Running head: R-Index

Quantifying Statistical Research Integrity: The Replicabilty-Index

Ulrich Schimmack

University of Toronto Mississauga

November, 2014

*Author's Not*e. I would like to thank Gregory Francis, Julia McNeil, Amy Muise, Michelle Martel, Elizabeth Page-Gould, Geoffrey MacDonald, Brent Donnellan, David Funder, Michael Inzlicht, and the Social-Personality Research Interest Group at the University of Toronto for valuable discussions, suggestions, and encouragement. Correspondence should be sent to Ulrich Schimmack, Department of Psychology, University of Toronto Mississauga, email: uli.schimmack@utoronto.ca.

Abstract

Researchers are competing for positions, grant money, and status. In this competition, researchers can gain an unfair advantage by using questionable research practices (QRPs) that inflate effect sizes and increase the chances of obtaining stunning and statistically significant results. To ensure fair competition that benefits the greater good, it is necessary to detect and discourage the use of QRPs. To this aim, I introduce a doping test for science; the replicability index (R-index), a quantitative measure of research integrity that can be used to evaluate the statistical replicability of a set of studies (e.g., journals, individual researchers' publications). I first discuss existing approaches to the detection of biased results and point out their limitations as a measure of scientific integrity. I then show how the R-index reveals the increase in the use of QRPs by comparing the R-index in the Journal of Abnormal and Social Psychology in 1960 to the R-index in the Attitudes and Social Cognition section of the Journal of Social and Personality Psychology in 2011. Like doping tests in sports, the availability of a scientific doping test should deter researchers from engaging in practices that advance their careers at the expense of everybody else. Demonstrating replicability should become an important criterion of research excellence that can be used by funding agencies and other stakeholders to allocate resources to research that advances science.

Keywords: Power, Publication Bias, Significance, Credibility, Sample Size, Questionable Research Methods, Replicability, Statistical Methods

Quantifying Statistical Research Integrity: The R-index

"Questionable research practices are the steroids of scientific competition" (John et al., 2012)

It has been known for decades that published results are likely to be biased in favor of authors' theoretical inclinations (Sterling, 1959). The strongest scientific evidence for publication bias stems from a comparison of the rate of significant results in psychological journals and the statistical power of published studies. Statistical power is the long-run probability to obtain a significant result, when the null-hypothesis is false (Cohen, 1988). The typical statistical power of psychological studies has been estimates to be around 60% (Sterling, Rosenbaum, & Weinkam, 1995). However, the rate of significant results in psychological journals is over 90% (Sterling, 1959; Sterling et al., 1995). The discrepancy between these estimates of power reveals that published studies are biased and that some findings may be simply false positive results, whereas other studies report inflated effect size estimates.

It has been overlooked that estimates of statistical power are also inflated by the use of questionable research methods. Thus, the commonly reported estimate that typical power in psychological studies is 60% is an inflated estimate of true power (Schimmack, 2012). If the actual power is less than 50%, it means that a typical study in psychology has a larger probability to fail (produce a false negative result) than to succeed (rejecting a false null-hypothesis). Conducting such low powered studies is extremely wasteful. Moreover, few researchers have resources to discard 50% of their empirical output. As a result, the incentive for the use of

questionable research practices that inflate effect sizes is strong. Not surprisingly, the use of questionable research practices is common (John et al., 2012). More than 50% of anonymous respondents reported selective reporting of dependent variables, dropping experimental conditions, or not reporting studies that did not support theoretical predictions.

The widespread use of QRPs undermines the fundamental assumption of science that scientific theories have been subjected to rigorous empirical tests. In violation of this assumption, QRPs allow researchers to find empirical support for hypotheses even when these hypotheses are false. The most dramatic example was Bem's (2011) infamous evidence of time-reversed causality (e.g., studying after a test can improve test performance). Although Bem reported nine successful studies, subsequent studies failed to replicate this finding and raised concerns about the integrity of Bem's studies (Schimmack, 2012). One possibility for false positive results could be that a desirable outcome occurred by chance and a researcher mistakes this fluke finding as evidence that a prediction was true. However, a fluke finding is unlikely to repeat itself in a series of studies. Statistically, it is highly improbable that Bem's results are simple type-I errors because the chance of obtaining 9 out of 10 type-I errors with a probability of .05 is less than 1 out of 53 billion (1 / 53,610,771,049). This probability is much smaller than the probability of winning the lottery (1 / 14 million). It is also unlikely that Bem simply failed to report studies with non-significant results because he would have needed 180 studies (9 x 20) to obtain 9 significant results because a type-I error of 5% implies that a significant result will occur, on average, for every 20 studies. With sample sizes of about 100 participants in reported studies, this would imply that Bem tested 18,000 participants. It is therefore reasonable to conclude that Bem used questionable research methods to produce his implausible and improbable results.

Although the publication of Bem's article in a flagship journal of psychology was a major embarrassment for psychologists, it provided an opportunity to highlight fundamental problems in the way psychologists produced and published empirical results (Wagenmakers @@@). There have been many valuable suggestions and initiatives to increase the integrity of psychological science (****). In this article, I want to propose another simple solution that can increase the integrity and productivity of science: I suggest that scientific organizations ban the use of questionable research practices, just like sports organizations ban the use of performance enhancing substances. Unfortunately, the world of professional sports also shows that doping bans are ineffective unless they are enforced by regular doping tests. Thus, a ban of questionable research practices needs to be accompanied by objective tests that can reveal the use of questionable research practices. The main purpose of this article is to introduce a statistical test that reveals the use of questionable research practices that can be used to enforce a ban of such practices. This quantitative index of research integrity can be used by readers, editors, and funding agencies to ensure that only rigorous empirical studies are published or funded. Before I introduce the integrity index, I review existing statistical tests of publication bias and point out their limitations as quantitative measures of research integrity.

