3

Psychological Science's Aversion to the Null, and Why Many of the Things You Think Are True, Aren't

Moritz Heene and Christopher J. Ferguson

Since you, as with all human beings, have the ability to foretell the future, you probably already know what we are going to say in this chapter. All right, we are being a little cheeky, but a recent article in a top journal in social and personality psychology did indeed claim that some people possess a form of ESP, the ability to foretell the future, at least to some small degree (Bem, 2011a; see Chapter 14). For instance, participants in the experiments showed an ability to predict the location of an erotic image in a larger frame without seeing it first, or were better at recalling words from a list of random words that they would later be asked to type than words they were not. Psychologist Daryl Bem reported on nine experiments in which the effect size of psi, that is, the power of its influence across all nine experiments, was equivalent to a standardized mean difference of $d = 0.22$, a relatively small effect, but not markedly different from the size of effects seen in much of social psychology. In a previous meta-analysis, which is a quantitative summary of research in the field, Bem and a parapsychology co-author (Bem & Honorton, 1994) suggested that the effect size of psi was greater than some important medical effects, such as taking aspirin to prevent heart attacks in people with a prior history of such attacks.

Nonetheless, it was the Bem (2011a) paper that set off some considerable "soul searching" in the field about what we are doing and how we assemble evidence for our theories (LeBel & Peters, 2011; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). The Bem (2011a) study is being parsed by both supporters and detractors (see Alcock, 2011a, 2011b; Bem, 2011b; see also, Chapter 14). We suggest here that the problems identified with the Bem paper were easy to detect due to its attention-grabbing topic, whereas many other papers with a similar level of flaws on mundane subjects may go undetected. We must ask if it is possible to publish a study

Psychological Science Under Scrutiny: Recent Challenges and Proposed Solutions, First Edition.
Edited by Scott O. Lilienfeld and Irwin D. Waldman.

in the top journal of social and personality psychology suggesting that we can read minds and foretell the future, something most of us understand *simply is not true*, then how many other theories survive not because they are "real" but simply because they are more plausible and thus never come under the scrutiny given to psi?

It is not our intent to be either critical or supportive of Bem (2011a); on the contrary, it is our assertion (similar to LeBel & Peters, 2011) that the Bem study is "low-hanging fruit" in which the topic allowed for particular scrutiny of methods that rarely occurs for articles on more mundane topics. We submit instead that the Bem study serves as a red flag that there are very likely many other theoretical ideas that students and the general public may believe are "true" without realizing that the data assembled for such theories may be as dubious as that for psi. Indeed, other paradigms once accepted as true have recently come under increased scrutiny and skepticism, including social priming (Pashler et al., 2012; see Chapter 9), which involves the belief that much of our behavior is automatically and unconsciously altered by subtle or even subliminal primes in the social environment, and beliefs in strong media violence effects (Ferguson, 2013; Granic, Lobel, & Engels, 2013). Indeed, false beliefs in weak hypotheses or theories are sufficiently widespread that some scholars have posited that many, if not most, of our beloved theories and the empirical publications that support them may be false (Ioannidis, 2005).

We wish to be clear at the outset of this chapter that it is not our intent to critique psychological research in the postmodernist sense; that is to say, we do not mean to imply that all knowledge is equal or that empiricism is a hopeless enterprise. Rather, we assert that psychological science remains rooted in some practices that mire it largely within the realm of *protoscience*, which we define as a knowledge-seeking endeavor that posits testable hypotheses, but which may inadvertently engage in practices that prevent the falsification of those hypotheses required of a true science. Unlike a *pseudoscience*, in which a particular belief is maintained despite convincing evidence to the contrary (National Science Foundation, 2002), a *protoscience* does not have a particular belief system as the end goal, and is thus at least open to change. Some elements of psychological science that are particularly rigid, ideological, or quasi-religious, perhaps in pursuit of particular advocacy goals, may indeed be pseudoscientific, but we certainly do not indict the entire field as such. We hope our chapter may elucidate certain practices in psychological science that impede it from it reaching its full potential. We focus particularly on psychological science's long-standing aversion to null results and how this aversion detracts from the emphasis on falsification necessary for a true science.

Falsification and Null Results

For a theory to be testable and, thereby, scientific, it must be possible to prove the theory wrong. Let us say, for instance, that we come up with the hypothesis that participating in role-playing action video games in which we wield a bow and arrow improves our archery skills in real life (the authors of this chapter could not hit

the broad side of a barn from 15 m, so there is much room for improvement). We perform a pretest with a target from 15 m and find that, on average, out of 30 shots in a given trial, we are able to hit the target 10 times. We might get a little practice effect if we ran multiple pretest trials, of course, but without formal instruction we will assume our ability levels out at 10 shots out of 30. Then we play lots of *Elder Scrolls: Oblivion*. We run a single post-test trial, and shoot an 11 out of 30! Success, right? Well, no, it is entirely possible that any given trial will shoot something other than 10, either above or below, due to chance.

So we will run 10 trials. They still average 11 out of 30, better than our previous 10 out of 30, but not by much. But is this difference big enough that we can say playing the video games had a systematic (i.e., non-random) influence? This is where inferential statistics come in. Put simply, these statistics estimate the likelihood that any difference may be due to chance alone (what is typically called the *null hypothesis*). We usually accept a 5% chance rate, hence the 0.05 standard. This means that we accept some error in reporting results, namely that 5% of positive findings will be due to chance, presuming that the null hypothesis is true. The more trials we run, the less our inferential statistics "think" any difference could be due to chance. So if we run 10 trials, our inferential statistics may calculate the chance rate as 50%. So, in statistical mumbo-jumbo, we "fail to reject the null hypothesis." If we run 100 trials, the statistics may say the chance rate is 15%, still unacceptable. So we run 100,000 trials (we have tenure, so we do not have much else to do). At this point, inferential statistics estimate that the odds of averaging 11 out of 30 instead of 10 out of 30 due to chance alone is now less than 1 in 100 (1% or $p = 0.01$). This is less than our 0.05 standard, so we reject the null hypothesis. Cue the press release "Training on role-playing games creates archery maniacs!" Of course, in a real experiment, the "trials" would be bow shots from individual participants in most cases, and there would not be any risk of a practice effect such as in our archery example.

This all sounds reasonable on the surface. However, some readers may have spotted the essential flaw in this system. We either "fail to reject the null" or "reject the null." Nowhere do we ever accept it. This approach, we argue, actually inverts proper falsification in science. In a falsifiable science, theories have either been proven false, or are *yet* to be been proven false. There are no "true" theories. Theories are never definitively proven to be true, but they garner support when they repeatedly withstand direct efforts to falsify them. But under the commonly applied system of misconceived inferential statistics and null-hypothesis significance testing (NHST), there are only theories that are "true" and theories that have not *yet* been proven true. If your inferential statistics provide an estimate of the treatment effect greater than could have occurred by chance with a probability greater than 5%, you just have not looked hard enough (see also Meehl, 1978)! Increase your sample size (essentially the equivalent to the number of trials in our archery example), or redesign the study with a more powerful stimulus and rerun it (which often can lead participants to ascertain what they are supposed to do, and pressuring them to do it, something called *demand characteristics*). Put simply, because of the way we handle our analyses, we have no clear way to say "You know what, this idea of mine is all bunk. Let me publish my

failed effort so other scientists will not waste their time on our mistake." In fact, studies that "fail to reject the null hypothesis" are typically difficult to get published. And if it is very difficult to get these published, it becomes difficult to falsify theories. And if it is very difficult to falsify theories, what we are doing is a *protoscience*, and not a proper science, or perhaps even an outright *pseudoscience* if our theories are beloved to us and we simply reject any null results outright.

We submit that there has evolved a culture in academic social sciences in which it is believed that "statistically significant" results, or those that reject the null hypothesis, are interpretable, but that results that "fail to reject the null" are not interpretable. Thus, it is common to hear that null results, studies that find no effect, may be due to "Type II error." It is quite simple to explain disappointing results as a "failure" to find a "true" population effect hidden in there somewhere. Null results are often explained away as difficult to understand, or with the implication that they could be manipulated to statistical significance through a bigger sample size or more powerful stimulus. We do not mean to imply that Type II errors never happen, quite the contrary (see Chapter 4). In fact, because psychological studies depend on random sampling, the significance tests applied in each of these studies would be expected at times to fail to reject a false null hypothesis (i.e., Type II error). By applying a statistical method called *a priori power analysis*, researchers could determine the Type II error rate. For instance, by setting the tolerable Type II error rate to 0.20, we would expect 20% failed rejections of the null hypothesis when it is false (assuming that the null hypothesis can ever realistically be considered "false"; see Cohen, 1992). But we believe that a persistent bias in explaining null results as Type II errors is one of the most problematic academic cultural influences to hold back the full scientific potential of academic social sciences.

Statistical Power

Let us take a step back to clarify the concept of statistical power, which is essential for the understanding of inferential statistics and the publication bias issue. Recall that we use empirical information as provided by a sample to draw conclusions about a parameter in the population. The population parameter under consideration can be a difference between population means of two or more groups, a correlation, a regression weight, a variance, etc., depending on the kind of information we want to obtain from the sample about the population.

After having obtained a population parameter estimate from a random sample, the question arises of how we can be sure that an estimated effect, for example a correlation of 0.20, from a finite sample is not a chance result – that is, how it can be explained by mere random sampling variation around a possible true correlation population parameter equal to zero. The short answer is that we cannot be sure. However, statistics enables us to estimate the degree of uncertainty. In inferential testing, this kind of uncertainty is expressed in terms of the probabilities of drawing the wrong or right conclusion about the effect in the population from samples.

To determine such probabilities, we need a probability model of how our data may have occurred. In inferential testing, the null hypothesis and its related sampling distribution serve as a model to explain how our observed result may have occurred. Typically, the null hypothesis, denoted as H_0, states that there is no effect in the population, for instance, that there is a zero mean difference between two populations of interest: H_0: $\mu_1 - \mu_2 = 0$. Of course, this hypothesis need not be true. So, the actual effect in the population might be different from zero. This alternative hypothesis, denoted H_A, may refer to any other parameter value. Typically, although social science researchers often have a hypothesis of the *direction* of the effect, they can be quite imprecise about hypothesizing a specific parameter value under the assumption that the H_A is true (i.e., holds in the population). Therefore, they usually formulate the H_A as H_A: $\mu_1 - \mu_2 > 0$, or H_A: $\mu_1 - \mu_2 < 0$ – that is, the mean difference is greater or smaller than zero. The direction (">0" or "<0") depends on the hypothesis derived from a theory. For instance, a learning theory may predict that, on average, an instructed learning group learns more than an uninstructed learning group (H_A: $\mu_1 - \mu_2 > 0$). Or, this theory may predict that an instructed learning group makes, on average, fewer mistakes (H_A: $\mu_1 - \mu_2 < 0$). For the sake of brevity, let us assume that H_A: $\mu_1 - \mu_2 > 0$.

Now, let us assume that we would draw random samples from the populations of interest (theoretically an infinite number of times) and conduct a significance test for each of these trials. If we then count the number of significant results for a given significance level α (e.g., 5%), we can draw either a correct or an incorrect conclusion, depending on whether the H_0 or H_A holds in the population. We then can distinguish between different errors. A Type I error occurs if the H_0 is true ($\mu_1 - \mu_2 = 0$), but it is rejected based on the basis of the significance test. The related error probability is denoted as α. The probability of correctly accepting the H_0 is then $1 - \alpha$ (i.e., inverse probability). A Type II error occurs when we accept the H_0, although it is not true ($\mu_1 - \mu_2 > 0$). Its related error probability is denoted β. The probability of *correctly* rejecting the H_0 under the assumption the alternative hypothesis H_A is true is then $1 - \beta$. Table 3.1 illustrates the possible errors and their associated probabilities.

Now, let us turn again to the issue of statistical power, which is the main focus of this section. Statistical power depends on three things: the effect size in the population, the sample size (technically, the standard error, which is, besides the variance of the outcome variable in the population, a function of sample size), and the chosen significance level. Interested readers may refer to statistics textbooks that give more

Table 3.1 Probabilities of correct and incorrect decisions.

		Reality	
Statistical decision based on the significance test		H_0 is true	H_0 is false
	Reject H_0	Incorrect decision: Type I error probability: α	Correct decision: $1 - \beta$
	Accept H_0	Correct decision: $1 - \alpha$	Incorrect decision: Type II error probability: β

detailed explanations of the dependency of power on these parameters. In general, the larger the population effect size, or the larger the sample size, or the higher the significance level, the greater the statistical power. For further clarification, consider the case of the one-sided *t*-test for independent samples, assuming equal but unknown population variances. The power is then defined as the probability to observe mean differences that are greater than the critical value W_C when the expected values of the populations 1 and 2 differ in the population. Or, more formally expressed: power $= P\ (T > W_C\ |\mu_1 \neq \mu_2)$. Referring to the example of a one-sided independent *t*-test (H_0: $\mu_1 = \mu_2$ vs. H_A: $\mu_1 > \mu_2$) and assuming equal, but unknown population variances in populations 1 and 2, the power is then defined as:

$$power = p\left(\frac{\mu_1 - \mu_2}{\sigma_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\mu_1 - \mu_2}{\sigma_{diff}} \geq W_c \right)$$

with μ_1: population mean of group 1,
μ_2: population mean of group 2,

$$\sigma_{pooled} = \sqrt{\frac{\sigma_1^2(n_1 - 1) + \sigma_2^2(n_2 - 1)}{n_1 + n_2 - 2}},$$

σ_{diff}: standard error of the mean difference,
n_1 and n_2: sample sizes of group 1 and 2, respectively.

For the sake of convenience and similarity to applied research, let us assume that a researcher has pre-experimentally fixed a critical value w_c, that is, set the significance level α to 5% and defined a specific hypothesis about the population mean difference (called delta, δ) in a standardized learning test under the alternative hypothesis $H_A: \mu_1 - \mu_2 = 0.6$. He or she thus states that the population mean difference is 0.6 standard deviations above the population mean difference under the null hypothesis: $H_0: \mu_1 - \mu_2 = 0$. Statistical power then depends on two parameters: the difference between the population means μ_1 and μ_2 and the sample sizes, which influences the standard error of the mean difference σ_{diff}, denoting the "typical" amount by which sample mean differences deviate from the population mean difference. Both properties are also illustrated in Figure 3.1, showing the sampling distributions of the mean differences under H_0 and H_A for the illustrative case of two populations differing in their mean scores on the hypothetical standardized learning test:

The blue area under the red curve to the right of the critical value, labeled as α, refers to the Type I error rate (i.e., 5% in this example). The bright red area under the blue curve to the left of critical value denotes the Type II error rate, labeled as β – that is, the probability of falsely *not rejecting* the *null* hypothesis when in fact the

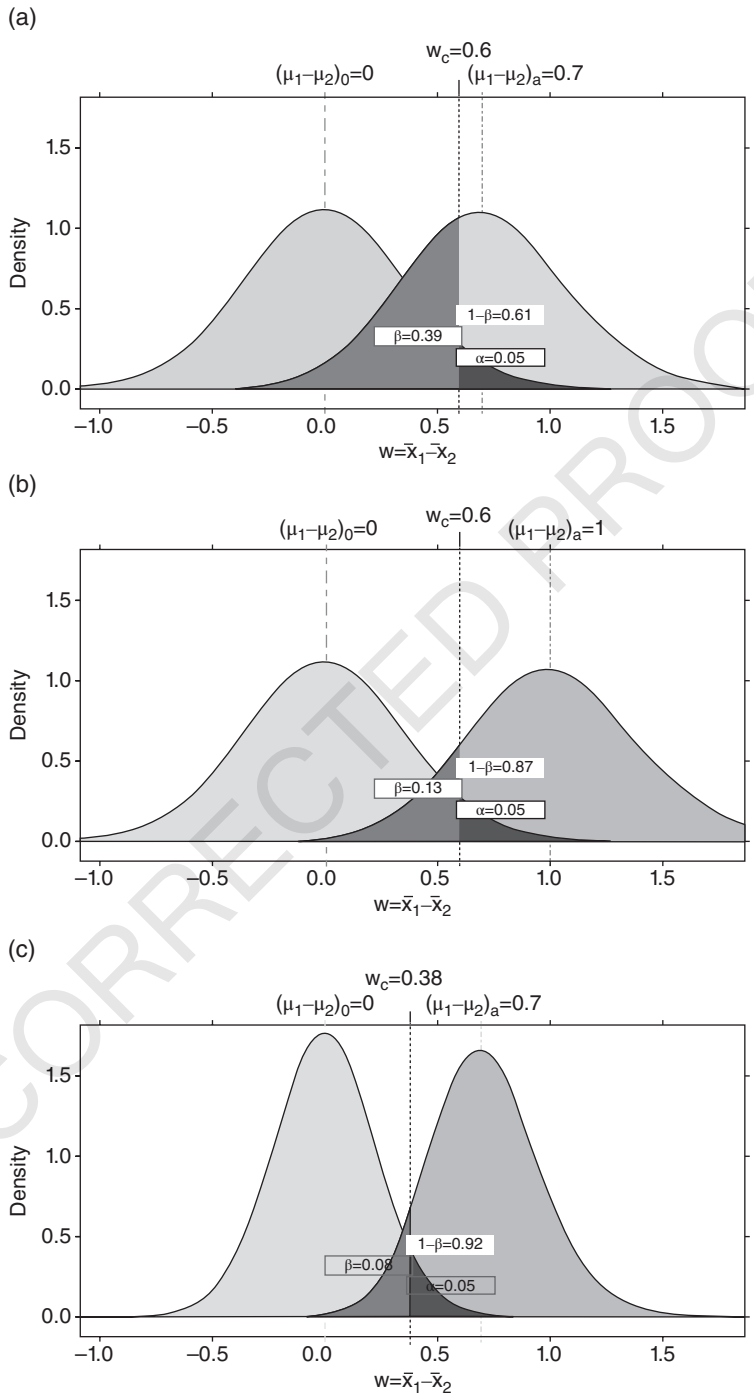

Figure 3.1 Illustration of parameters affecting power. (a) Power for population effect size $\delta = 0.7$ and sample sizes of $n_1 = n_2 = 20$. (b) Effect of greater population effect size ($\delta = 1$) on power. (c) Effect of increasing sample size to $n_1 = n_2 = 50$ on power.

alternative hypothesis is true. Hence, $1-\beta$, the reverse probability, then defines the pink area under the density curve of the sampling distribution under the H_1. This area under the curve defines the power, that is, the probability of correctly rejecting the null hypothesis (H_0: $\mu_A = \mu_B$) if the specific alternative (H_1: $\mu_A > \mu_B$) is true. As illustrated in Figure 3.1a, for a population effect size of $\delta = 0.7$, sample sizes of $n_1 = n_2 = 20$, and a significance level of 0.05 (corresponding to a critical mean difference of $w_c = 0.60$), the power is 0.61, and thus not enough to reliably reject the null hypothesis. Now, how can we improve the situation, that is, how can we increase power? As we can see from Figure 3.1b, a greater mean difference of $\delta = 1$ in the population implies a higher power, resulting in an increase in power of 0.87. In fact, while this part of the illustration is of theoretical interest, it is far-fetched from a practical point of view, just because we cannot manipulate population effect sizes. The only feasible method to increase power is to increase sample sizes. This example is shown in Figure 3.1c. By increasing the sample size from $n_1 = n_2 = 20$ to 50, the standard errors (i.e., the related standard deviations of both sampling distributions) decrease, and, thus, the pink area under the H_1 distribution *increases*, thereby increasing power to 0.92. One should also notice that the reduced standard error of the mean difference also implies an increased precision of population mean difference estimates coming from different samples, as Figure 3.1c also illustrates. Note that these estimates are now grouped closer around the population mean difference of 0.70 than those in Figure 3.1a. It is thus less likely to observe mean differences that are far off the true population mean difference of 0.7, as is the case in Figure 3.1a. Thus, increasing sample size results in two desirable properties. On the one hand, power is increased, and the trustworthiness of the decision to reject or not reject the null hypothesis based on a significance test result is increased. On the other hand, the precision of population parameter estimates is increased because the standard error is reduced.

One of the troubles with significance testing occurs when one tries to interpret a significant result from a study associated with low statistical power. Let us take the case of a researcher finding a significant mean difference ($p < 0.05$), and let the power, that is, the probability of rejecting the H_0 when it is false, be 0.50. Therefore, if one repeats the study under identical conditions with random samples, in the long run, 50% of studies will yield a significant result ($p < 0.05$). What does this actually imply for the validation of the psychological theories underlying those results? In samples with reduced power, tests of statistical significance examining a given theory under consideration will vary greatly from one sample to another, producing a pattern of apparent contradictions in the published literature. Given the low power of the study, we do not possess the required level of statistical confidence to reject the null hypothesis, or, in other words, to trust our significant result. The moral of the tale is that we are none the wiser, because we failed to decrease uncertainty by applying a significance test.

The reader may think that we have constructed a far-too-extreme and unfair example by choosing a power of 0.50. However, this value is typical, and arguably charitable, for psychology. Studies have typically found a median estimated power to

find a medium-sized effect (i.e., Cohen's $d = 0.5$ in the population) of around 35% (Bakker, Van Dijik, & Wicherts 2012, p. 544). Consider also the work by Richard, Bond, and Stokes-Zoota (2003), who reported an average power of 20% in social psychology, as well as that by Button et al. (2013), who investigated 730 neuroimaging studies in 49 meta-analyses and found a median statistical power of about 21% (see also Chapter 11).

A critical reader may object that the power problem in psychology is, in fact, troublesome in regard to the robustness of conclusions derived from published studies, but does not logically imply publication bias in the field. Although this is true, the problem of publication bias becomes more apparent if we observe more significant results from a body of research than we would predict from the power of the studies. To give a crude illustration: if a power analysis of published research predicted 50% significant results, but we observed 80 significant out of 100 published studies, then this result would be too good to be true. Indeed, there is increasing evidence that published literature in psychology shows this excess of significant findings (Button et al., 2013; Francis, 2013, 2014; Levine, Asada, & Carpenter, 2009; Maxwell, 2004; O'Boyle, Banks, & Gonzalez-Mulé, 2014). Put another way, the problem of power is not merely that *true effects* are being missed due to small sample sizes, but that, with inadequate power, it is more likely that significant results will be false positives. Thus, it is difficult to have confidence in a positive result, a problem known as the "winner's curse" (Button et al., 2013). Our suspicion, however, supported by the data already discussed, is that issues of inadequate power more often come into play for null results than for results that manage, even if by chance, to reach statistical significance.

The File-Drawer Problem

The result of psychological science's aversion to the null is a problem that has been known for quite some time. It is often called the *file-drawer problem* or, more technically, *publication bias* (Rosenthal, 1979). Put simply, publication bias arises whenever the probability that a study is published depends on the statistical significance of its results (Scargle, 2000; Schonemann & Scargle, 2008). "The extreme view of this problem, the 'file drawer problem', is that the journals are filled with the 5% of the studies that show Type I errors [rejection of a true null hypothesis], while the file drawers back at the lab are filled with the 95% of the studies that show nonsignificant (e.g., $p > 0.05$) results" (Rosenthal, 1979, p. 638). In fact, studies published in journals (and reported to the general public) most often support a psychological theory rather than refute it even if the theory in real-life is false (Fanelli, 2010, 2012; Ioannidis, 2005). As noted earlier, certain statistical arguments have been developed to resist null results, but we argue that there is something of an emotional bias toward positive findings. Certainly, as scholars, we tend to think our ideas are clever, and are biased toward results that find those ideas to be "true." If we create the "Ferguson and Heene Theory of Everything," we will be quite disposed toward

research results that support it and negatively inclined or even hostile toward results that refute it, which could express itself when we serve as peer reviewers for papers. Psychological science benefits, at least in the short-term, by being able to present nifty theories as "true" to the general public. Look at all the cool stuff we can do (with our grant money, news headlines, kids we send to others as psychology majors, etc.)! In contrast, a failed theory is a blow to psychological science, or at least that may be the perception.

The result is a remarkable string of apparent successes for psychological theories in published articles. For example, Fanelli (2010) found that theory-supportive results are far more prevalent in psychology and psychiatry than in the so-called hard sciences (91.5% vs. 70.2% in the space sciences, for instance). That is to say, in the hard sciences, with their arguably more standardized and rigorous methods, scientists acknowledged being wrong about their hypotheses about 30% of the time in published studies, but this happened less than 10% of the time in social sciences. Are social scientists really that much smarter than are physicists and chemists? We suspect the more likely explanation is that social sciences are more adverse to publishing null results, and the fluidity of social science methods makes it easier for scholars, even those acting in good faith, to "nudge" their results to support their preexisting beliefs. Although the problem of publication bias has been identified for some time, it appears to be getting worse, not better (Fanelli, 2012). Other researchers have confirmed this sobering finding (Kepes & McDaniel, 2013).

The prevalence of publication bias is, in general, difficult to estimate. There are statistical tools available for detecting publication bias in the context of meta-analyses (e.g., Fritz, Scherndl, & Kühberger, 2012; Ioannidis & Trikalinos, 2007). Meta-analysis is a statistical procedure designed to combine studies in a research domain to ascertain the average effect size and its heterogeneity across studies. The underlying premise of meta-analysis is that, for a "true" effect in nature, there will be variation in estimates of that effect across studies stemming from random sampling variation, and combining them can yield a better approximation of the "true" effect. Nonetheless, in the presence of publication bias, that average effect size is likely to be spurious or at least upwardly biased, as the failed replications were never included in the analysis. There are some more basic problems with how meta-analysis is used in this fashion, but we will return to this idea shortly. Ferguson and Brannick (2012) found that approximately 41% of meta-analyses reported some evidence for publication bias and, using a conservative statistical analysis, they found evidence for publication bias in roughly 25% of published meta-analyses. Searching for unpublished studies to include in meta-analyses actually tended not to help, and often made matters worse, because unpublished studies are not indexed (aside from dissertations that may be indexed in dissertation databases such as Digital Dissertations) in publicly available databases. Such searches also tended to suffer from selection bias, as indicated by overrepresentation of the meta-analytic authors themselves in included unpublished studies in proportion to their representation in published studies. This problem is compounded by the fact that statistical approaches for detecting publication bias only detect one type of bias, namely the bias *across* articles in a field.

But what also matters are biases *within* an article due to *outcome reporting bias* to create statistically significant findings. Such bias raises particular concern because it undermines the theoretical conclusions of the article. For instance, earlier we indicated that running more participants increases the likelihood of inferential statistics demonstrating that an outcome is statistically significant at the magical 0.05 level. So let us imagine we conduct an experiment testing the hypothesis that eating broccoli increases anxiety (our personal experience with broccoli suggests this hypothesis may be about right). In a well-structured randomized experimental design, we run 200 participants, some eating broccoli, some not, then give them a measure of state anxiety as the outcome variable. We find an effect size in terms of r (the correlation coefficient is often used as an easy-to-understand effect size estimate, even in experimental studies such as this one) as $r = 0.157$, with $p = 0.08$. That, sadly, does not quite fall below the 0.05 mark. So now what? Pack up our broccoli and go home? No, that $p = 0.08$ is tantalizingly close to $p = 0.05$, so what we do (not that we should, but it is what folks often do; Simmons, Nelson, & Simonsohn, 2011) is just run more participants until that $p = 0.08$ becomes a $p = 0.049$ or lower. Hence, we add 50 more participants, and voila! Our results are now significant at $p = 0.05$, even with the same effect size of $r = 0.157$. So was the effect false with a sample of 200 and now "real" with a sample of 250? As a result of this property of inferential statistics, it is not uncommon to see large studies publish tiny effect sizes, a luxury that small studies cannot afford. Unfortunately, small studies tend to produce more extreme (i.e., unexpected) results since the standard error of a statistic from small samples is larger. Those surprising effects are typically easier to publish because they seem to reveal something new (Schooler, 2011). Because the average published observed effect sizes from such small sample sizes will be larger than those from larger samples having smaller standard errors, a negative correlation in meta-analyses between sample sizes and effect sizes can be indicative of publication bias, and this is just the correlation we often observe in social sciences (e.g., Fritz et al., 2012). Indeed, some of the beloved theories that students believe to be "true" may be the beneficiaries of publication bias and actually may be "false."

The methodological flexibility problem

As noted earlier, most means of detecting publication bias rely on methods aimed at detecting bias across studies. Nevertheless, publication bias arises not just from journal editors' predilection for publishing positive rather than negative findings, but also from researchers' all-too-human desire to support their beloved hypotheses rather than refute them. A recent study by Franco, Malhotra, and Simonovits (2014) illustrated both points. They found that, in TESS (Time-sharing Experiments for the Social Sciences), a National Science Foundation–sponsored program, "strong results are 40 percentage points more likely to be published than are null results and 60 percentage points more likely to be written up" (p. 1502). The tendency to publish and to write up only results supporting one's hypotheses then leads to bias *within*

studies. As a result, many investigators engage in dubious research practices that "tilt the machine," changing null results to positive results without necessarily requiring substantial increases in sample size (cf., Fang, Steen, & Casadevall, 2012, and John, Loewenstein, & Prelec, 2012, for findings on the prevalence of questionable research practices and scientific misconduct; see also Chapter 5, this volume). Although some of these dubious practices involve outright fraud, the vast majority are not intentionally dishonest but rather involve humans doing what they naturally do, namely, engaging in confirmation bias – i.e., valuing evidence that supports their beliefs over evidence that does not, and nudging the statistical system until they find the results they want or expect to see.

For instance, let us return to the hypothesis that eating broccoli increases anxiety. Let us say you are absolutely sure (as we are) that this is true, but you are not able to increase your sample size (granted, if you are doing that to fall below $p = 0.05$ this is itself a dubious research practice; Simmons et al., 2011). Perhaps you have exhausted your pool of undergraduate researchers, or you are due to turn in the draft of your dissertation tomorrow. Well, you have got options! You could look for "outliers" and, finding them, kick them loose and rerun your analyses. Or you could add a plausible covariate, such as levels of depression (which tend to be moderately to highly correlated with anxiety), or remove a covariate. Or perhaps you might convince yourself that one of the items on your anxiety measure does not load well with the others, and recalculate the anxiety outcome without that item. Or you could dichotomize your outcome into "high-anxiety" and "low-anxiety" subgroups rather than using continuous anxiety scores. Or perhaps you had two separate anxiety outcomes and found statistical significance for one but not the other – so you only report the significant outcome. The options go on and on like this for a rather simple study. Of course, any of these choices might be defensible, and that is part of the issue. It is easy for researchers to convince themselves that they are doing the right thing rather than being fraudulent. However, if these changes to the design occur with the conscious or unconscious hope of changing a null result into a statistically significant one, they are questionable research practices.

Because they do not depend on sample size, these practices are harder to detect in the context of traditional publication bias, although they certainly contribute to a kind of publication bias, this time initiated by authors rather than journals. Some surveys of psychological researchers indicate that questionable research practices are common. For instance, John et al. (2012) found that approximately 63%–67% of researchers admitted to failing to report all outcomes in a published paper, presumably indicating that they found inconsistent results, but published only those that fit their hypothesis (see also Chapter 5). In a clever analysis, O'Boyle et al. (2014) examined changes in manuscripts between indexed dissertations and published papers and found that studies often underwent systematic transformations from dissertation to final published product. Specifically, inconsistent outcomes were often dropped, as were mentions of unsupported hypotheses. Moreover, the direction of predicted hypotheses were reversed, and data analyses often appeared to be altered. These practices can grossly distort our knowledge.

In John et al.'s survey, other practices were less common. About 55% of respondents reported "significance chasing," that is, adding more participants until falling below $p = 0.05$. About 35% reported "data snooping," namely, examining how excluding or including some data might influence the results and *then* picking which data to include. Recently, Francis (2014) found that 82% of studies published in the journal *Psychological Science* between 2009 and 2012 succeeded in showing a significant effect "at a rate much higher than is appropriate for the estimated effects and sample sizes" (p. 2). It is by now well known that such practices can create spurious and even absurd results in the data (Simmons et al., 2011). For example, in the field of aggression, lack of standardization in many commonly employed laboratory aggression measures has now clearly been demonstrated to result in potentially spurious results (Elson, Mohseni, Breuer, Scharkow, & Quandt, 2014).

All of these issues point to a broad and systemic cultural ethos in which scientists are trying too hard to support their *a priori* hypotheses. Most of this is presumably in good faith. Such problems are not unique to the social sciences, although because social sciences typically lack the standardization of measurement common to harder sciences, such practices may be easier in the social sciences. But this issue is one reason we refer to psychology as a *protoscience*. This is not to say that some scholars are not conducting rigorous science, or that others have become so fully wedded to ideological positions that they are practicing little more than *pseudoscience*. We are certain that the majority of psychological scholars are committed to scientific principles. Nevertheless, this dedication competes with cultural pressures to produce positive findings that may activate the natural human ability to "nudge" things in a particular direction while convincing oneself that this is the proper thing to do. As such, in the absence of standardized measures and procedures for analyzing data, much of social science runs the risk of remaining a protoscience. In the next section, we discuss potential avenues for improving this state of affairs.

Moving From Protoscience to Science

Develop statistical tools for the analysis of null effects

Much of the resistance to null effects arises from uncertainly about whether null results are "true" or due to Type II errors. Techniques for analyzing null effects remain in their infancy. Some techniques for the testing of null results – that is to say, tests designed specifically to lend support to the null hypothesis – exist (e.g., Levine, Weber, Park, & Hullett, 2008), although they remain underutilized and are not part of standard statistical packages. Bayesian statistics may also be employed to examine relative support for the null and alternative hypotheses, although utilization of such statistics among research psychologists remains minimal.

Dissatisfaction with null-hypothesis significance testing (NHST) and the absence of a mechanism for evaluating null results have often left researchers in the mire of trying to analyze effect sizes. In 1999, the APA's Wilkinson Task force released

recommendations for the reporting of effect sizes alongside of traditional NHST. We agree that such information is valuable, although the crucial question of how researchers should interpret an effect size has often been poorly grounded and left to as much subjectivity as NHST (Cortina & Landis, 2011). As Thompson (2001) noted, the rigid use of benchmarks for "weak," "moderate," and "strong" (see Cohen, 1992, for suggested effect sizes as well as warnings that the rigid use of such recommendations would be problematic) in interpreting effect sizes is not clearly superior to NHST, yet the absence of any benchmarks at all leaves effect size interpretation up to the individual author's subjective judgment. Because it may be human nature for authors to be biased toward the subjective importance of their findings, we can observe a plethora of "small is big" arguments, in which scholars (including those who advocate for psi) compare their results with medical effects (often using flawed statistics; see Meyer et al., 2001, as an example, and Ferguson, 2009, for a discussion of the flawed statistics used to make such comparisons) or concoct other reasons why a result is important so long as it is (a) larger than $r = 0.00$, and (b) "statistically significant." Thus, arguments for the interpretation of effect sizes have hardly ushered in an era of scientific caution, given that scholars tend to interpret effect sizes to fit, by happy coincidence, with their pre-existing hypotheses. As such, statistical methods for a careful examination of null results, which would argue against the potential for Type II errors, would be welcome.

Improve the use of meta-analysis

One issue we mentioned earlier is the problematic use of meta-analysis, a statistical technique designed to surmount sampling error to estimate a true effect in the population by combining existing studies in a field in one analysis. The rationale for meta-analysis is compelling. If we expect individual studies to vary in effect sizes due to sampling error, combining them should help us to see past that error.

In contrast, if publication bias is as prevalent as it often seems to be, meta-analysis will tend to provide a spurious and biased estimate of the population parameters (cf., Scargle, 2000; Schonemann & Scargle, 2008). In a related vein, meta-analyses do not "know" which studies may be biased due to poor methods (and meta-analytic authors may inject their own biases when trying to decide which studies have poor methods and which are good). As such, the old "garbage in, garbage out" (GIGO) critique of meta-analysis, stemming from questionable research practices, simple errors, or sloppy methodology, remains a serious issue. Further, to the extent that a median effect size produced through meta-analysis is accepted as "true," meta-analyses (similar to any literature review) may do more harm than good by presenting a field as more consistent than it actually is and omitting failed replications that get lost in the shuffle. Failed replications rarely have much impact in psychology (see Chapters 1 and 2), and meta-analysis can therefore be employed as a tool to *resist* falsification by taking an "average effect size wins" approach, rather than taking failed replications seriously.

This problem is particularly prevalent in studies that rely on bivariate correlations (correlations between two variables) rather than effect size estimates that better control for potential confounds. Most scholars agree that it makes sense in individual studies to control for theoretically relevant confounds, but, in meta-analyses that rely on bivariate correlations, those controls are lost (Pratt & Cullen, 2000). Thus, it is possible for every study examining the correlation between X and Y to conclude that, although X and Y are correlated, this is probably explained by variable Z, which when controlled diminishes the correlation between X and Y to a very small magnitude that is nonsignificant. Yet, a meta-analysis of the bivariate correlation between X and Y is unable to control for Z, and thus may provide a spuriously high effect size estimate. This is obviously a serious problem for meta-analysis (and other literature reviews) that has long gone ignored. Some techniques have been developed for attempting to achieve unbiased effect size estimates by controlling for publication bias (e.g., Nelson, Simonsohn, & Simmons, 2014); however, these methods remain preliminary.

Publish replications, including failed replications

Replication is a key feature of science, yet psychology has often been replication-averse (see Chapters 1 and 2, this volume). Direct replications of studies are difficult to publish, and failed replications more difficult still. The November 2012 issue of *Perspectives on Psychological Science* (PPS) addressed this issue of a replication crisis in psychology head-on. Further, PPS has introduced a new section of the journal devoted to replication projects, which we view as an excellent step forward (Association for Psychological Science, 2013).

Journal editors could also make explicit their willingness to publish replication studies, including failed replications, and null results more specifically. Without such open calls, many scholars may justifiably continue to assume that null results remain unwelcome.

Change the academic culture

Perhaps the biggest hurdle is the need to change the academic culture. We should move toward an academic culture in which scholars accept that null results of adequately powered studies are of equal or perhaps greater importance to assessing statistical significance than positive results. Although there may be good reasons to carefully evaluate null results, similar to statistically significant results, with some degree of skepticism, null results should not be held to a higher standard than statistically significant results. Of course, there may be many reasons why null results are obtained, and not all of these necessarily represent a critical failure of a theory. Nevertheless, given the current incentive structure of scientific publishing, and the strong affection for many theories from their developers and supporters,

dismissal biases in the interpretation of null results are already substantial (see Edwards & Smith, 1996, for a discussion of "disconfirmation bias").

We still hear of journal editors who decline to publish null results, and we worry greatly about the damage this does to our understanding of multiple scientific phenomena. As more null results are published, this may gradually change, but scientific organizations such as the Association for Psychological Science (APS) and American Psychological Association can take the lead in promoting openness to null results in the journals that they publish (such as *Perspectives on Psychological Science*, published by the APS).

Ultimately, the academic culture will change not based on a sudden spark or landmark event, but based on thousands of individual decisions among scholars who make insistent efforts to see null results published, or who are able to pull back from the temptation of methodological flexibility in converting null results to those that are statistically significant. We believe that, with the investment of the psychological community and continued attention to this matter, our field can improve and, with an openness to null results, make a concerted move beyond protoscience into genuine science.

Glossary

Falsification: A quality of scientific theories wherein it is possible to empirically demonstrate that a theory is incorrect. A theory that cannot possibly be proven incorrect is unfalsifiable and, thus, unscientific.

Meta-analysis: A statistical tool for estimating the weighted average effect size of all studies in a particular domain.

Methodological flexibility: Also called "researcher degrees of freedom." Refers to the ability for researchers, even acting in good faith, to convert null results to statistically significant results via manipulations of their methodology or statistical analysis.

Null result: Broadly speaking, a study result that fails to support the study hypotheses by suggesting that any observed effects are likely due to chance.

Null-hypothesis significance testing: Standard statistical analyses in the social sciences designed to test the probability that an observed effect is due to chance.

Publication bias: The tendency for journals to demonstrate preference for publishing statistically significant findings rather than null results.

References

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

Alcock, J. (2011b). Response to Bem's comments. *Skeptical Inquirer.* Retrieved from: http://www.csicop.org/specialarticles/show/response_to_bems_comments

Association for Psychological Science. (2013). Registered replication reports. Retrieved from: http://www.psychologicalscience.org/index.php/replication

Bakker, M., Dijk, A. van, & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554. doi: 10.1177/1745691612459060

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

Bem, D. (2011b). Response to Alcock's "back from the future: Comments on Bem." *Skeptical Inquirer.* Retrieved from: http://www.csicop.org/specialarticles/show/response_to_alcocks_back_from_the_future_comments_on_bem

Bem, D., & Honorton, C. (1994). Does psi exist? Replicable evidence for an anomalous process of information transfer. *Psychological Bulletin*, *115*, 4–18.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376.

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

Cortina, J. M., & Landis, R. S. (2011). The earth is not round (p = 0.00). *Organizational Research Methods*, *14*(2), 332–349. doi: 10.1177/1094428110391542

Edwards, K., & Smith, E. E. (1996). A disconfirmation bias in the evaluation of arguments. *Journal of Personality and Social Psychology*, *71*, 5–24.

Elson, M., Mohseni, R., Breuer, J., Scharkow, M., & Quandt, T. (2014). Press CRTT to measure aggressive behavior: The unstandardized use of the competitive reaction time test in aggression research. *Psychological Assessment*, *26*(2), 419–432. doi:10.1037/a0035569

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLoS ONE*, *5*(4), e10068.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*, 891–904.

Fang, F. C., Steen, R. G., & Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, *109*(42), 17028–17033. doi:10.1073/pnas.1212247109

Ferguson, C. J. (2009). Is psychological research really as good as medical research? Effect size comparisons between psychology and medicine. *Review of General Psychology*, *13*(2), 130–136.

Ferguson, C. J. (2013). Violent video games and the Supreme Court: Lessons for the scientific community in the wake of Brown v EMA. *American Psychologist*, *68*(2), 57–74.

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling and implications for the use of meta-analyses. *Psychological Methods*, *17*(1), 120–128.

Francis, G. (2013). Replication, statistical consistency, and publication bias. *Journal of Mathematical Psychology*, *57*(5), 153–169. doi: 10.1016/j.jmp.2013.02.003

Francis, G. (2014). The frequency of excess success for articles in Psychological Science. *Psychonomic Bulletin & Review*, 1–8. Retrieved from http://www1.psych.purdue.edu/~gFrancis/Publications/Francis2014PBR.pdf

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505.

Fritz, A., Scherndl, T., & Kühberger, A. (2012, April). Correlation between effect size and sample size in psychological research: Sources, consequences, and remedies. *10th Conference of the Austrian Psychological Society, Graz, Austria*.

Granic, I., Lobel, A., & Engels, R. (2013). The benefits of playing video games. *American Psychologist*, *69*(1), 66–78. doi: 10.1037/a0034857

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS.Med*, *2*, e124. Retrieved 7/14/09 from: http://www.plosmedicine.org/article/info. doi: 10.1371/journal.pmed.0020124

Ioannidis, J. A., & Trikalinos, T. A. (2007). The appropriateness of asymmetry tests for publication bias in meta-analyses: A large survey. *Canadian Medical Association Journal*, *176*(8), 1091–1096. doi: 10.1503/cmaj.060410

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. doi: 10.1177/0956797611430953

Kepes, S., & McDaniel, M. A. (2013). How trustworthy is the scientific literature in industrial and organizational psychology? *Industrial and Organizational Psychology: Perspectives On Science and Practice*, *6*(3), 252–268. doi: 10.1111/iops.12045

LeBel, E. P., & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology*, *15(*4), 371–379. doi: 10.1037/a0025172

Levine, T. R., Asada, K. J., & Carpenter, C. (2009). Sample sizes and effect sizes are negatively correlated in meta-analyses: Evidence and implications of a publication bias against nonsignificant findings. *Communication Monographs*, *76*(3), 286–302. doi: 10.1080/03637750903074685

Levine, T., Weber, R., