Running head: INCREDIBILITY INDEX

The Ironic Effect of Significant Results on the Credibility of Multiple Study Articles

Ulrich Schimmack

University of Toronto Mississauga

June 2012

*Author's Not*e. I would like to thank Daryl Bem, Roy Baumeister, Brian Connelly, Alex Kogan, Leyla Javam, Bonnie Le, Raymond Mar, Amy Muise, Scott Maxwell, Elizabeth Page-Gould, and Jonathan Schooler for helpful comments during the writing of this article. This work was made possible by a standard research grant of the Canadian Social Sciences and Humanities Research Council (SSHRC). Correspondence should be sent to Ulrich Schimmack, Department of Psychology, University of Toronto Mississauga, email: uli.schimmack@utoronto.ca.

Abstract

Cohen (1962) pointed out the importance of statistical power for psychology as a science, but statistical power has not improved. At the same time, the number of studies has increased from a single study to multiple studies within a single article. It has been overlooked that multiple study articles are severely underpowered because power decreases as a function of the number of statistical tests that are being conducted (Maxwell, 2004). The discrepancy between the expected number of significant results and the actual number of significant results in multiple study articles undermine the credibility of the reported results and it is likely that questionable research practices contributed to the reporting of too many significant results (Sterling, 1959). The problem of low power in multiple study articles is illustrated using Bem's (2011) article on extrasensory perception and Gailliot et al.'s (2007) article on glucose and self-regulation. I conclude with several recommendations that can increase the credibility of scientific evidence in psychological journals. One major recommendation is to pay more attention to the power of studies to produce positive results without the help of questionable research practices and to request that authors justify sample sizes with *a priori* predictions of effect sizes. It is also important to publish replication studies with non-significant results, if these studies had high power to replicate a published finding.

Keywords: Power, Publication Bias, Significance, Credibility, Sample Size

The Ironic Effect of Significant Results on the Credibility of Multiple Study Articles

"Less is more except of course for sample size" (Cohen, 1990)

In 2011, the prestigious *Journal of Personality and Social Psychology* published an article that provided empirical support for extrasensory perception (Bem, 2011). The publication of this controversial article created vigorous debates in psychology departments, the media, and science blogs. In response to this debate, the acting editor and the editor-in-chief felt compelled to write an editorial accompanying the article. The editors defended their decision to publish the article by noting that Bem's studies were performed according to standard scientific practices in the field of experimental psychology and that it would seem inappropriate to apply a different standard to studies of extrasensory perception (Judd & Gawronski, 2011).

Others took a less sanguine view. They saw the publication of Bem's (2011) article as a sign that the scientific standards guiding publication decisions are flawed and that Bem's article serves as a glaring example of these flaws (Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). In a nutshell, Wagenmakers et al. (2011) argued that the standard statistical model in psychology is biased against the null-hypothesis; that is, only findings that are statistically significant are submitted and accepted for publication. This bias leads to the publication of too many positive (i.e., statistically significant) results.

The observation that scientific journals, not only in psychology, publish too many statistically significant results is by no means novel. In a seminal article, Sterling (1959) noted that selective reporting of statistically significant results can produce literatures that "consist in

substantial part of false conclusions" (p. 30). Three decades later, Sterling, Rosenbaum, and Weinkam (1995) observed that "the practice leading to publication bias have not changed over a period of 30 years" (p. 108). Recent articles indicate that publication bias remains a problem in psychological journals (Fiedler, 2011; Kerr, 1998; John, Loewenstein, & Prelec, 2012; Simmons, Nelson, and Simonsohn, 2011; Strube, 2006; Vul, Harris, Winkielman, & Pashler, 2009; Yarkoni, 2010). Other sciences have the same problem (Yong, 2012). For example, medical journals have seen an increase in the percentage of retracted articles (Steen, 2011a, 2011b), and there is the concern that a vast number of published findings may be false (Ioannidis, 2005). However, a recent comparison of different scientific disciplines suggests that the bias is stronger in psychology than in some of the older and harder scientific disciplines at the top of a hierarchy of sciences (Fanelli, 2010).

It is important that psychologists use the current crisis as an opportunity to fix problems in the way research is being conducted and reported. The proliferation of eye-catching claims based on biased or fake data can have severe negative consequences for a science. A recent *New Yorker* article (Lehrer, 2010) warned the public that "all sorts of well-established, multiply confirmed findings have started to look increasingly uncertain. It's as if our facts were losing their truth: claims that have been enshrined in textbooks are suddenly unprovable." If students who read psychology textbooks and the general public lose trust in the credibility of psychological science, psychology loses its relevance because objective empirical data are the only feature that distinguishes psychological science from other approaches to the understanding of human nature and behavior. It is therefore hard to exaggerate the seriousness of doubts about the credibility of research findings published in psychological journals.

In an influential article, Kerr (1998) discussed one source of bias, namely hypothesizing after the results are known (HARKing). The practice of HARKing may be attributed to the high costs of conducting a study that produces a non-significant result that cannot be published. To avoid this negative outcome, researchers can design more complex studies that test multiple hypotheses. Chances increase that at least one of the hypotheses will be supported, if only because type-I error increases (Maxwell, 2004). As noted by Wagenmakers et al. (2011), generations of graduate students were explicitly advised that this questionable research practice is how they should write scientific manuscripts (Bem, 2000).

It is possible that Kerr's article undermined the credibility of single study articles and added to the appeal of multiple study articles (Diener, 1998; Ledgerwood & Sherman, 2012). After all, it is difficult to generate predictions for significant effects that are inconsistent across studies. Another advantage is that the requirement of multiple significant results essentially lowers the chances of a type-I error; that is, the probability of falsely rejecting the null-hypothesis. For a set of five independent studies, the requirement to demonstrate five significant replications essentially shifts the probability of a type-I error from $p < .05$ for a single study to $p < .0000003$ (i.e., $.05^5$) for a set of five studies. This is approximately the same stringent criterion that is being used in particle physics to claim a true discovery (Castelvecchi, 2011).

It has been overlooked, however, that researchers have to pay a price to meet more stringent criteria of credibility. To demonstrate significance at a more stringent criterion of significance it is necessary to increase sample sizes to reduce the probability of making a type-II error (failing to reject the null-hypothesis). This probability is called *beta*. The inverse probability (1 – *beta*) is called *power*. Thus, to maintain high statistical power to demonstrate an effect with a more stringent alpha-level requires an increase in sample sizes, just like physicists

had to build a bigger collider to have a chance to find evidence for smaller particles like the Higgs-Bosson particle. Yet, there is no evidence that psychologists are using bigger samples to meet more stringent demands of replicability (Cohen, 1992; Maxwell, 2004; Rossi, 1990; Sedlmeier & Giegerenzer, 1989). This raises the question of how researchers are able to replicate findings in multiple study articles, despite modest power to demonstrate significant effects even within a single study. Researchers can use questionable research practices (e.g., snooping, not reporting failed studies, dropping dependent variables, etc.; Strube, 2006; Simons et al., 2011) to dramatically increase the chances of presenting a false-positive result. Moreover, a survey of researchers indicated that these practices are common (John et al., 2012), and the prevalence of these practices has raised concerns about the credibility of psychology as a science (Yong, 2012).

An implicit assumption in the field appears to be that the solution to these problems is to further increase the number of positive replication studies that need to be presented to ensure scientific credibility (Ledgerwood & Sherman, 2012). However, the assumption that many replications with significant results provide strong evidence for a hypothesis is an illusion that is akin to the Texas sharpshooter fallacy (Milloy, 1995). Imagine a Texan farmer named Joe. One day he invites you to his farm and shows you a target with nine shots in the bull's eye and one shot just outside the bull's eye. You are impressed by his shooting abilities until you find out that he cannot repeat this performance when you challenge him to do it again. Over some beers, Joe tells you that he first fired 10 shots at the barn and then drew the targets after the shots were fired. One problem in science is that reading a research article is a bit like visiting Joe's farm. Readers only see the final result, without knowing how the final results were created. Is Joe a sharpshooter who drew a target and then fired 10 shots at the target? Or was the target drawn after the fact?

The reason why multiple study articles are akin to a Texan sharpshooter is that psychological studies have modest power (Cohen, 1962; Rossi, 1990; Sedlmeier & Giegerenzer, 1989). Assuming 60% power for a single study, the probability to obtain 10 significant results in 10 studies is less than 1% ($.6^{10} = .006 = 0.6\%$). I call the probability to obtain only significant results in a set of studies *total power*. Total power parallels Maxwell's (2004) concept of *all-pair power* for multiple comparisons in Analysis of Variance designs. Figure 1 illustrates how total power decreases with the number of studies that are being conducted. Eventually, it becomes extremely unlikely that a set of studies produces only significant results. This is especially true if a single study has modest power. When total power is low, it is incredible that a set of studies yielded only significant results. To avoid the problem of incredible results, researchers would have to increase the power of studies in multiple study articles.

Table 1 shows how the power of individual studies has to be adjusted to maintain 80% total power for a set of studies. For example, to have 80% total power for five replications, the power of each study has to increase to 96%. Table 1 also provides information about the total sample size across all studies that is required to achieve 80% total power, assuming a simple between-group design, an alpha-level of .05 (two-tailed), and Cohen's (1992) guidelines for a small ($d = .2$), moderate, ($d = .5$), and strong ($d = .8$) effect.

In sum, my main proposition is that psychologists have falsely assumed that increasing the number of replications within an article increases credibility of psychological science. The problem of this practice is that a truly programmatic set of multiple studies is very costly and few researchers are able to conduct multiple studies with adequate power to achieve significant results in all replication attempts. Thus, multiple study articles have intensified the pressure to

use questionable research methods to compensate for low total power and may have weakened rather than strengthened the credibility of psychological science.

What is the Allure of Multiple Study Articles?

One apparent advantage of multiple study articles is to provide stronger evidence against the null-hypothesis (Ledgerwood & Sherman, 2012). However, the number of studies is irrelevant because the strength of the empirical evidence is a function of the total sample size rather than the number of studies. The main reason why aggregation across studies reduces randomness as a possible explanation for observed mean differences (or correlations) is that *p*-values decrease with increasing sample size. The number of studies is mostly irrelevant. A study with 1,000 participants has as much power to reject the null-hypothesis as a meta-analysis of 10 studies with 100 participants if it is reasonable to assume a common effect size for the 10 studies. If true effect sizes very across studies, power decreases because a random effect model may be more appropriate (Schmidt, 2010; but see Bonett, 2009). Moreover, the most logical approach to reduce concerns about type-I error is to use more stringent criteria for significance (Mudge, Baker, Edge, & Houlahan, 2012). For controversial or very important research findings, the significance level could be set to $p < .001$ or, as in particle physics, to $p < .0000005$. It is therefore misleading to suggest that multiple study articles are more credible than single study articles. A brief report with a large sample ($N = 1{,}000$) provides more credible evidence than a multiple study article with five small studies ($N = 40$, total $N = 200$).