Fail-Safe-N

The problem of publication bias was first recognized when psychologists started to conduct meta-analyses in the 1970s. The main purpose of a meta-analysis is to conduct a more powerful test of hypotheses by pooling the results of several small studies. To achieve this goal, it is necessary that the set of original studies is an unbiased sample of all studies that tested a specific hypothesis. If the set of original studies is biased in favor of studies that support a hypothesis, a meta-analysis would produce biased results. Rosenthal (1979) developed the Fail-

Safe-N approach to examine biases in meta-analyses. I illustrate Fail-Safe N with Bem's (2011) studies of extrasensory perception.

Bem reported 10 statistical results (see Bem, 2011, Table 7). The first step in calculating Fail-Safe-N is to convert p-values into z-scores using the inverse function of the standard normal distribution. For example, a one-tailed p-value of .010 corresponds to a z-score of 2.326 because 90% of the normal distribution is on the left side of z = 2.326 and 10% of the distribution is on the right side of it. The second step is to sum the z-values. For Bem's set of studies the sum is 21.171. The third step is to divide the sum score by the square root of the number of studies ($10^{1/2} = 3.162$). The resulting z-score is $z_{sum} = 6.695$. Fail-Safe-N uses this z-score to calculate how many studies with a null-result (z = 0) would have to be missing to produce an overall z-score that is no longer significant. The formula for Fail-Safe N is $N_{fs} = (z_{sum} / 1.65)^2 - k$, where 1.65 is the z-score for alpha = .05 (one-tailed) and k is the number of z-scores that were summed. For Bem's set of studies Fail-Safe N equals 155 studies. It seems implausible that Bem conducted 155 additional studies with non-significant results. Thus, Fail-Safe N would suggest that Bem's studies provide statistically significant evidence for time-reversed causality even if he did not report all studies.

Fail-Safe N has numerous limitations. One major limitation is that Fail-Safe N does not consider the use of questionable research methods. If questionable research methods were used to obtain significant results, Fail-Safe N overestimates the number of studies that would be required to render published evidence inconclusive. For example, Simons, @@@ (2011) demonstrated that questionable research practices can boost power by over 50%. Thus, it takes only 2 studies rather than 20 studies to produce a significant result. With a false-positive rate of 50%, Bem would have needed only 18 studies to produce 9 studies with significant results using

QRPs, while dropping studies that failed to produce significant results even with the use of QRPs. Another major limitation of Fail-Safe N as a measure of integrity is that it does not test whether a set of studies is biased or not. Rather, the main aim is to assess whether the evidence is sufficient to reject the null-hypothesis even if an unknown amount of bias contaminated the data.

p	z
.010	2.326
.009	2.336
.007	2.457
.014	2.197
.014	2.197
.037	1.787
.039	1.762
.096	1.305
.029	1.896
.002	2.878
Sum	21.171

The weaknesses of Fail-Safe N may explain why Fail-Safe-N typically fails to reveal biases, just like it failed to detect that Bem's (2011) results are biased and fail to replicate in other laboratories. Thus, Fail-Safe-N is akin to a doping test that does not detect the use of performance enhancing drugs. This creates the appearance of a clean sport, when doping is rampant. In conclusion, Fail-safe N failed as a doping test of science and may rather have created the illusion of integrity.

Eggert/Stanley Bias-Regression

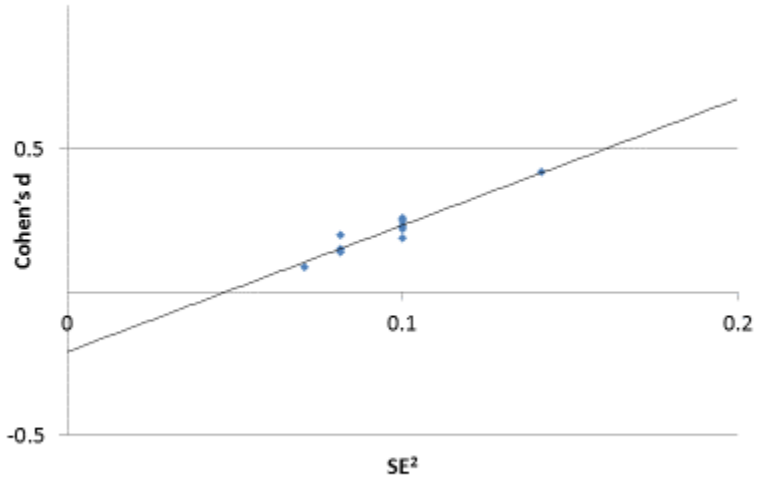
The first true bias test for meta-analyses examines the relation between sample size and effect size (Egger, Smith, Schneider, & Minder, 1997; Stanley & Doucouliagos, 2009). The logic of this test relies on the fact that it is easier to obtain statistically significant results in larger

samples because sampling error decreases with increasing sample size. In contrast, small samples require large effect sizes to produce significant results. As a result, it is more likely that researchers sue QRPs in small samples because the true effect size may be smaller than the effect size needed to produce a significant result. For example, a between-group study with two conditions and 20 participants in each cell, requires an effect size of d = .64 (i.e., .64 standard deviations mean difference between groups) to be significant (d = .64; SE = .32, t(38) = .64/.32 = 2.02, *p* = .05 (two-tailed). If the true effect size is smaller (e.g., a moderate effect size of d = .5), significance can only be achieved with the help of luck or questionable research practices that lead to inflated effect-size estimate in the sample.

The use of QRPs in smaller samples can be detected because it would produce a negative correlation between observed effect sizes and sample sizes or by means of a positive correlation between observed effect sizes and sampling error. Using sampling error as the predictor has the additional advantage that it becomes possible to obtain a bias-corrected estimate of the true effect size (Stanley ***). I used Bem's (2011) data to illustrate the approach. The first step is to compute sampling error as a function of sample size. The next step is to regress the effect sizes onto sampling errors. A significant positive regression coefficient suggests the presence of bias. As sampling error is an inverse function of sample size, sampling error approaches zero as sample size increases towards infinity. Thus, the intercept provides an estimate of the effect size when sampling error is zero; in other words it is an unbiased estimate of the effect size in the population.

However, the relation between sampling error and bias is non-linear because questionable research practices are no longer needed when samples are large enough to have high statistical power. Based on simulations, Stanley recommends to regress effect sizes on the squared standard

error to estimate the population effect size. Figure 1 shows the results for Bem's (2011) data. The square of the standard error (SE^2) is a nearly perfect predictor of the variation in effect sizes (Cohen's d), $r^2 = .92$ and the relationship is highly significant despite the small sample size, $r(10) = .96$, $t(8) = 9.64$, $p = 0.000011$. This strong relationship between sampling error and effect size suggests that questionable research practices influenced the results. The estimate of the true effect size is -.207 with a 95% confidence interval that ranges from -.31 to -.10. Thus, while the original effect size suggested a positive ESP effect with an effect size of $d = .22$, the unbiased effect size estimate is negative. It is unlikely that the true effect size is negative, but the finding that the sign of the effect changes when a bias-correction is applied casts doubt on the credibility of the evidence in the original article.



Meta-regression is a powerful tool to examine biases in meta-analysis, but it has a number of limitations as a quantitative measure of research integrity. One problem is that the method requires a relatively large number of studies to demonstrate a reliable association

between effect size and sample size. For example, 67 studies would be needed to have 80% power to detect a moderate correlation of *r* = .30 between effect size and sampling error. Another problem is that the regression approach requires meaningful variability in sample sizes. If all studies used small samples and questionable research methods, the set of studies lacks unbiased studies with large samples that can reveal biases in the studies with smaller samples.

A more serious problem is that meta-regression may lead to false conclusions when effect sizes are heterogeneous. For example, a set of studies might examine small, moderate, and large effects (d = .2, .5, & .8). Researchers might conduct a prior sample sizes to plan their studies and would recruit large, moderate, and small samples, respectively (Ns = 788, 128, 52). This would produce a strong correlation between effect size and sampling error. However, the correlation would reveal good research practices (matching sampling error to effect sizes) rather than the use of questionable research practices (matching effect sizes to sampling error). To use the analogy of doping in sports, a strong correlation between the importance of a competition and performance can reveal the use of doping or it may suggest that athletes train harder for more important competitions; or both.

In conclusion, the regression approach is most suitable for meta-analyses of studies that examine a common effect, but it is problematic to use this method to examine the integrity of sets of studies with heterogeneous effect sizes. Importantly, meta-regression was developed for the former purpose and the authors never proposed to use it as a quantitative measure of research integrity.

P-Curves

The p-curve method relies on the distribution of p-values to examine the integrity of a set of studies (Simonsohn, Nelson, & Simmons, 2013). The foundation of this test is the observation

that p-values are uniformly distributed when the null-hypothesis is true. This is a bit counterintuitive, but a simple example is helpful to illustrate this point. A p-value of .05 means that 5% of the area under the curve is in the tails. A p-value of .04 means that 4% of the area under the curve is in the tails. As a result, 1% (i.e., 5% - 4%) of the area under the curve is between p = .05 and p = .04. This holds for any difference in p-values (e.g., .04 to .03, .03 to .02, .01 to .00). Thus, a p-value is as likely to be between .04 and .03 as it is to be between.01 to .00; when the null-hypothesis is true. When the null-hypothesis is false, the probability of p-values in the tails increases as a function of the power of a study and p-values in the tail are no longer uniformly distributed. Rather, p-values are now right-skewed so that p-values in the interval between .05 and .04 are less likely to occur than p-values in the region between .01 and .00.

The P-curve method uses two statistical tests to test the null-hypothesis that a set of p-values is uniformly distributed. One approach divides the tail regions into two equal intervals. One interval covers the range of p-values from .050 to .025. The other interval covers the area from .025 to 0. A chi-square test is used to test the null-hypothesis that p-values are equally distributed across the two intervals. The test can yield three outcomes. First, the null-hypothesis is not rejected. In this case, the test is inconclusive. Second, the test can be significant because p-values in the upper interval (i.e., .050 > p > 025) are more frequent than those in the lower interval (i.e., .025 > p > 000). This outcome suggests that questionable research practices contributed to the p-values more than a real effect. Third, the test can be significant because p-values in the lower interval are more frequent than those in the upper interval. This outcome is interpreted as evidence that a real effect contributed to the p-values more than questionable research practices.

Table 2 illustrates this approach with Bem's (2011) data. Out of 10 studies, one study produced a non-significant result. Non-significant results are discarded because the method only considered p-values between .05 and .00. The other 9 studies yielded 7 p-values in the lower interval and 2 p-values in the upper interval. The binomial probability of 7 or more p-values in the extreme region is $p = .090$. Thus, the P-curve method suggests that Bem's results are more strongly influenced by a true effect than by questionable research methods, if a criterion of $p < .10$ is used or the test would be inconclusive if a more stringent criterion were used.

p	Region
.010	1
.009	1
.007	1
.014	1
.014	1
.037	0
.039	0
.096	-
.029	1
.002	1

The test does not suggest that questionable research practices contributed to Bem's results. The failure of the p-curve method to reveal the use of questionable research methods in Bem's studies can be attributed to several shortcomings of this approach. First, the test has low power to reject the null-hypothesis. To detect a moderate effect size (deviation from equal proportions) with a liberal false-positive criterion of 10%, 18 studies are needed to achieve 80% power. The second problem of this approach is the assumption that questionable research practices are likely to produce p-values just below the critical value (i.e, in the .050 to .0250 range). However, some questionable research practices capitalize on chance and chance produces p-values that are equally distributed across the full range of p-values from .050 to .000.

For example, the use of multiple depended variables allows researchers to capitalize on chance with each additional variable that is being tested and each test is as equally likely to produce a p-value in the upper or the lower interval. Finally, if the test reveals significantly more p-values in the lower interval, it merely shows that the set of studies contains some studies with real effects. This is particularly problematic when a set of studies has heterogeneous effect sizes. For example, a set of 100 studies may contain 20 studies with high statistical power that produced highly significant results. All of these studies contribute to the frequency of studies with p-values in the lower interval. The remaining 80 studies relied on questionable research methods to produce significant results, with half of the p-values (40) falling into the upper interval and half (40) falling into the lower interval. The resulting distribution of 40 p-values in the upper interval and 60 p-values in the lower interval is statistically significant. Chi-square (df = 1) = 4, p = .046. The test correctly reveals that some of the studies have evidential value; that is, reveal correct rejections of the null-hypothesis. However, the test does not reveal that 80% of the studies used questionable research methods. The reason is that the test provides no information about the use of questionable research methods.

A second test examines how extreme p-values are. Each p-value is compared to the critical value for statistical significance; typically p = .05. More extreme p-values are more likely to be based on real effects. Extremity is measured with a ratio of the p-value divided by the significance criterion. For example, a p-value of .025 yields a ratio of .025/.05 = .5. A p-value of .001 yields a ratio of .001/.05 = .02. These ratios are log-transformed using the formula ln (p/.05) * -2. The sum of these values is chi-square distributed with the number of studies times two as the degrees of freedom. For the two p-values of .025 and .001, the chi-square test suggests that the studies have evidential value, chi-square (df = 1) = 9.21, p = .056. This test is

more powerful to reject the null-hypothesis that p-values are uniformly distributed. However, the gain in power is achieved because a few highly significant p-values can compensate for many p-values that were obtained by means of questionable research practices. For example, a set of studies may contain 20 studies with strong significant results, p = .001. The remaining 80 studies were intensely p-hacked and produced a just significant p-value of .03. The chi-square test correctly shows that the set of studies contains some studies with credible findings, chi-square (df = 200) = 238, p = .033. However, the test does not reveal that 80% of the studies were obtained by means of questionable research practices.

In conclusion, p-curves reveal whether a set of studies provide some credible evidence. This may be useful for small sets of studies that examine a specific question. However, it is not useful to test for the presence of some evidential value in journals or other heterogeneous sets of studies. The conclusion can only be that there is some evidential value or that the test had insufficient power to detect it because the null-hypothesis that all published findings are false is implausible, a priori. P-curves are not useful as a doping test for science because the test was not designed to reveal the presence of bias.

The R-index

The R-index is based on power theory (Cohen, 1988). Statistical power is defined as the long-run probability of obtaining statistically significant results in a series of studies. A study with 50% power is expected to produce 50 significant results and 50 non-significant results. In the short-run, the actual number of significant results can underestimate or overestimate the true power of a study. Importantly, underestimation is as likely as overestimation. However, Sterling (1959) was the first to observe that scientific journals report more significant results than the actual power of studies justifies. In other words, there is a discrepancy between two estimates of

statistical power. A simple count of the percentage of significant results in journals would suggest that psychological studies have over 90% statistical power to reject the null-hypothesis. However, power estimates based on sample sizes and effect sizes in these studies suggest that power is at best 60% (Giegerenzer & Sedelmeier, 1995). The discrepancy between these estimates of power reveals a systematic bias because these estimates should converge in the long run. Discrepancies between the two estimates of power can be tested for significance. A significant difference suggests that publication bias or questionable research practices contributed to the discrepancy. However, even non-significant differences have practical implications because deviations imply that published studies reported effect sizes that were inflated by random sampling error.

Although the logic of the R-index is simple to understand, it has been difficult to implement it because the true power of a set of studies is unknown and statisticians have suggested that it is fruitless to conduct post-hoc power analysis of published studies (Henning & Hoenig, 2001). The argument against power estimation is that post-hoc power estimates depend on observed effect sizes and observed effect sizes are imprecise estimators of true effect sizes. This is a valid argument against post-hoc power estimation in single studies. However, when post-hoc power is estimated based on a set of studies, power estimates become more precise as the number of studies increases.

Ioannidis and Trikalinos (2007) were the first to use power analysis to examine biases in meta-analysis. To estimate power, they first conducted a meta-analysis to obtain the (weighted) average effect size. The weighted average was then used as an estimate of the true effect size in the set of studies. Power of individual studies was then estimated as a function of the estimated true effect size and sample size. This approach works well for meta-analyses of studies that

examine the same effect. However, the approach can lead to false conclusions when the true effect sizes vary across studies (Ioannidis & Trikalinos, 2007). Schimmack (2012) developed a different approach that does not require pooling of effect sizes across studies. In this approach, post-hoc power is first determined for each study as a function of the design, sample size, and the observed effect size of this study. Thus, no assumptions are being made that effect sizes in one study are related to effect sizes in other studies. In a second step, the post-hoc power estimates of individual studies are averaged to estimate the average power of the set of studies. Importantly, no assumption is being made that this average reflects true power because true power can vary across studies. It is merely assumed that the average of the observed studies corresponds to the average of the true power. The average observed power is then used to compare the expected number of significant results against the observed number of significant results. The probability of obtaining at least as many significant results as were actually observed, is called the Incredibility Index (IC-Index). For example, if the average power of a set of 20 studies is 60% (12) and the actual rate of significant results is 90% (18), the IC-Index is .996 or 99.6%. This result indicates that the probability of obtaining more than the reported two non-significant results is over 99%. Thus, it is highly unlikely that the discrepancy between the two estimates of average power (i.e., 60% vs. 90%) is just due to chance. Either non-significant results were excluded due to publication bias or researchers used QRPs to produce more significant results than the true power of their studies would allow. Table 3 applies this method to Bem's summary of his 10 studies, reported in Table 7 (Bem, 2011).

N	d	EP	sig.
100	.25	.705	1
150	.20	.850	1
100	.26	.705	1
100	.23	.705	1
100	.22	.705	1
150	.15	.850	1
150	.14	.850	1
200	.09	.927	0
100	.19	.705	1
50	.42	.456	1

The average estimated power is .748 or 75%. The actual percentage of significant results is 90%. The binomial probability of obtaining fewer than the reported nine significant results in a set of 10 studies with 75% power is 76%. Schimmack (2012) noted that Bem considered the non-significant study a failure due to a methodological flaw rather than a type-II error. If this study is excluded from the analysis, the IC-index increases to 94% (see Schimmack, 2012, for a detailed discussion of Bem's studies).

The IC-index has several problems as a doping test for science. The main problem is that the IC-index tests the probability that the discrepancy between two power estimates is due to chance or not. Like all other statistical tests, the answer to this question is a function of effect size (the magnitude of the discrepancy) and the number of observations. As the number of observations increases, even small discrepancies are unlikely to be due to chance. For example, as noted earlier, the IC-index for 20 studies with 60% power and 18 significant results (discrepancy 90% - 60% = 30%) is 99.6%. The same IC-index is obtained if a set of 100 studies with 90% power contains 98 significant results (discrepancy 98% - 90% = 8%). In this example, the difference in sample sizes masks the fact that the discrepancy between the estimates of power is much larger in the first example (.90 - .60 = .30) than in the second example (.98 - .90 = .08).

In large sets of studies (e.g., an entire volume of a journal), the IC-index is useless because it will merely reveal the well-known presence of publication bias and QRPS. To quantify research integrity it is more useful to focus on the effect size, just like effect sizes are more important than statistical significance in research publications. Thus, I propose the discrepancy between the percentages of significant results minus the estimated average power as an integrity index. As effect sizes are independent of sample sizes, a set of studies with a smaller discrepancy has more integrity than a set of studies with a larger discrepancy.

A second problem of the IC-Index is that it used the mean of estimated power to estimate the average power of a set of studies. This is problematic because sampling error in power estimates is not normally distributed (Yuan & Maxwell, 2005). For example, when the true power is close to the upper value of 100%, observed power is more likely to underestimate than to overestimate true power. To overcome this problem, I suggest using the median as an estimate of average power. The median is unbiased because in each study it is equally likely that the observed effect size underestimates or overestimates the true effect size. Thus, it is equally likely that a power estimate underestimates or overestimates true power. While the amount of underestimation and overestimation is not symmetrically distributed, the direction of bias is known to be equally distributed on both sides of true power. Simulations confirm that the median provides an unbiased estimate of true power even when power is high. In sum, I propose to quantify research integrity with the formula

R-index = Percentage of Significant Results – Median (Estimated Power)

Example 1: Journal of Abnormal and Social Psychology (1960)

Over 50 years ago, Cohen (1962) examined the statistical power in the 1960 volume of the Journal of Abnormal and Social Psychology, which later split into two separate journals; one

of them being the Journal of Personality and Social Psychology that published Bem's (2011) article. Cohen pointed out that the sample sizes in the published studies had only sufficient power to detect large effects. To reach this conclusion, Cohen relied on a priori effect sizes, rather than estimating effect sizes and power of the actual studies (Sedlmeier & Giegerenzer, 1989). Sedlmeier and Giegerenzer selected 20 articles in the 1960 volume and estimated the median effect size, which was $r = .31$. With this estimate of the median effect size, median power is estimated to be just 48% (Sedlmeier & Giegerenzer, 1989). Thus, one would expect that 50% of the studies reported non-significant results. To follow these seminal studies of power, I used the 1960s volume of Journal of Abnormal and Social Psychology to introduce the R-index.

To avoid concerns about selection bias (cheery picking), I used citations as a selection criterion and I focused on the most cited articles. As most articles in these days reported only a single study, I analyzed 18 articles to obtain 20 estimates of power; two articles reported two studies. The logic of using the most highly cited studies is that these studies are most influential. It has also been argued that science is self-correcting and a low citation count might be an indicator of studies with low quality. For science as a process it is important to know the integrity of widely cited articles that continue to have an impact on current science.

One problem for the calculation of the R-index is that a single study contains multiple statistical tests. Some of these tests are not theoretically relevant (e.g., a manipulation check shows a significant effect of a mood manipulation on a mood-measure) and can be ignored. However, often a study also contains multiple tests of theoretical importance. For example, an overall ANOVA is followed up with post-hoc tests or a manipulation should produce significant effects on multiple dependent variables. As these tests are often not independent, it would be

problematic to treat each statistical test as an independent observation. To address this problem I first estimate power for each individual test and then use the median value as an estimate of power for this study. For example, the most cited study by Allport reported 51 statistical tests and 21 statistical tests that I considered to be theoretically important. For each statistical test, I computed the exact p-value and then converted p-values into z-scores. This approach closely follows Rosenthal's use of z-score in meta-analysis. The advantage of z-scores is that it is easy to compute post-hoc power for z-scores (Henning and Hoenig,). For example, the Excel formula is

EP = 1 - Norm.Dist (Norm.INV(p_{1t}, 0, 1) – Norm.INV(p_c, 0,1))

Typically, the critical p-value is .025 because 2.5% of the error rate is in one tail and 2.5% of the error rate is in the other tail. Thus, a significant result requires a one-tailed p-value that is more extreme than 2.5% of p-values in one of the tails. To illustrate this approach, consider a t-test with a strong effect size of d = .8 and n = 20 in each cell. This yields a t-value of 2.53 with 38 degrees of freedom. The two-tailed p-value is .016 and the one-tailed p-value is .008. Using the formula above, power is estimated as 67%.

EP = 1 – Norm.Dist (-2.41 – (-1.96)) = 1 – Norm.Dist (-.45) = 1 - .33 = .67

Using this approach, I first estimated power for each statistical test of a theoretical hypothesis and used the median power as an estimate of the average power in a study. The Top 10 articles reported 20 independent studies. Table 3 shows the median sample sizes, z-scores, p-values, and estimated powers. In addition, it shows the percentage of significant results for each study.

N	p	z	Sig	Power
379	3.78E-05	4.1208	0.952381	0.984646
55	0.000102	3.886829	1	0.973002
45	0.002632	3.831457	1	0.897535
92	0.002032	3.150943	0.875	0.877952
30	0.022258	2.948106	1	0.746485
61	0.003336	2.934938	0.659794	0.835214
171	0.005	2.807034	1	0.801522
77	0.005	2.807034	0.933333	0.801522
30	0.00586	2.755508	1	0.786852
16	0.01	2.575829	0.5625	0.731008
49	0.0125	2.497705	1	0.704622
58	0.016806	2.39498	0.875	0.667956
204	0.02	2.326348	0.666667	0.642961
120	0.02	2.326348	0.818182	0.642961
80	0.024718	2.245775	1	0.612489
40	0.026964	2.21216	1	0.59955
256	0.028708	2.187476	0.8	0.706305
88	0.032	2.177443	0.642857	0.584766
130	0.042625	2.034298	0.5	0.529527
24	0.180036	1.34096	0	0.268005
69	0.014653	2.536767	0.814286	0.719744

Median estimated power was 72%. The average percentage of significant results was 81%. The R-index rewards power and punishes bias due to the use of questionable research practices. It can range from 0 (100% significant results with 50% observed power) to 100, 100% significant results with 100% power. The R-index for the set of articles from the Journal of Abnormal and Social Psychology in 1960 is 72% - (81% - 72%) = 72% - 9% = 63%. The discrepancy between replication rate and median observed power indexes the inflation in effect sizes and observed power. Therefore, the R-index can also be considered a rough approximation

of true median power. An R-index of 63% would suggest that studies had a true power of 60%, which is consistent with other estimates of power in the 1960s issue of Journal of Abnormal and Social Psychology (Cohen, 1962; Sedlmeier & Giegerenzer, 1989).

Example 2: Journal of Personality and Social Psychology: Attitudes and Social Cognition (2011)

The second example focussed on the journal that published Bem's (2011) article. There are several reasons for focussing on this issue. First, it has been proposed that Bem's (2011) article revealed deeper flaws in research practices in psychology. Second, the editors wrote an editorial to justify the publication of this controversial article. One reason was that the article met the scientific criteria of the journal and that rejecting it could not be justified on scientific grounds. Third, I expected that integrity decreased since 1960. The main reason is that publication standards changed. Whereas most articles in 1960 were single-study articles, it was the norm in 2011 to publish multiple-study articles. The requirement to demonstrate evidence for a phenomenon in multiple studies had important implications for power. The reason is that demonstrating a phenomenon repeatedly lowers the type-I error rate (Schimmack, 2012). If the type-I error rate in one study is 5%, the type-I error rate in two studies is only 0.25% ($.5^2$). As power decreases with more stringent criteria to rule out type-I errors, it is only possible to meet the criterion of successful repeated replications by studying stronger effects, increasing sample sizes, or using more questionable research methods. However, neither effect sizes nor sample sizes have changed since 1960 (Sedlmeier & Giegerenzer, 1989). Thus, the most plausible explanation for successful replications in multiple-study articles is an increase in the use of questionable research practices.

The issue contained 16 articles in addition to Bem's (2011) article on extrasensory perception. One article was retracted because the authors noticed problems with the coding of behavioral data. This article reported seven studies, but only six studies tested statistical hypotheses against the null-hypothesis. All statistical tests supported predictions. Median power of individual studies ranged from .56 in Study 7 to .99 in Study 5. The estimated median power of all studies was 96%, suggesting that the statistical analyses are legitimate (@@@ remove later, if you can have control over the data, you do not need to fudge the data analysis @@@). I kept this article in the estimation of the integrity index because excluding a study without statistical bias could create a bias in the integrity index. The 16 articles reported 60 studies; an average of 3.75 studies per article. The median sample size was N = 75 participants; compared to N = 69 in 1960. Median power was 69%. This estimate is also unchanged from the estimate in 1960 (72%). However, the percentage of significant results was much higher in 2011 than in 1960 (99.5% vs. 81%). In fact, only two studies reported non-significant results for a test of theoretically predicted effects, and even these studies averaged over 80% significant results. Accordingly, the percentage of significant results creates the illusion that studies had over 80% power. The discrepancy between the observed power and the percentage of significant results in 2011 is much higher than the percentage in 1960 (31% vs. 11%). This finding confirms the prediction that multiple study articles have ironically produced less credible results because in the absence of a real increase in power, researchers needed to use more QRPs to produce a series of studies with consistent significant results (Schimmack, 2012).

The amount of bias also has implications for the estimation of the true power of studies. As bias leads to inflated effect size estimates, it also leads to inflated estimates of power. When bias is present, true median power is less than the observed median power. This implies that true

power in 1960 was higher than in 2011. Although it is difficult to estimate true power, the R-index rewards high power and punishes bias. The R-index in 1960 is higher than in 2011, indicating that the use of QRPs has decreased the credibility of published results. The R-index in 2011 is 38%, compared to an R-index of 63% in 1960. As the R-index is an approximation of true medina power, this finding also implies that true power has decreased and is below 50%. That is, a study is more likely to fail than to provide evidence for a hypothesis that is actually true.

It is also instructive to compare the observed R-index to indices for different hypothetical scenarios. For example, the extreme scenario in which the null-hypothesis is always true, but only significant results are published would produce a median p-value of .025. This translates into a median observed power of 61%. With 100% success rate, the bias is 39% and the R-index is 22%. In contrast, if researchers would follow Cohen's recommendation and conduct studies with 80% power and publish only significant results, the median p-value would be .005 and median observed power would be

Example 3: Prediction of Empirical Replicability

The Open Science Foundation (OSF) started to conduct empirical replication studies of studies published in 2008 in three top psychology journals (i.e., Psychological Science, Journal of Personality and Social Psychology, & Journal of Experimental Psychology: Learning, Memory, and Cognition). In November 2014, 25 statistical hypothesis tests from 24 articles were reported. I computed the observed power of all 25 tests. I recorded whether a study was replicated. However, replication studies often had larger samples, increasing the chances of a successful replication. I therefore also examined whether the effect size in a replication study would have been significant with the identical sample size as the original study. The different

ways of measuring success in replication studies resulted only in one discrepancy. In this case, the successful replication would not have been successful with the original sample size. As power predicts replicability in studies with identical sample size, I used the adjusted success rate as the criterion. Median observed power in the original studies was 72%, whereas the actual success rate was 100%, resulting in an inflation of significant results 28%. The R-Index for this set of studies is 43%. Importantly, these results are consistent with the analysis of the JPSP:ASC section in Example 2. Thus, the sample of studies in the replication project appears to be representative.

The actual replication rate is 28% and the median power in the replication studies is 12% (5% power implies the null-hypothesis is true and all significant results are type-I errors). Thus, there is no evidence that failed replication studies are statistically biased against successful replications (e.g., median Power = 40% vs. success Rate 10%). These results confirm that observed power in published studies is inflated by doping.

Observed power also predicts success in individual studies, r = .46. The median power of successful replication is .96 (mean = .87). The median power of failed replications is 67%. It is important to remember that median power is 61% when the null-hypothesis is true and failed studies are not reported. Thus, there is little evidence that failed replications are due to problems with the replication attempts. The lack of successful replications is entirely consistent with the explanation that original findings were only significant due to the use of QRPs that inflate effect sizes that cannot be replicated in unbiased replication attempts.

It is also interesting to examine the influence of research design on replication success. Within-subject (WS) designs can have high power with small sample sizes, whereas between-subject (BS) designs often need large samples to achieve power. The set of 25 studies included

11 BS studies. None of the BS studies was successfully replicated. Thus finding should not be interpreted as evidence that the null-hypothesis in all of these studies is true. The main problem is that the original studies did not have sufficient power to produce significant results when the true effect size is small. Given sample sizes ranging from N = 30 to 236 (median = 59), the median power of BS studies to detect a small (d = .2) effect was 12%. Thus, only one of the 11 BS studies should have produced significant results with a true effect size of d = .2 in all studies.

In conclusion, the replication project has provided valuable information about the empirical replicability of psychological studies. I used information to validate the R-Index. An R-Index of 43% suggests that no more than 50% of original studies can be replicated in a set of exact replications with the same statistical power as the original studies. The actual replication rate of 28% is consistent with this prediction. Moreover, statistical power and bias is a predictor of replication success of individual studies. The R-Index of successful replications was 92% (96% - 4% = 92%). The R-Index of failed replications was 34% (67% - 33% = 34%). The R-Index is conservative and less precise when true power is low. More research is needed to optimize prediction of empirical replicability, but the results provide a simple explanation for empirical replication failures: reported results in original articles are inflated by questionable research practices. This is not an isolated or minor problem. In fact, the majority of studies, especially BS studies with small samples have low replicability.

Example 4: The Multiple Lab Project

In the fourth example, I computed the R-Index for the Mani-Labs project (@@@). In this project, an international team of researchers replicated 13 psychological studies in several laboratories. The main finding of this project was that 10 of the 13 studies were successfully replicated in several labs; a replication rate of 77%. The success rate is notably higher than the

success rate in the ongoing replication project (i.e., 28%). One possible explanation for this discrepancy is that many-labs replication studies had higher power. To test this hypothesis, I computed the observed power of 12 studies. One of the original studies provided insufficient information about effect size. Table @@@ shows the results. Median observed power was 86%. As all original studies produced significant results (for one study, significance was marginal, p = .10), bias is 14%. If the marginal result is counted as a non-significant result, bias is only 6%. The R-Index is 72% or 80%, if the marginal result is counted as non-significant. Thus, the R-Index correctly predicts that studies in the many-labs project were more likely to replicate than studies in the open science project. The reason is simply that the studies in the many-labs project had higher power. A comparison with the analysis of JPSP:ASC shows that the many-labs studies are not representative of studies in contemporary social psychology (R-Index of 38% vs. 72%). In fact, many of these studies are based on classic findings that have been replicated many times before the replication attempt in the many-labs project. The median publication year of studies that successfully replicated is 1995 (range 1941 to 2009). In contrast, the median publication year of the three failed replication studies is 2011 (range 2010 to 2013). Observed power of original studies also predicted replicability. The median observed power of studies that were successfully replicated was 97%. In contrast, the median observed power of failed replication studies was 51%. The correlation between observed power and replication success was r = .71. The many labs also reported the percentage of labs that successfully replicated the original study. Observed power in the original studies predicted replication rate with *r* = .79. The largest discrepancy was observed for the Norms of Reciprocity study (Observed Power = 92%, Replication Rate = 36%). A simple explanation for this discrepancy is that the sample size in the original study was much larger than the typical sample size in

replication attempts. In conclusion, the R-Index provides valuable information about the outcome of the many-labs project. First, it shows that the set of studies is not representative of contemporary social psychology. Second, it correctly predicts a high rate of successful replications that is close to the actual success rate (72% vs. 75). Third, it shows again that studies with high observed power (> 80%) are likely to replicate, whereas studies with low observed power (< 80%) have a high failure rate.

Conclusion

It has been widely recognized that questionable research practice are threatening the foundations of science. This manuscript introduces the R-Index as a statistical tool to assess the replicability of published results. Results are replicable if the original studies had sufficient power to produce significant results. A study with 80% power is likely to produce a significant result in 20% of all attempts without the need for questionable research practices. In contrast, a study with 20% power can only produce significant results with the help of inflated effect sizes. In 20% of all attempts, luck alone will be sufficient to inflate effect sizes. In all other cases, researchers have to hide failed attempts in file drawers or use questionable statistical practices to inflate effect sizes. The R-Index reveals the presence of questionable research practices when observed power is lower than the rate of significant results. The R-Index has two components. It increases with observed power because studies with high power are more likely to replicate. The second component is the discrepancy between the percentage of significant results and observed power. The greater the discrepancy, the more questionable research practices have contributed to success and the more observed power overestimates true power.

References

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. [Article]. *Journal of Personality and Social Psychology, 71*(2), 230-244. doi: 10.1037//0022-3514.71.2.230

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. [Article]. *Journal of Personality and Social Psychology, 100*(3), 407-425. doi: 10.1037/a0021524

Bem, D. J., & Honorton, C. (1994). DOES PSI EXIST - REPLICABLE EVIDENCE FOR AN ANOMALOUS PROCESS OF INFORMATION-TRANSFER. [Review]. *Psychological Bulletin, 115*(1), 4-18. doi: 10.1037//0033-2909.115.1.4

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71*(3), 425-440.

Cohen, J. (1990). Things I have learned (so far). [Article]. *American Psychologist, 45*(12), 1304-1312.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159.

Cohen, J. (1994). The earth Is round (P-Less-Than.05). *American Psychologist, 49*(12), 997-1003.

Egger, M., & Smith, G. D. (1998). Meta-analysis - Bias in location and selection of studies. [Article]. *British Medical Journal, 316*(7124), 61-66.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. [Article]. *British Medical Journal, 315*(7109), 629-634.

Francis, G. (2012b). Too good to be true: Publication bias in two prominent studies from experimental psychology. [Article]. *Psychonomic Bulletin & Review, 19*(2), 151-156. doi: 10.3758/s13423-012-0227-9

Galak, J., LeBoeuf, R. A., Nelson, L. D., & Simmons, J. P. (2013). Correcting the Past: Failures to Replicate Psi. *Journal of Personality and Social Psychology*.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*(6), 1464-1480.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. [Article]. *American Statistician, 55*(1), 19-24. doi: 10.1198/000313001300339897

Ioannidis, J. P. A., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. [Article]. *Clinical Trials, 4*(3), 245-253. doi: 10.1177/1740774507079441

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. [Article]. *Psychological Science, 23*(5), 524-532. doi: 10.1177/0956797611430953

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. [Article]. *Group Dynamics-Theory Research and Practice, 6*(1), 101-115.

Rosenthal, R. (1979). The File Drawer Problem and Tolerance for Null Results. [Article]. *Psychological Bulletin, 86*(3), 638-641.

Schimmack, U. (2012). The Ironic Effect of Significant Results on the Credibility of Multiple-Study Articles. [Article]. *Psychological Methods, 17*(4), 551-566. doi: 10.1037/a0029487

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies. [Article]. *Psychological Bulletin, 105*(2), 309-316.

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.

Simonsohn, U., Nelson, L. D., & Simmons, J. (2013). P-Curve: A Key to the File Drawer. *Journal of Experimental Psychology: General*.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance: Or vice versa. [Article]. *Journal of the American Statistical Association, 54*(285), 30-34. doi: 10.2307/2282137

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice-versa. [Editorial Material]. *American Statistician, 49*(1), 108-112.