2017). Sensitivity, specificity, and predictive Values: Foundations, pliabilities, and pitfalls in Research and Practice. *Frontiers in public health*, *5*, 307. https://doi.org/10.3389/fpubh.2017.00307

Valverde-Albacete, F. J., & Peláez-Moreno, C. (2014). 100% Classification accuracy considered harmful: The normalized information transfer factor explains the accuracy paradox. *PLoS ONE*, *9*(1), e84217. https://doi.org/10.1371/journal.pone.0084217

Wasserstein, R. L., Schirm, A. L., & Lazar, A. N. (2019). Moving to a world beyond “$p < 0.05$”. *The American Statistician*, *73*(1). https://doi.org/10.1080/00031305.2019.1583913

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS ONE*, *6*(11), e26828. https://doi.org/10.1371/journal.pone.0026828