The main appeal of multiple study articles seems to be that they can address other concerns (Ledgerwood & Sherman, 2012). For example, one advantage of multiple studies could be to test the results across samples from diverse populations (Henrich, Heine, & Norenzayan, 2010). However, many multiple-study articles are based on samples drawn from a narrowly

defined population (typically students at the local university). If researchers were concerned about generalizability across a wider range of individuals, multiple study articles should examine different populations. However, it is not clear why it would be advantageous to conduct multiple independent studies with different populations. To compare populations it would be preferable to use the same procedures and to analyze the data within a single statistical model with population as a potential moderating factor. Moreover, moderator tests often have low power. Thus, a single study with a large sample and moderator variables is more informative than articles that report separate analyses with small samples drawn from different populations.

Another attraction of multiple study articles appears to be the ability to provide strong evidence for a hypothesis by means of slightly different procedures. However, even here single studies can be as good as multiple study articles. For example, replication across different dependent variables in different studies may mask the fact that studies included multiple dependent variables and researchers picked dependent variables that produced significant results (Simmons et al., 2011). In this case, it seems preferable to demonstrate generalizability across dependent variables by including multiple dependent variables within a single study and reporting the results for all dependent variables. One advantage of a multi-method assessment in a single study is that the power to demonstrate an effect increases for two reasons. First, while some dependent variables may produce non-significant results in separate small studies due to low power (Maxwell, 2004), they may all show significant effects in a single study with the total sample size of the smaller studies. Second, it is possible to increase power further by constraining coefficients for each dependent variable or by using a latent variable measurement model to test whether the effect is significant across dependent variables rather than for each one independently.

Multiple study articles are most common in experimental psychology to demonstrate the robustness of a phenomenon using slightly different experimental manipulations. For example, Bem (2011) used a variety of paradigms to examine extrasensory perception. Demonstrating a phenomenon in several different ways can show that a finding is not limited to very specific experimental conditions. Analogously, if Joe can hit the bull's eye nine times from different angles, with different guns, and in different light conditions, Joe truly must be a sharpshooter. However, the variation of experimental procedures also introduces more opportunities for biases (Ioannidis, 2005). The reason is that variation of experimental procedures allows researchers to discount null-findings. Namely, it is possible to attribute non-significant results to problems with the experimental procedure rather than to the absence of an effect. In this way, empirical studies no longer test theoretical hypotheses because they can only produce two results: either they support the theory ($p < .05$) or the manipulation did not work ($p > .05$). It is therefore worrisome that Bem (2011) noted that "like most social-psychological experiments, the experiments reported here required extensive pilot testing" (p. 421). If Joe is a sharpshooter, who can hit the bull's eye from different angles and with different guns, why does he need extensive training before he can perform the critical shot?

The freedom of researchers to discount null-findings leads to the paradox that conceptual replications across multiple studies give the impression that an effect is robust followed by warnings that experimental findings may not replicate because they depend "on subtle and unknown factors" (Bem, 2011, p. 422). If experimental results were highly context dependent, it would be difficult to explain how studies reported in research articles nearly always produce the expected results. One possible explanation for this paradox is that sampling error in small samples creates the illusion that effect sizes vary systematically, although most of the variation is

random. Researchers then pick studies that randomly produced inflated effect sizes and may further inflate them by using questionable research methods to achieve significance (Simmons et al., 2011). The final set of studies that worked is then published and gives a false sense of the effect size and replicability of the effect (you should see the other side of Joe's barn). This may explain why research findings initially seem so impressive, but when other researchers try to build on these seemingly robust findings, it becomes increasingly uncertain whether a phenomenon exists at all (Ioannidis, 2005; Lehrer, 2010). At this point, a lot of resources have been wasted without providing credible evidence for an effect.

To increase the credibility of reported findings, it would be better to use all of the resources for one powerful study. For example, the main dependent variable in Bem's (2011) study of extrasensory perception (ESP) was the percentage of correct predictions of future events. Rather than testing this ability 10 times with $N = 100$ participants, it would have been possible to test the main effect of ESP in a single study with 10 variations of experimental procedures and use the experimental conditions as a moderating factor. By testing one main effect of ESP in a single study with $N = 1{,}000$, power would be greater than 99.9% to demonstrate an effect with Bem's *a priori* effect size. At the same time, the power to demonstrate significant moderating effects would be much lower. Thus, the study would lead to the conclusion that ESP does exist, but that it is unclear whether the effect size varies as a function of the actual experimental paradigm. This question could then be examined in follow up studies with more powerful tests of moderating factors.

In conclusion, it is true that a programmatic set of studies is superior to a brief article that reports a single study if both articles have the same total power to produce significant results (Ledgerwood & Sherman, 2012). However, once researchers use questionable research practices

to make up for insufficient total power, multiple study articles lose their main advantage over single study articles, namely to demonstrate generalizability across different experimental manipulations or other extraneous factors. Moreover, the demand for multiple studies counteracts the demand for more powerful studies (Cohen, 1962; Maxwell, 2004; Rossi, 1990) because limited resources (e.g., subject pool of PSY100 students) can only be used to increase sample size in one study or to conduct more studies with small samples. It is therefore likely that the demand for multiple studies within a single article has eroded rather than strengthened the credibility of published research findings (Steen, 2011a, 2011b), and it is problematic to suggest that multiple study articles solve the problem that journals publish too many positive results (Ledgerwood & Sherman, 2012). Ironically, the reverse may be true because multiple study articles provide a false sense of credibility.

Joe the Magician: How Many Significant Results are Too Many?

Most people enjoy a good magic show. It is fascinating to see something and to know at the same time that it cannot be real. Imagine that Joe is a well-known magician. In front of a large audience he fires nine shots from impossible angles, blindfolded, and seemingly through the body of an assistant, who miraculously does not bleed. You cannot figure out how Joe pulled of the stunt, but you know it was a stunt. Similarly, seeing Joe hit the bull's eye 1000 times in a row raises concerns about his abilities as a sharp shooter and suggests that some magic is contributing to this miraculous performance. Magic is fun, but it is not science. The problem is that some articles in psychological journals appear to be more magical than one would expect on the basis of the normative model of science (Kerr, 1998).

To increase the credibility of published results, it would be desirable to have a diagnostic tool that can distinguish between credible research findings and those that are likely to be based

on questionable research practices. Such a tool would also help to counteract the illusion that multiple study articles are superior to single study articles, without leading to the erroneous reverse conclusion that single study articles are more trustworthy. Articles should be evaluated on the basis of their total power to demonstrate consistent evidence for an effect. As such a single-study article with 80% (total) power is superior to a multiple study article with 20% total power, but a multiple study article with 80% total power is superior to a single study article with 80% power.

The Incredibility Index

The idea to use power analysis to examine bias in favor of theoretically predicted effects and against the null-hypothesis was introduced by Sterling et al. (1995). Ioannidis and Trikalinos (2007) provided a more detailed discussion of this approach for the detection of bias in meta-analyses. Ioannidis and Trikalinos (2007) exploratory test estimates the probability of the number of reported significant results given the average power of the reported studies. Low *p*-values suggest that there are too many significant results, suggesting that questionable research methods contributed to the reported results. In contrast, the inverse inference is not justified because high *p*-values do not justify the inference that questionable research practices did not contribute to the results. For example, HARKing can contribute to significant results in single study articles with high power (Kerr, 1998). As such, low *p*-values suggest that reported results are incredible, but high values do not suggest that results are credible. In other words, a high *p*-value is necessary to ensure credibility, but it is not sufficient. To emphasize this asymmetry in inferential strength, I suggest to reverse the exploratory test and to focus on the probability of obtaining more non-significant results than were reported in a multiple study article and to call this index the *Incredibility Index* (IC-index). Higher values of the IC-index indicate that there is a

surprising lack of non-significant results (a.k.a., shots that missed the bull's eye). The higher the IC-index is, the more incredible the observed outcome becomes. The term incredible conveys the meaning that a set of results is unbelievable or improbable, just as it is improbable to hit a bull's eye 10 times in a row. Incredible results do not imply a specific reason for the incredible event. Too many significant results could be due to faking, fudging, or fortune. Thus, the statistical demonstration that a set of reported findings is incredible does not prove that questionable research methods contributed to the results in a multiple study article. However, even when questionable research methods did not contribute to the results, the published results are still likely to be biased because fortune helped to inflate effect sizes and produce more significant results than total power justifies.

Ioannidis and Trikalinos (2007) suggested a criterion of $p < .10$, which is equivalent to an IC-index > .90, as a criterion to infer that the results of a meta-analysis are biased. They justified this relatively liberal criterion by means of low power of the exploratory test when the set of studies in a meta-analysis is small. This liberal criterion may be reasonable for meta-analysis where there are multiple reasons for bias. However, when the method is used to evaluate multiple study articles, evidence for bias implies that researchers used questionable research methods. In this context, a 90% criterion may be too liberal.

I think that it is counterproductive to use a strict criterion to make dichotomous decisions about bias in single articles. The problem of this approach is apparent in Francis's (2012a) examination of bias in an article with five studies by Balcetis and Dunning (2010). Francis (2012a) calculated that the probability of the reported significant effects in all five studies was $p = .076$ (IC-index = .924). Using a criterion of $p < .10$, he concluded that "the proper interpretation of the experimental findings is that they are non-scientific or anecdotal" (p. 176).

In response, Balcetis and Dunning (2012) used a different approach to compute the probability of bias (the differences are elaborated later), and obtained a p-value of .163 (IC-index = .837). As this value does not meet the criterion of $p < .10$, the authors concluded that Francis's conclusions are unwarranted.

The focus on the $p < .10$ criterion (IC-index > .90) distracts from the main problem of multiple study articles that total power is low, no matter whether the point estimate is 8% or 16% (when all results are significant, the IC-index is the inverse of total power). It seems unwise to embark on a research project that has a less than 20% chance to produce the desired outcome that all five studies produce a significant result.

In sum, I propose to use the probability that a set of studies yielded more non-significant results than were reported to evaluate the credibility of reported results in a study. I call this probability the incredibility index because it becomes increasingly incredible that all repeated attempts are successful. High IC-values raise concerns about the credibility of published research findings, but I do not recommend a fixed value that can be used to make dichotomous decisions about bias. In addition, I recommend to compute total power and to make high total power a requirement for multiple study articles because low total power undermines the credibility of successful replications.

Computation of the Incredibility Index

To understand the basic logic of the IC-index, it is helpful to consider a concrete example. Imagine a multiple study article with 10 studies with an average observed effect size of $d = .5$ and 84 participants in each study (42 in two conditions, total $N = 840$) and all studies produced a significant result. At first sight these 10 studies seem to provide strong support against the null-hypothesis. However, a *post-hoc* power analysis with the average effect size of *d*

= .5 as estimate of the true effect size reveals that each study had only 60% power to obtain a significant result. That is, even if the true effect size were *d* = .5, only 6 out of 10 studies should have produced a significant result. The IC-index quantifies the probability of this outcome using binomial distribution theory.

From the perspective of binomial probability theory, the scenario is analogous to an urn problem with replacement with six green balls (significant) and four red balls (non-significant). The binomial probability to draw at least 1 red ball in 10 independent draws is 99.4%. (http://stattrek.com/tables/binomial.aspx). That is, 994 out of 1000 multiple study articles with 10 studies and 60% average power should have produced at least one non-significant result in one of the 10 studies. It is therefore incredible if an article reports 10 significant results because only 6 out of 1000 attempts would have produced this outcome simply due to chance alone.

One of the main problems for power analysis in general and the computation of the IC-index in particular is that the true effect size is unknown and has to be estimated. There are three basic approaches to the estimation of true effect sizes. In rare cases, researchers provide explicit *a priori* assumptions about effect sizes (Bem, 2011). In this situation, it seems most appropriate to use an author's stated assumptions about effect sizes to compute *post-hoc* power analyses with the sample sizes of each study. A second approach is to average reported effect sizes either by simply computing the mean value or by weighting effect sizes by their sample sizes. Averaging of effect sizes has the advantage that *post-hoc* effect size estimates of single studies tend to have large confidence intervals. The confidence intervals shrink when effect sizes are aggregated across studies. However, this approach has two drawbacks. First, averaging of effect sizes makes strong assumptions about the sampling of studies and the distribution of effect sizes (Bonett, 2009). Second, this approach assumes that all studies have the same effect size, which is unlikely

if a set of studies used different manipulations and dependent variables to demonstrate the generalizability of an effect. Ioannidis and Trikalinos (2007) were careful to warn readers that "genuine heterogeneity may be mistaken for bias" (p. 252).

To avoid the problems of average effect sizes, it is promising to consider a third option. Rather than pooling effect sizes, it is possible to conduct post-hoc power analysis for each study. Although each *post-hoc* power estimate is associated with considerable sampling error, sampling errors tend to cancel each other out and the IC-index for a set of studies becomes more accurate without having to assume equal effect sizes in all studies. Unfortunately, this does not guarantee that the IC-index is unbiased because power is a non-linear function of effect sizes. Yuan and Maxwell (2005) examined the implications of this non-linear relationship. They found that the IC-index is conservative (i.e., underestimates incredibility) when the true effect sizes are in the small to moderate range (d < .5). A meta-analysis of social psychological experiments produced an average effect size of *d* = .4, which is in the small (*d* = .2) to moderate (*d* = .5) range (Richard et al. 2003). Thus, the IC-index seems appropriate to evaluate the credibility of multiple study articles for articles with typical effect sizes. When power is moderate (*d* = .5), the IC-index is unbiased. However, for strong effect sizes, the IC-index may provide inflated estimates of incredibility, especially in small samples where observed effect sizes vary widely around the true effect size.

Overestimation of incredibility is unlikely, however, because Yuan and Maxwell's (2005) simulation did not take publication bias into account. The main reason for an inflated IC-index for strong effects in small samples is that power will be underestimated more when observed effect sizes are lower than the true effect size than the overestimation of power when observed effect sizes are inflated. For example, a researcher may expect a strong true effect (*d* = .8) and

conduct a between-subject study with $N = 20$ in each condition, which is a common sample size for small studies in psychology (Simmons et al., 2011). Due to the strong effect size, the study has 70% power. However, in actual studies, the observed effect size will differ due to sampling error. If the observed effect size is $d = .6$, estimated power decreases to 46%, a decrease by 24% points. In contrast, an equivalent inflation of the true effect size by .2 points ($d = 1$) raises the power estimate to 87%, an inflation by only 17%. The difference between the amount of underestimation (24%) and overestimation (17%) leads to the inflation of the IC-index. However, a power-value of 46% implies that the observed effect is not significant at the conventional $p = .05$ significance level because 50% power corresponds to a p-value of .05 (Hoenig & Heisey, 2001). If the reported results do not include non-significant results it is likely that the reported effect sizes are inflated because even strong effects are likely to produce non-significant results in small samples.

Another reason why the IC-index is likely to be conservative is that Yuan and Maxwell (2005) assumed that effect sizes are not systematically related to sample sizes. However, multiple study articles often show negative correlations between sample size and effect sizes. For example, in Bem's (2011) article on ESP, the correlation between effect size and sample size is $r = -.91$ (Alcock, 2011). There are two reasons for negative correlations between effect sizes and sample sizes. One reason is that researchers make *a priori* predictions about heterogeneous effect sizes in different studies and use power analysis to plan sample sizes to have sufficient power to obtain significant effects. However, it is rare to find any explicit rationale for varying sample sizes in multiple study articles. For example, Bem (2011) explicitly stated that he "set 100 as the minimum number of participants/sessions for each of the experiments reported in this article" (p. 409). He provided no rationale for the fact that the actual sample sizes ranged from $N = 50$ to $N =$

200 and he did not explain why this variation in effect sizes correlates strongly with the observed effect sizes. For example, studies 8 and 9 are classified together as retroactive facilitation of recall and differ only in a minute detail. It is therefore not clear why one would expect *a priori* a stronger effect in Study 9 that would justify reducing the sample size from $N = 100$ to $N = 50$, which implies a reduction of *a priori* power from 80% to 54%.

The second reason for negative correlations between sample sizes and effect sizes is that researchers use questionable research methods to achieve a significant result. A negative correlation between effect sizes and sample sizes is a common tool in meta-analysis to detect bias (Sutton & Higgins, 2008) because smaller samples have less power. Thus, effect sizes in small samples have to be inflated more to achieve a significant result. If bias produced an artificial correlation between effect sizes and sample sizes, the IC-index is conservative because *post-hoc* power of large effects in small samples is overestimated. To illustrate this, it is instructive to use Bem's (2011) Study 9 with 50 participants that yielded an observed effect size of .42. This is nearly double the average effect size of all studies, his *a priori* effect size of $d = .25$, and the effect size in the nearly identical Study 8 ($d = .19$). By using the observed effect size in Study 9, the *post-hoc* power estimate increases to 90%, whereas power for the other estimation procedures ranges from 37% to 54%. A value of 90% power would have a small effect on the IC-index and the IC-index might underestimate incredibility, if large effects in small samples are due to bias.

Finally, the IC-index is conservative because the binomial distribution formula assumes that each study has the same power. This assumption is rarely fulfilled because effect sizes and sample sizes often vary across studies. The simplifying assumption of equal power in all studies makes the IC-index conservative because this special case maximizes the expected value of

significant results. In contrast, variation in sample sizes makes it more likely that smaller studies have less power and are more likely to produce a non-significant result.

In sum, it is possible to use reported effect sizes to compute *post-hoc* power and to use *post-hoc* power estimates to determine the probability of obtaining a significant result. The *post-hoc* power values can be averaged and used as the probability for a successful outcome. It is then possible to use binomial probability theory to determine the probability that a set of studies would have produced equal or more non-significant results than were actually reported. This probability is called the IC-index.

Example 1: Extrasensory Perception (Bem, 2011)

There are several reasons to choose Bem's (2011) article as an example. First, Bem's (2011) article may have been a tipping point for the current scientific paradigm in psychology (Wagenmakers et al., 2011). Second, editors explicitly justified the publication of Bem's (2011) article on the grounds that it was subjected to a rigorous review process, suggesting that it meets current standards of scientific practices (Judd & Gawronski, 2011). In addition, the editors hoped that the publication of Bem's (2011) article and Wagenmakers et al.'s (2011) critique would stimulate "critical further thoughts about appropriate methods in research on social cognition and attitudes" (p. 406). Finally, Francis (2012b) used a high IC-index to conclude that Bem's (2011) article does "not tell us anything scientifically useful" (p. 154).

A first step in the computation of the IC-index is to define the set of effects that are being examined. This may seem trivial, when the IC-index is used to evaluate the credibility of results in a single article, but multiple study articles contain many results and it is not always obvious that all results should be included in the analysis (Maxwell, 2004). For example, Bem's (2011) Study 1 contained five types of stimuli, namely erotic pictures, romantic but non-erotic pictures,

positive pictures, negative pictures, and neutral pictures. Only erotic pictures produced a significant result. To compute the IC-index, it is necessary to decide whether the non-significant results for the other four types of stimuli should be treated as failed replications or whether these effects should be ignored because the theory only predicted effects for erotic stimuli.

This decision should be guided by theoretical predictions. However, Bem (2011) does not provide a clear explanation for the fact that only erotic pictures produced a significant effect. He merely states "this first experiment adopts this traditional protocol, using erotic pictures as explicit reinforcement for correct 'precognitive' guesses" (p. 408). His summary of results in Table 7 ignores the other types of stimuli. Moreover, assuming independence and 80% power, it is actually unlikely that five attempts produce only one significant result (i.e., the probability to produce more than one significant effect is 99.3%). On the other hand, other successful studies did use non-erotic stimuli, and five types of stimuli increase the type-I error from 5% to 25%. Given this uncertainty about the results in Study 1, I decided to limit the computation of the IC-index to the results reported in Bem's (2011) Table 7, which included only the erotic stimulus condition.

Another question is whether Study 7 should be included. This study produced a non-significant result. Including this study would increase the credibility of the reported results (decrease the IC-index) because low total power makes it likely that some studies produce a non-significant result. However, it is only justified to treat non-significant results as failed replications if they are presented as such. If the non-significant effect is attributed to factors that are unrelated to the theory, the finding does not constitute a failed replication and should be excluded from the computation of the IC-index. Bem's (2011) interpretation of Study 7 is ambiguous. On the one hand, he blamed his modification of the experimental procedure for the

failed replication: “It was my (wrongheaded) hunch that supraliminal exposures would be more likely to produce boredom after 10 exposures than would the subliminal exposures successfully used in our original retrospective habituation experiments” (p. 418). On the other hand, he included Study 7 in his summary table, including the non-significant p-value of .096 (one-tailed). Moreover, Study 7 produced a significant moderator effect, which makes little sense if the condition did not produce a main effect (there is no ESP, but extraverts show it more). Francis (2012b) included Study 7 in his calculation of the IC-index. For illustrative purposes, I computed IC-indices with and without Study 7.

Another decision concerns the number of hypotheses that should be examined. Just like multiple studies reduce total power, tests of multiple hypotheses within a single study also reduce total power (Maxwell, 2004). Francis (2012b) decided to focus only on the hypothesis that ESP exists; that is, can the average individual can foresee the future. However, Bem (2011) also made predictions about individual differences in ESP. Therefore, I examined total power and credibility for all 19 effects reported in Table 7 (11 ESP effects & 8 personality effects).

After selecting the set of observed effects, it is necessary to make assumptions about the true effect sizes. Bem’s (2011) article provides several options to do so. The first approach is to use Bem’s (2011) *a priori* assumptions about effect sizes. Bem (2011) explicitly hypothesized that the main effect of ESP is weak (d = .25), and used a meta-analysis to predict a weak effect of individual differences in extraversion (r = .09). A second approach is to pool the observed effect sizes. Bem (2011) reports a mean d = .22 for the main effects. He does not report the average correlation with personality measures, but it is easy to compute the average correlation (r = .09). As sample sizes vary, it is also possible to compute the weighted average under the assumption that larger samples produce more accurate estimates of the true effect size (Francis, 2012b). The

weighted averages are similar (main: d = .21; moderator: r = .12). Finally, I used the observed effect sizes of each study.

I used G*Power 3.1.2 to obtain *post-hoc* power on the basis of effect sizes and sample sizes (Faul, Erdfelder, Buchner, & Lang, 2009). Table 2 shows the results for the main effects (i.e., average ESP performance). Total power estimates range from 1% to 13%, depending on the different assumptions about the true effect size. Despite this variation, all estimates show that the set of studies was severely underpowered to produce consistent support for ESP. The IC indices vary more dramatically from 54% to 99%. When the non-significant effect in Study 7 is included, even small differences in estimates of the true effect size produce large variation in IC-indices (54% to 92%). Bem's (2011) *a priori* effect size of d = .25 produces an IC-index of 54% that does not raise any concerns about the credibility of the results. Francis used the weighted average effect size and an IC-index of 92% to conclude that there is positive evidence for bias (IC-index > .90). IC-indices are reduced substantially by excluding the non-significant results of Study 7. The different estimation methods also produce more consistent results, ranging from 85% to 99%. The reason is that total power is hardly affected by excluding this study, but it is less credible that 9 out of 9 attempts rather than 9 out of 10 attempts were successful.

Table 3 shows the results for the moderating effects of personality. At first sight, these results may seem more credible than the significant main effects because Bem (2011) reported that three of the eight correlations were not significant. However, this finding has to be evaluated in the context of the power of Bem's (2011) studies to demonstrate personality effects. I used a one-tailed 5% significance level because Bem had an *a priori* hypothesis about a positive correlation, which leads to lower IC-indices than a two-tailed test. With Bem's (2011) hypothesis that the effect would be small (r = .09), an *a priori* power analysis showed that a sample size of

N = 761 is required to achieve 80% power. For Bem's planned sample size of 100 participants, power is only 23%. Thus, even a single study in Bem's (2011) article is severely underpowered; that is, it had a high probability (77%) to fail to confirm Bem's prediction that extraversion moderates ESP.

The non-significant studies produced a problem for the estimation of *post-hoc* power using effect sizes of individual studies because these studies also showed correlations with the reverse sign of the predicted effect. For these studies, it is inappropriate to use the observed effect size as an estimate of the true effect size. I set the true effect size for these studies to zero, which makes the IC-index even more conservative (an alternative approach would be to use negative power values for the computation of average power). All estimation methods produce total power values close to zero or zero, indicating that Bem was virtually guaranteed to obtain non-significant results. IC-indices of the reported outcome (i.e., 5 out of 8 significant results) vary from 76% to 97%. The use of effect sizes of individual studies produced the most conservative estimate, in part because this approach is conservative when the true effect sizes are small (Yuan & Maxwell, 2005).

There are other worrying signs that undermine the credibility of Bem's (2011) moderator results. Most problematic is the unusual approach to the measurement of extraversion. Rather than using a validated extraversion measure consistently across all studies, the measures varied across studies. In one study, the measure was created by "converting two items from Zuckerman's Sensation Seeking Scale (1974) into true/false statements" (p. 417). In the next step, these items were not combined in the traditional manner by averaging the two scores, but the items were scored so that "participants who endorsed both statements, were defined as erotic

stimulus seekers" (p. 417). The high IC-indices suggest that this questionable approach to the measurement of extraversion contributed to the incredibly high number of significant results.

To evaluate the total credibility of main effects and moderator effects, it is possible to compute the inverse IC-indices (1 – IC-index), multiply the two probabilities, and subtract the resulting probability from 1 (1 – [[1-IC-main] * [1-[IC-moderator]]). Using the lowest IC-values produces a total IC-index of 89%. Using the highest IC-values produces a total IC-index of 99.97%.

In sum, it is unlikely that Bem (2011) conducted 10 studies, ran 19 statistical tests of planned hypotheses, and obtained 14 statistically significant results. Yet, the editors felt compelled to publish the manuscript because "we can only take the author at his word that his data are in fact genuine and that the reported findings have not been taken from a larger set of unpublished studies showing null effects" (Judd & Gawronski, 2011, p. 406). The IC-index provides quantitative information about the credibility of this assumption, and would have provided the editors with objective information to guide their decision. More importantly, awareness about total power could have helped Bem to plan fewer studies with higher total power to provide more credible evidence for his hypotheses.

Example 2: Sugar High - When Rewards Undermine Self-Control

Because Bem's (2011) article is exceptional in that it examined a controversial phenomenon, I used another nine-study article that was published in the prestigious *Journal of Personality and Social Psychology* to demonstrate that low total power is also a problem for articles that elicit less skepticism because they investigate less controversial hypotheses. Gailliot et al. (2007) examined the relation between blood glucose levels and self-regulation. I also chose this article because it has attracted a lot of attention (142 citations in *Web of Science* as of May,

2012; average 24 citations per year) and it is possible to evaluate the replicability of the original findings on the basis of subsequent studies by other researchers (Dvorak & Simons, 2009; Kurzban, 2010).

The article tested three hypotheses. Studies 1 and 2 examined whether self-regulation lowers glucose levels. Studies 3-6 examined whether glucose levels after a self-regulation task predict performance on a second task that requires cognitive resources. Studies 7-9 examined the influence of experimental manipulations of glucose levels after self-regulation on a variety of dependent variables. As studies tested different hypotheses and used different designs, it is problematic to assume a common effect size for all studies. For example, the size of the effect of self-regulation on blood-glucose levels is independent of the size of the effect of blood glucose levels on self-regulation. Nevertheless, I computed a variety of IC-indices using different estimation methods of the true effect size for instructive purposes. I was not able to use *a priori* effect sizes because the authors did not make quantitative predictions about effect sizes.

Table 4 shows that sample sizes were modest, ranging from $N = 12$ to 102. Four studies had sample sizes of $N < 20$, which Simmons et al. (2011) considered to require special justification. The total N is 359 participants. Table 1 shows that this total sample size is sufficient to have 80% total power for four large effects or two moderate effects and insufficient to demonstrate a small effect. Notably, Table 4 shows that all 9 reported studies produced significant results. To examine the credibility of this outcome, I computed total power and IC-indices. In the special case when all outcomes are significant, the IC-index is the inverse of total power. Effect sizes of the various paradigms were converted into d-values. Effect sizes varied considerably from $d = .23$ to $d = 1.53$. This variation was strongly negatively related to sample sizes, $r = -.79$, without any explicit explanations for this correlation. The average effect size is *d*

= .89, and the weighted average is d = .61. Total power values for all three estimation methods are less than 1% and all IC-indices are greater than 99%. This indicates that from a statistical point of view, Bem's (2011) evidence for ESP is more credible than Gailliot et al.'s (2007) evidence for a role of blood-glucose in self-regulation.

Several follow-up studies provide an opportunity to evaluate the IC-index as a measure of credibility. Kurzban (2010) noted that the effect sizes in Studies 1 and 2 seemed to be too large given existing knowledge about the glucose consumption of the entire brain. Therefore, Kurzban (2010) requested Gailliot et al.'s (2007) data. He was informed that the data from Study 1 were corrupted and not available, but he obtained the data from studies 3 to 6, which also provided information about the influence of self-regulation on glucose consumption although these results were not reported in the original article. The main finding was that across the four studies, the results showed no significant decrease in glucose levels. This shows that one reason for too many significant results in Gailliot et al.'s article is simply the omission of statistical tests that did not confirm theoretical predictions (Kerr, 1998; Maxwell, 2004). At the same time, the failure to replicate the effect in these studies does not provide strong support for the null-hypothesis because the power in these studies is very small.

A more powerful replication study with N = 180 participants provides more conclusive evidence (Dvorak & Simons, 2009). This study actually replicated Gailliot et al.'s (1997) findings in Study 1. At the same time, the study failed to replicate the results for studies 3-6 in the original article. Dvorak and Simons (2009) did not report the correlation, but the authors were kind enough to provide this information. The correlation was neither significant in the experimental group, $r(90)$ = .10, nor in the control group, $r(90)$ = .03. Even in the total sample it did not reach significance, $r(180)$ = .11. It is therefore extremely likely that the original

correlations were inflated because a study with a sample of $N = 90$ has 99.9% power to produce a significant effect if the true effect size is $r = .5$. Thus, Dvorak and Simons (2009) results confirm the prediction of the IC-index that the strong correlations in the original article are incredible.

In conclusion, Gailliot et al. (2007) had limited resources to examine the role of blood-glucose in self-regulation. By attempting replications in nine studies they did not provide strong evidence for their theory. Rather, the results are incredible and difficult to replicate, presumably because the original studies yielded inflated effect sizes. A better solution would have been to test the three hypotheses in a single study with a large sample. This approach also makes it possible to test additional hypotheses, such as mediation (Dvorak & Simons, 2009). Thus, Example 2 illustrates that a single powerful study is more informative than several small studies.

General Discussion

Fifty years ago, Cohen (1962) made a fundamental contribution to psychology by emphasizing the importance of statistical power to produce strong evidence for theoretically predicted effects. He also noted that most studies at that time had only sufficient power to provide evidence for strong effects. Fifty years later, power analysis remains neglected. The prevalence of studies with insufficient power hampers scientific progress in two ways. First, there are too many type-II errors that are often falsely interpreted as evidence for the null-hypothesis (Maxwell, 2004). Second, there are too many false positive results (Sterling, 1959; Sterling et al., 1995). Replication across multiple studies within a single article has been considered a solution to these problems (Ledgerwood & Sherman, 2012). The main contribution of this article is to point out that multiple study articles do not provide more credible evidence simply because they report more statistically significant results. Given the modest power of

individual studies, it is even less credible that researchers were able to replicate results repeatedly in a series of articles than that they obtained a significant effect in a single study.

The demonstration that multiple study articles often report incredible results might help to reduce the allure of multiple study articles. This is not to say that multiple study articles are intrinsically flawed or that single study articles are superior. However, more studies are only superior if total power is held constant, yet limited resources create a trade-off between the number of studies and total power of a set of studies. To maintain credibility, it is better to maximize total power rather than number of studies. In this regard, it is encouraging that some editors no longer consider number of studies as a selection criterion for publication (Smith, 2012).

Subsequently, I first discuss the puzzling question of why power continues to be ignored despite the crucial importance of power to obtain significant results without the help of questionable research methods. I then discuss the importance of paying more attention to total power to increase the credibility of psychology as a science. Due to space limitations, I will not repeat many other valuable suggestions that have been made to improve the current scientific model (Schooler, 2011; Simmons et al., 2011; Spellman, 2012; Wagenmakers et al., 2011). In my discussion, I will refer to Bem's (2011) and Gailliot et al.'s (2007) articles, but it should be clear that these articles merely exemplify flaws in the current scientific practices in psychology.

Why Do Researchers Continue to Ignore Power?

Although Cohen's (1992) *Power Primer* article has been cited over 4,660 times (December 2011, Web of Science), it received only 25 citations in the *Journal of Personality and Social Psychology*, 4 citations in the *Journal of Experimental Psychology: General*, 22 citations in the *Journal of Abnormal Psychology*, 14 citations in the *Journal of Applied Psychology*, 7

citations in *Behavioral Neuroscience*, and 19 citations in *Developmental Psychology*. The references for the free and user-friendly software G*Power (Erdfelder, Faul, & Buchner, 1996; Faul et al., 2009; Faul, Erdfelder, Lang, & Buchner, 2007) received only 7 citations in all of the journals listed above. The low number of citations across journals form different areas of psychology suggests that power continues to be ignored in many areas of psychology.

Cohen (1992) could not understand why psychologists ignored his contribution to psychological science and wondered about the "the passive acceptance of this state of affairs by editors and reviewers" (p. 155). Maxwell (2004) proposed that researchers ignore power because they can use a shot-gun approach. That is, if Joe sprays the barn with bullets, he is likely to hit the bull's eye at least once. For example, experimental psychologists may use complex factorial designs that test multiple main effects and interactions to obtain at least one significant effect (Maxwell, 2004). Psychologists who work with many variables can test a large number of correlations to find a significant one (Kerr, 1998). Although studies with small samples have modest power to detect all significant effects (low total power), they have high power to detect at least one significant effect (Maxwell, 2004). The main problem of the shot-gun approach is that journals are filled with type-II errors and that results will be inconsistent across articles as different effects will emerge as significant in different studies.

The shotgun model is unlikely to explain incredible results in multiple study articles because the pattern of results in a set of studies has to be consistent. This has been seen as the main strength of multiple study articles (Ledgerwood & Sherman, 2012). However, low total power in multiple study articles makes it improbable that all studies produce significant results and increases the pressure on researchers to use questionable research methods to comply with the questionable selection criterion that manuscripts should report only significant results. A

simple solution to this problem would be to increase total power to avoid having to use questionable research methods. It is therefore even more puzzling why the requirement of multiple studies has not resulted in an increase in power.

One possible explanation is that researchers do not care about effect sizes. Researchers may not consider it unethical to use questionable research methods that inflate effect sizes as long as they are convinced that the sign of the reported effect is consistent with the sign of the true effect. For example, the theory that implicit attitudes are malleable is supported by a positive effect of experimental manipulations on the Implicit Association Test, no matter whether the effect size is d = .8 (Dasgupta & Greenwald, 2001) or d = .08 (Joy-Gaba & Nosek, 2010), and the influence of blood-glucose levels on self-control is supported by a strong correlation of r = .6 (Gailliot et al., 2007) and a weak correlation of r = .1 (Dvorak & Simons, 2009). The problem is that in the real world effect sizes matter. For example, it matters whether exercising for 20 minutes twice a week leads to a weight loss of 1 pound or 10 pounds. Unbiased estimates of effect sizes are also important for the integrity of the field. Initial publications with stunning and inflated effect sizes produce underpowered replication studies even if subsequent researchers use *a priori* power analysis. As failed replications are difficult to publish, inflated effect sizes are persistent and can bias estimates of true effect sizes in meta-analyses. Failed replication studies in file-drawers also waste valuable resources (Spellman, 2012). In comparison to one small (N = 40) published study with an inflated effect size and 9 replication studies with non-significant replications in file-drawers (N = 360), it would have been better to pool the resources of all 10 studies for one strong test of an important hypothesis (N = 400).

A related explanation is that true effect sizes are often likely to be small to moderate, and that researchers may not have sufficient resources for unbiased tests of their hypotheses. As a

result, they have to rely on fortune (Wegner, 1992) or questionable research methods (Simmons et al., 2011; Vul et al., 2009) to report inflated observed effect sizes that reach statistical significance in small samples. For example, an effect of blood-glucose on self-control is likely to be small, $r < .10$ (Dvorak & Simons, 2009). Thus, even a single test of this hypothesis would require a sample size of more than $N = 782$ participants to have 80% power to obtain a significant result.

Another reason for ignoring power could be that humans are risk-averse (Kahneman & Tversky, 1979) and that conducting a single powerful test is riskier than conducting several underpowered tests of a hypothesis. For example, a simple between-subject experiment with $N = 40$ (cell size $n = 20$), has only 33% power to find a moderate effect of $d = .5$. This would mean every third study produces a significant effect, whereas the other two studies are wasted and produce a false-negative result. Yet, if all 120 participants of the three studies were used for a single study, power increases only to 78%. Thus, the researcher who conducts a single study with high power has a lower chance to obtain a significant result than the researcher who tries three times with smaller samples. The main problem of using multiple small samples is that published effect sizes will be inflated and that researchers who conduct replication studies with samples that produced a significant result one out of three times are more likely to fail to replicate a result than to obtain a successful replication. These researchers will be discouraged and lose interest in a phenomenon, although the phenomenon is actually real.

Another explanation is that researchers prefer small samples to large samples *because* small samples have less power. When publications do not report effect sizes, sample sizes becomes an imperfect indicator of effect sizes because only strong effects reach significance in small samples. This has led to the flawed perception that effect sizes in large samples have no

practical significance because even effects without practical significance can reach statistical significance (cf. Royall, 1986). This line of reasoning is fundamentally flawed and confounds credibility of scientific evidence with effect sizes. Large samples are able to show that small effects are significant, but they can also show that large effects are significant. Moreover, tighter confidence intervals in large samples make it possible to estimate true effect sizes more precisely in large samples. Finally, large effects in small samples are likely to be inflated and may be difficult to replicate. It is therefore incorrect to use significance in small samples to make inferences about practical significance. Practical significance should be assessed in large studies with high ecological validity (Mitchell, 2012).

The most probable and banal explanation for ignoring power is poor statistical training at the undergraduate and graduate level. Discussions with colleagues and graduate students suggest that power analysis is mentioned, but without a sense of importance. Research articles also reinforce the impression that power analysis is not important as sample sizes vary seemingly at random from study to study or article to article. As a result, most researchers probably do not know how risky their studies are and how lucky they are when they do get significant and inflated effects. I hope that this article will change this, and that readers take total power into account when they read the next article with five or more studies and ten or more significant results and wonder whether they witnessed a sharpshooter or saw a magic show.

Finally, it is possible that researchers ignore power simply because they follow current practices in the field. Few scientists are surprised that published findings are too good to be true. Indeed, a common response to presentations of this work has been that the IC-index only shows the obvious. Everybody knows that researchers use a number of questionable research practices to increase their chances of reporting significant results, and a high percentage of researchers

admit to using these practices, presumably because they do not consider them to be questionable (John et al., 2012). The benign view of current practices is that successful studies provide all of the relevant information. Nobody wants to know about all the failed attempts of alchemists to turn base metals into gold, but everybody would want to know about a process that actually achieves this goal. However, this logic rests on the assumption that successful studies were really successful and that unsuccessful studies were really flawed. Given the modest power of studies, this conclusion is rarely justified (Maxwell, 2004).

In clinical drug trials, it would be extremely problematic to disregard failed studies (Bekelman, Li, & Gross, 2003; Cipriani et al., 2009; Kirsch et al., 2008). If only successful studies are used as evidence for effectiveness of a new drug, the only gold that has been created ends up in the hands of investors in companies that profit from selling ineffective drugs. Studies with low power can also often fail to demonstrate effectiveness in conditions that may produce stronger effects or could have other advantages (e.g., a drug could have fewer side effects or would be cheaper to produce). As a result, it is misguided to focus only on a select set of studies with significant results.

In medical drug trials, the occurrence of failed studies is actually very common. An astonishing 50% of stage III drug trials, the last hurdle before a drug can be approved and sold, produce non-significant results (https://www.mckinseyquarterly.com/ Why_drugs_fall_short_in _late -stage_trials_1879). This is even more astonishing as effectiveness is tested in stage-II drug trials. This finding essentially shows a 50% failure rate to replicate effects that were significant during stage II testing. The rate of failure is especially common for drugs that are based on novel mechanisms, which makes these studies more similar to studies published in top psychological

journals that place a premium on new discoveries. In contrast to 50% failure rates in drug trials, the failure rate in psychological journals is close to zero (Sterling et al., 1995).

Low total power and high IC-indices suggest that the main reason for this low failure rate is not that psychological research is more robust. A more likely explanation is that psychological discoveries are never subjected to rigorous tests equivalent to stage-III drug trials. To improve the status of psychological science it will be important to elevate the scientific standards of the field. Rather than pointing to limited resources as an excuse, researchers should allocate resources more wisely (spend less money on underpowered studies) and conduct more relevant research that can attract more funding.

I think it would be a mistake to excuse the use of questionable research practices by pointing out that false discoveries in psychological research have less dramatic consequences than drugs with little benefits, huge costs, and potential side effects. Students and the general public are interested in psychology because they assume that it provides scientific answers to questions about human behavior. However, psychological research can only fulfill this promise if the scientific process produces credible findings. Therefore, I disagree with Bem's (2000) view that psychologists should "err on the side of discovery" (p. 5). The aim of scientific inquiry is to reduce errors in humans' understanding of themselves and the world, not to commit the right type of error.

Recommendations for Improvement

Use Power in the Evaluation of Manuscripts

Power analysis is slowly becoming more common (Maxwell, 2004). Granting agencies often ask that researchers plan studies with adequate power (Fritz & MacKinnon, 2007). However, power analysis is ignored when researchers report their results. The reason is probably

that (*a priori*) power analysis is only seen as a way to ensure that a study produces a significant result. Once a significant finding has been found, low power no longer seems to be a problem. After all, a significant effect was found (in one condition, for male participants, after excluding two outliers, $p < .07$, one-tailed).

One way to improve psychological science is to require researchers to justify sample sizes in the method section. For multiple study articles, researchers should be asked to compute total power. If a study has 80% total power, researches should also explain how they would deal with the possible outcome of a non-significant result. Maybe it would change perception of research contributions when a research article reports 10 significant results, although power was only sufficient to obtain six. Implementing this policy would be simple. Thus, it is up to editors to realize the importance of statistical power and to make power an evaluation criterion in the review process (Cohen, 1992). Implementing this policy could change the hierarchy of psychological journals. Top journals would no longer be the journals with the most inflated effect sizes, but rather journals with the most powerful studies and the most credible scientific evidence.

Reward Effort Rather than Number of Significant Results

Another recommendation is to pay more attention to the total effort that went into an empirical study rather than the number of significant *p*-values. The requirement to have multiple studies with no guidelines about power encourages a frantic empiricism in which researchers will conduct as many cheap and easy studies as possible to find a set of significant results. It is simply too costly for researchers to invest in studies with observation of real behaviors, high ecological validity, or longitudinal assessments that take time and may produce a non-significant result. Given the current environmental pressures, a low quality/high quantity strategy is more

adaptive and will ensure survival (publish or perish) and reproductive success (more graduate students who pursue a low quality/high quantity strategy) (Ledgerwood & Sherman, 2012).

A common misperception is that multiple study articles should be rewarded because they required more effort than a single study. However, the number of studies is often a function of the difficulty to conduct research. It is therefore extremely problematic to assume that multiple studies are more valuable than single studies. A single longitudinal study can be costly, but can answer questions that multiple cross-sectional studies cannot answer. For example, one of the most important developments in psychological measurement has been the development of the Implicit Association Test (Greenwald, McGhee, & Schwartz, 1998). A widespread belief about the IAT is that it measures implicit attitudes that are more stable than explicit attitudes (Gawronski, 2009), but there exist hardly any longitudinal studies of the stability of implicit attitudes. One possible explanation for the gap in this literature is that in an environment that rewards multiple studies, it is too costly to conduct a single longitudinal study that is not publishable in journals that favor multiple studies. Similarly, Baumeister, Vohs, and Funder (2009) observed a lack of studies of actual behavior. The reason may be that these studies are costly and that editors favor multiple reaction time studies over a single study of real behavior.

A simple way to change the incentive structure in the field is to undermine the false belief that multiple study articles are better than single-study articles. Often multiple studies are better combined into a single study. For example, one article published four studies that were identical "except that the exposure duration—suboptimal (4 ms) or optimal (1 s)—of both the initial exposure phase and the subsequent priming phase was orthogonally varied" (Murphy, Zajonc, & Monahan, 1995; p. 589). In other words, the four studies were four conditions of a 2 x 2 design. It would have been more efficient and informative to combine the information of all studies in a

single study. In fact, after reporting each study individually, the authors report the results of a combined analysis. “When all four studies are entered into a single analysis, a clear pattern emerges” (p. 600). Although this article may be the most extreme example of unnecessary multiplicity, other multiple study articles could also be more informative by reducing the number of studies in a single article. Apparently, readers of scientific articles are aware of the limited information gain provided by multiple study articles because citation counts show that multiple study articles do not have more impact than single study articles (Haslam et al., 2008). Thus, editors should avoid using number of studies as a criterion for accepting articles.

Allow Publication of Non-Significant Results

The main point of the IC-Index is to alert researchers, reviewers, editors, and readers of scientific articles that a series of studies that produced only significant results is not a cause for celebration, nor strong evidence for the demonstration of a scientific discovery; at least not without a power analysis that shows the results are credible. Given the typical power of psychological studies, non-significant findings should be obtained regularly and the absence of non-significant results raises concerns about the credibility of published research findings. Most of the time, biases may be benign and simply produce inflated effect sizes, but occasionally it is possible that biases may have more serious consequences (e.g., demonstrate phenomena that do not exist).

The IC-Index implies that a perfect pattern of results includes some imperfection because nothing is absolutely perfect. A perfectly planned set of five studies, where each study has 80% power, is expected to produce one non-significant result. It is not clear why editors sometimes ask researches to remove studies with non-significant results. Science is not a beauty-contest and a non-significant result is not a blemish.

This wisdom is captured in the Japanese concept of *wabi-sabi*, in which beautiful objects are designed to have a superficial imperfection as a reminder that nothing is perfect. Based on this conception of beauty, a truly perfect set of studies is one that echoes the imperfection of reality by including failed studies or studies that did not produce significant results. Even if these studies are not reported in great detail, it might be useful to describe failed studies and explain how they informed the development of studies that produced significant results. Another possibility is to honestly report that a study failed to produce a significant result with a sample size that provided 80% power and that the researcher then added more participants to increase power to 95%. This is different from snooping (looking at the data until a significant result has been found), especially if it is stated clearly that the sample size was increased because the effect was not significant with the originally planned sample size, and the significance test is adjusted to take into account that two significance tests were performed.

The IC-index rewards honest reporting of results because reporting of null-findings render the number of significant results more consistent with the total power of the studies. In contrast, a high IC-index can undermine the allure of articles that report more significant results than the power of the studies warrants. In this way, *post-hoc* power analysis could have the beneficial effect that researchers finally start paying more attention to *a priori* power.

Limited resources may make it difficult to achieve high total power. When total power is modest, it becomes important to report non-significant results. One way to report non-significant results would be to limit detailed discussion to successful studies, but to include studies with non-significant results in a meta-analysis. For example, Bem (2011) reported a meta-analysis of all studies covered in the article. However, he also mentions several pilot studies and a smaller study that failed to produce a significant result. To reduce bias and increase credibility, pilot

studies or other failed studies could be included in a meta-analysis at the end of a multiple study article. The meta-analysis could show that the effect is significant across an unbiased sample of studies that produced significant and non-significant results. This overall effect is functionally equivalent to the test of the hypothesis in a single study with high power. Importantly, the meta-analysis is only credible if it includes non-significant results.

It is also important that top journals publish failed replication studies. The reason is that top-journals are partially responsible for the contribution of questionable research practices to published research findings. These journals look for novel and groundbreaking studies that will garner many citations to solidify their position as top journals. As everywhere else (e.g., investing), the higher pay-off comes with a higher risk. In this case, the risk is publishing false results. Moreover, the incentives for researchers to get published in top journals or get tenure at ivy-league universities increases the probability that questionable research practices contribute to articles in the top journals (Ledford, 2010). Stapel faked data to get a publication in *Science*, not to get a publication in *Psychological Reports*.

There are positive signs that some journal editors are recognizing their responsibility for publication bias (Dirnagl & Lauritzen, 2010). The medical journal *Journal of Cerebral Blood Flow and Metabolism* created a section that allows researchers to publish studies with disconfirmatory evidence so that this evidence is published in the same journal. One major advantage of having this section in top journals is that it may change the evaluation criteria of journal editors towards a more careful assessment of the type-I error when they accept a manuscript for publication. After all, it would be quite embarrassing to publish numerous articles that erred on the side of discovery, if subsequent issues reveal that these discoveries were illusory. It could also reduce the use of questionable reshearch practices by researchers eager to

publish in prestigious journals, if there was a higher likelihood that the same journal will publish failed replications by independent researchers. It might also motivate more researches to conduct rigorous replication studies, if they can bet against a finding and hope to get a publication in a prestigious journal.

The IC-index can be helpful in putting pressure on editors and journals to curb the proliferation of false positive results because it can be used to evaluate editors and journals in terms of the credibility of the results that are published in these journals. As everybody knows the value of a brand rests on trust and it is easy to destroy this value when consumers lose that trust. Journals that continue to publish incredible results and suppress contradictory replication studies are not going to survive, especially given the fact that the internet provides an opportunity for authors of repressed replication studies to get their findings out (Spellman, 2012). It is therefore important that top journals change their policy not to publish replication studies, if they want to maintain their top rating (Ritchie, Wiseman, & French, 2012b).

Whereas increasing awareness about the importance of replication studies is a step in the right direction, the current debate continues to ignore the importance of statistical power for replication studies. For example, Ritchie, Wiseman, and French (2012a) tried to replicate Bem's Study 9. The title emphasizes that there were "three unsuccessful attempts" and the abstract informs readers that they conducted "three pre-registered, independent attempts to exactly replicate" this study and that "*all three* [italics added] attempts failed to produce significant effects" (p. 1). However, the fact that three replication studies failed to replicate the original experiment is not as impressive as it sounds when the power of the replication studies is taken into account. The authors chose the experiment with the smallest sample size ($N = 50$) and by exact replication they also mean that they used the same small sample size. As noted earlier, this

study had the lowest power to detect the *a priori* effect size of $d = .25$, namely 56% power. Thus, only 1.68 experiments should have produced a significant effect, if Bem's hypothesis is true. The binomial probability of 0 out 3 significant effects is 91%. Thus, it does not reach the common 95% criterion that is typically used to infer a significant effect. Moreover, the authors increased the risk of making a type-I error that would provide further evidence for Bem's hypothesis from 5% to 15% by conducting three tests. It would have been better to combine the data of the three identical replication studies and conduct a single test of the hypothesis. With a sample size of $N = 150$, power would be 93% to produce a significant result and a non-significant result would provide stronger evidence without inflating the chances of making a type-I error. Thus, unnecessary multiplicity is as much a problem for replication studies as it is for original studies. In this regard, Dvorak and Simons's (2009) single study article with a large sample provides a positive example for a powerful replication study. Their study with $N = 90$ participants had 99.97% power to replicate the strong correlations ($r = .5$) reported by Gailliot et al. (2007). Alas, the authors did not report this finding in their original article, presumably because it would have undermined their chances to publish their exemplary study.

To increase the chances of publication for so-called negative results it might be helpful to reconsider the notion of studies with non-significant results as failed or unsuccessful studies. If psychologists were paying more attention to effect sizes, it would be possible to frame replication studies as studies that test effect sizes in previous studies. For example, Gailliot et al.'s (2007) article provided the first evidence on the correlation between blood-glucose levels and self-control. Their article suggested a strong effect, $r = .5$, which was significant in four studies. A replication study by Dvorak and Simons's (2009) produced a non-significant difference from the null-hypothesis that the true correlation is zero. However, it is also possible

to test the hypothesis that the replication study replicated the original finding of a strong correlation. A significance tests shows a significant difference, Chi^2 (N = 90, df = 1) = 12.75, p < .001. Thus, the replication study successfully refuted the null-hypothesis that the replication study produced the same effect size as the original studies. Thus, replication studies with high power can provide strong evidence that original studies produced inflated effect sizes. This finding should be publishable, unless psychologists are not concerned about the magnitude of psychological effects.

Another solution would be to ignore p-values altogether and to focus more on effect sizes and confidence intervals (Cumming & Finch, 2001). Although it is impossible to demonstrate that the true effect size is exactly zero, it is possible to estimate true effect sizes with very narrow confidence intervals. For example, a sample of N = 1,100 participants would be sufficient to demonstrate that the true effect size of ESP is 0 with a narrow confidence interval of plus or minus .05. If an even more stringent criterion is required to claim a null-effect, sample sizes would have to increase further, but there is no theoretical limit to the precision of effect size estimates. No matter whether the focus is on p-values or confidence intervals, Cohen's recommendation that bigger is better, at least for sample sizes, remains true because larger samples are needed to obtain narrow confidence intervals (Goodman & Berlin, 1994).

Conclusion

Changing paradigms is a slow process. It took decades to unsettle the stronghold of behaviorism as the main paradigm in psychology. Despite Cohen's (1962) important contribution to the field 50 years ago and repeated warnings about the problems of underpowered studies, power analysis remains neglected (Maxwell, 2004; Rossi, 1990; Sedlmeier, & Gigerenzer, 1989). I hope the IC-index can make a small contribution towards the goal of improving the scientific

standards of psychology as a science. Bem's (2011) article is not going to be a dagger in the heart of questionable research practices, but it may become the historic marker of a paradigm shift. There are positive signs in the literature on meta-analysis (Sutton & Higgins, 2008), in the search for better statistical methods (Wagenmakers, 2007), the call for more open access to data (Schooler, 2011), changes in publication practices of journals (Dirnagl & Lauritzen, 2010), and increasing awareness of the damage caused by questionable research practices (Francis, 2012a, 2012b; John et al., 2012; Kerr, 1998; Simmons et al., 2011) to be hopeful that a paradigm shift may be underway. Even the Stapel debacle, where a prominent psychologists admitted to faking data (Heatherton, 2010), may have a healthy effect on the field. After all, faking increases the type-I error by 100% and is clearly considered unethical. If questionable research practices can increase type I-error by up to 60% (Simmons et al., 2011), it becomes difficult to maintain that these widely used practices are questionable, but not unethical.

During the reign of a paradigm, it is hard to imagine that things will ever change. However, for most contemporary psychologists it is also hard to imagine that there was a time when psychology was dominated by animal research and reinforcement schedules. Older psychologists may have learned that the only constancy in life is change. I was fortunate enough to witness historic moments of change such as the falling of the Berlin Wall in 1989 and the end of behaviorism when Skinner gave his last speech at the convention of the American Psychological Association in 1990. In front of a packed auditorium, Skinner compared cognitivism to creationism. There was dead silence, made more audible by a handful of grey-haired members in the audience who applauded him. I can only hope to live long enough to see the time when Cohen's valuable contribution to psychological science will gain the prominence that it deserves. A better understanding of the need for power will not solve all problems, but it

will go a long way towards improving the quality of empirical studies and the credibility of results published in psychological journals. Learning about power not only empowers researchers to conduct studies that can show real effects without the help of questionable research practices, it also empowers them to be critical consumers of published research findings. Knowledge about power is power.

References

Alcock, J.E. (2011). Back from the future: Parapsychology and the Bem affair. *Skeptical Inquirer, 35*. http://www.csicop.org/specialarticles/show/back_from_the_future

Balcetis, E., & Dunning, D. (2010). Wishful seeing: More desired objects are seen as closer. *Psychological Science, 21*, 147-152. doi: 10.1177/0956797609356283

Balcetis, E., & Dunning, D. (2012). A false-positive error in search in selective reporting: A refutation of Francis. *i-Perception, 3*.

Bonett, D. G. (2009). Meta-analytic interval estimation for standardized and unstandardized mean differences. *Psychological Methods, 14*, 225-238. doi: 10.1037/a0016619.

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science, 2*, 396-403. doi:10.1111/j.1745-6916.2007.00051.x

Bekelman, J. E., Li, Y., & Gross, C. P. (2003). Scope and impact of financial conflicts of interest in biomedical research - A systematic review. *Jama-Journal of the American Medical Association, 289*, 454-465. doi:10.1001/jama.289.4.454

Bem, D. J. (2000). Writing an empirical article. In R. J. Sternberg (Ed.), *Guide to publishing in psychological journals* (pp. 3-16). Cambridge, England: Cambridge University Press.

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

Castelvecchi, D. (2011). Has the Higgs been discovered? Physicists gear up for watershed announcement. *Scientific American*. http://www.scientificamerican.com /article.cfm?id=higgs-lhc.

Cipriani, A., Furukawa, T. A., Salanti, G., Geddes, J. R., Higgins, J. P. T., Churchill, R., . . . Barbui, C. (2009). Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. *Lancet, 373,* 746-758. doi:10.1016/S0140-6736(09)60046-5

Cohen, J. (1962). Statistical power of abnormal-social psychological research: A review. *Journal of Abnormal Psychology, 65*, 145-153. doi:10.1037/h0045186

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

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159. doi:10.1037//0033-2909.112.1.155

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61*, 532-574.

Dasgupta, N., & Greenwald, A. G. (2001). On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology, 81*(5), 800-814. doi:10.1037//0022-3514.81.5.800

Diener, E. (1998). Editorial. *Journal of Personality and Social Psychology, 74*, 5-6. doi:10.1037/h0092824

Dirnagl, U., & Lauritzen, M. (2010). Fighting publication bias: Introducing the Negative Results section. *Journal of Cerebral Blood Flow and Metabolism, 30*(7), 1263-1264. doi:10.1038/jcbfm.2010.51

Dvorak, R. D., & Simons, J. S. (2009). Moderation of resource depletion in the self-control strength model: Differing effects of two modes of self-control. *Personality and Social Psychology Bulletin, 35*(5), 572-583. doi: 10.1177/0146167208330855

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods Instruments & Computers, 28*(1), 1-11. doi:10.3758/BF03203630

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *Plos One, 5*. e10068. doi:10.1371/journal.pone.0010068

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160. doi: 10.3758/brm.41.4.1149

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191. doi:10.3758/BF03193146

Fiedler, K. (2011). Voodoo correlations are everywhere: Not only in neuroscience. *Perspectives on Psychological Science, 6*(2), 163-171. doi:10.1177/1745691611400237

Francis G. (2012a). The same old New Look: Publication bias in a study of wishful seeing. *i-Perception, 3*, 176-178.

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

Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science, 18*(3), 233-239. doi:10.1111/j.1467-9280.2007.01882.x

Gailliot, M. T., Baumeister, R. F., DeWall, C. N., Maner, J. K., Plant, E. A., Tice, D. M., & Schmeichel, B. J. (2007). Self-control relies on glucose as a limited energy source: Willpower is more than a metaphor. *Journal of Personality and Social Psychology, 92*(2), 325-336. doi:10.1037/0022-3514.92.2.325

Gawronski, B. (2009). Ten frequently asked questions about implicit measures and their frequently supposed, but not entirely correct answers. *Canadian Psychology-Psychologie Canadienne, 50*(3), 141-150. doi:10.1037/a0013848

Goodman, S. N., & Berlin, J. A. (1994). The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine, 121*, 200-206.

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. doi:10.1037//0022-3514.74.6.1464

Haslam, N., Ban, L., Kaufmann, L., Loughnan, S., Peters, K., Whelan, J., & Wilson, S. (2008). What makes an article influential? Predicting impact in social and personality psychology. *Scientometrics, 76*, 169-185.

Heatherton, T. (2010). Official SPSP communiqué on the Diederik Stapel debacle. http://danaleighton.edublogs.org/2011/09/13/official-spsp-communique-on-the-diederik-stapel-debacle/ (retrieved January 1, 2012).

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*, 61-83. doi:10.1017/S0140525X0999152X

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

Ioannidis, J. P. A. (2005). Why most published research findings are false. *Plos Medicine, 2*(8), 696-701. doi:10.1371/journal.pmed.0020124

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

John, L., Loewenstein, G. F., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science*. doi:10.2139/ssrn.1996631

Joy-Gaba, J. A., & Nosek, B. A. (2010). The surprisingly limited malleability of implicit racial evaluations. *Social Psychology, 41*(3), 137-146. doi:10.1027/1864-9335/a000020

Judd, C. M., & Gawronski, B. (2011). Attitudes and Social Cognition. *Journal of Personality and Social Psychology, 100*, 406. doi: 10.1037/0022789

Kahneman, D., & Tversky, A. (1979). Prospect theory: Analysis of decision under risk. *Econometrica, 47*(2), 263-291.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*, 196-217. doi:10.1207/s15327957pspr0203_4

Kirsch, I., Deacon, B. J., Huedo-Medina, T. B., Scoboria, A., Moore, T. J., & Johnson, B. T. (2008). Initial severity and antidepressant benefits: A meta-analysis of data submitted to the food and drug administration. *Plos Medicine, 5*(2), 260-268. doi:10.1371/journal.pmed.0050045

Kurzban, R. (2010). Does the brain consume additional glucose during self-control tasks? *Evolutionary Psychology, 8*, 244-259. Retrieved from http://www.epjournal.net/wp-content/uploads/ep08244259.pdf

Ledford, H. (2010). Harvard probe kept under wraps. *Nature, 466*(7309), 908-909. doi:10.1038/466908a

Ledgerwood, A., & Sherman, J. W. (2012). Short, Sweet, and Problematic? The Rise of the Short Report in Psychological Science. *Perspectives on Psychological Science, 7*(1), 60-66. doi: 10.1177/1745691611427304

Lehrer, J. (2010). The truth wars off. *New Yorker.* p. 52. Retrieved from http://www.newyorker.com/reporting/2010/12/13/101213fa_fact_lehrer.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9*(2), 147-163. doi:10.1037/1082-989X.9.2.147

Milloy, J. S. (1995). *Science without sense: The risky buisness of public health research*. Cato Institute. http://www.junksciencearchive.com/news/sws/sws-chapter5.html. Retrieved 2011-12-21.

Mitchell, G. (2012). Revisiting Truth or Triviality: The external validity of research in the psychological laboratory. *Perspectives on Psychological Science, 7*(2), 109-117. doi: 10.1177/1745691611432343

Mudge, J. F., Baker, L. F., Edge, C. B., & Houlahan, J. E. (2012). Setting an optimal alpha that minimizes errors in null hypothesis significance tests. *Plos One, 7*. doi: e3273410.1371/journal.pone.0032734

Murphy, S. T., Zajonc, R. B., & Monahan, J. L. (1995). Additivity of non-conscious affect: Combined effects of priming and exposure. *Journal of Personality and Social Psychology, 69*, 589-602.

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. doi:10.1037/1089-2680.7.4.331

Ritchie, S. J., Wiseman, R., & French, C. C. (2012a). Failing the future: Three unsuccessful attempts to replicate Bem's 'retroactive facilitation of recall' effect. [Article]. *Plos One, 7*(3). doi: e3342310.1371/journal.pone.0033423

Ritchie, S. J., Wiseman, R., & French, C. C. (2012b). Replication, replication, replication. [Editorial Material]. *Psychologist, 25*(5), 346-348.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 Years. *Journal of Consulting and Clinical Psychology, 58*(5), 646-656. doi:10.1037//0022-006X.58.5.646

Royall, R. M. (1986). The effect of sample size on the meaning of significance tests. [Editorial Material]. *American Statistician, 40*(4), 313-315. doi: 10.2307/2684616

Schmidt, F. (2010). Detecting and correcting the lies that data tell. *Perspectives on Psychological Science, 5*, 233-242. doi:10.1177/1745691610369339

Schooler, J. (2011). Unpublished results hide the decline effect. *Nature, 470*(7335), 437-437. doi:10.1038/470437a

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies. *Psychological Bulletin, 105*(2), 309-316. doi:10.1037//0033-2909.105.2.309

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. doi:10.1177/0956797611417632

Spellman, B. A. (2012). Introduction to the special section: Data, data, everywhere ... Especially in my file drawer. *Perspectives on Psychological Science, 7*(1), 58-59. doi: 10.1177/1745691611432124

Smith, E. R. (2012). Editorial. *Journal of Personality and Social Psychology, 102*, 1-3.

Steen, R. G. (2011a). Retractions in the scientific literature: Do authors deliberately commit research fraud? *Journal of Medical Ethics, 37*(2), 113-117. doi: 10.1136/jme.2010.038125

Steen, R. G. (2011b). Retractions in the scientific literature: Is the incidence of research fraud increasing? *Journal of Medical Ethics, 37*(4), 249-253. doi: 10.1136/jme.2010.040923

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance: Or vice versa. *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. *American Statistician, 49*(1), 108-112. doi:10.2307/2684823

Strube, M. J. (2006). SNOOP: A program for demonstrating the consequences of premature and repeated null hypothesis testing. *Behavior Research Methods, 38*(1), 24-27. doi:10.3758/BF03192746

Sutton, A. J., & Higgins, J. P. I. (2008). Recent developments in meta-analysis. *Statistics in Medicine, 27*(5), 625-650. doi:10.1002/sim.2934

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science, 4*(3), 274-290. doi:10.1111/j.1745-6924.2009.01125.x

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14*(5), 779-804. doi:10.3758/BF03194105

Wagenmakers, E. J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of Psi: Comment on Bem (2011). *Journal of Personality and Social Psychology, 100*(3), 426-432. doi:10.1037/a0022790

Wegner, D. M. (1992). The premature demise of the solo experiment. *Personality and Social Psychology Bulletin, 18*(4), 504-508. doi:10.1177/0146167292184017

Yarkoni, T. (2009). Big correlations in little studies: Inflated fMRI correlations reflect low statistical power: Commentary on Vul et al. (2009). *Perspectives on Psychological Science, 4*(3), 294-298. doi:10.1111/j.1745-6924.2009.01127.x

Yong, E. (2012). Bad copy. *Nature, 485*(7398), 298-300.

Yuan, K. H., & Maxwell, S. (2005). On the post hoc power in testing mean differences. *Journal of Educational and Behavioral Statistics, 30*(2), 141-167. doi: 10.3102/10769986030002141

Table 1

# of Studies	Power / Study	Total N		
		Large	Moderate	Small
1	.800	52	128	788
2	.894	136	336	2068
3	.928	228	570	3522
4	.946	328	824	5096
5	.956	440	1090	6750
6	.963	540	1368	8472
7	.968	658	1652	10234
8	.972	768	1952	12064
9	.976	900	2250	13914
10	.978	1020	2560	15820

Note. Effect Sizes: Large (*d* = .8), Moderate (*d* = .5), Small (*d* = .2); Total *N* = Total number of participants for all studies; Power / Study = Power of each study to achieve 80% power for a set of studies. Sample sizes are based on a simple between-group design with alpha = .05, two-tailed using G-Power.

Table 2

Sample Sizes, Effect Sizes, Power, Total Power, and IC-Indices for Bem's Main Effects

				Power			
Study#	N	*d*	Sig	Prior	AVG	WAVG	Individual
1	100	.25	1	.80	.71	.60	.55
2	150	.20	1	.92	.85	.75	.92
3	100	.26	1	.80	.71	.60	.80
4	100	.23	1	.80	.71	.60	.80
5	100	.22	1	.80	.71	.60	.80
6a	150	.15	1	.92	.85	.75	.92
6b	150	.14	1	.92	.85	.75	.92
7	200	.09	0	.97	.93	.85	.35
8	100	.19	1	.80	.71	.60	.80
9	50	.42	1	.54	.46	.36	.54
Average							
All	120	.22	0.90	.83	.75	.65	.74
Exclude 7	111	.19	1.00	.81	.73	.62	.78
Total Power							
All				.13	.05	.01	.03
Exclude 7				.14	.05	.01	.09
IC-Index							
All				.54	.76	.92	.78
Exclude 7				.85	.94	.99	.89

Note. Prior = A prior effect size, $d = .25$, AVG = Average Effect size, $d = .22$, WAVG = Weighted Average Effect Size, $d = .19$; Individual = effect sizes of each study

Table 3

Sample Sizes, Effect Sizes, Power, Total Power, and IC-Indices for Bem's Personality Effects

				Power			
Study#	N	*r*	Sig	Prior	AVG	WAVG	Individual
1	100	.18	1	.23	.24	.34	.56
2	150	.17	1	.29	.31	.44	.67
3	100	-.05	0	.23	.24	.34	.00
4	100	-.07	0	.23	.24	.34	.00
6b	150	.24	1	.29	.31	.44	.91
7	200	.16	1	.34	.37	.54	.63
8	100	.22	1	.23	.23	.34	.73
9	50	-.10	0	.15	.16	.21	.00
Average	119	.09	0.63	.25	.26	.37	.44
Total Power				< .001	< .001	< .001	0
IC-Index				.97	.97	.87	.76

Note. Prior = A prior effect size, r = .09, AVG = Average Effect size, r = .09, WAVG = Weighted Average Effect Size, r = .12; Individual = effect sizes of each study

Table 4

Sample Sizes, Effect Sizes, Power, Total Power, and IC-Indices for Gailliot et al. (2007)

Study#	N	Test	*d*	Sig	AVG	WAVG	Individual
1	102	ANOVA	0.25	1	.99	.86	.71
2	38	Mod. Reg.	0.35	1	.77	.45	.58
3	15	Corr	1.50	1	.39	.21	.70
4	12	Corr	1.35	1	.29	.16	.51
5	23	Corr	1.00	1	.55	.30	.61
6	17	Corr	0.95	1	.43	.23	.45
7	62	ANCOVA	0.50	1	.93	.66	.60
8	72	*t*-test	0.65	1	.96	.72	.64
9	18	Diff. Corr	1.57	1	.43	.23	.74
Average	40	-	0.89	1.00	.64	.42 -	.62
Total Power					.01	< .01 -	.01
IC-Index					.99	> .99 -	.99

Note. Effect sizes were based on transformation into d-values, original effect sizes were d-values for ANOVA, incremental R^2 for moderated regression (Mod. Reg.) and ANCOVA, *r* for correlations, and *q* for difference between correlations (Diff. Corr); sig = 1 for significant, 0 for non-significant AVG = Average Effect size, *d* = .89, WAVG = Weighted Average Effect Size, *d* = .61; Individual = effect sizes of each study

Figure Captions

Figure 1

Total Power in a set of studies as a function of power in a single study.

Note. Total Power = Probability to obtain only significant effects in a set of studies.

Figure 1

Total Power as a Function of Number of Studies and Power of Individual Studies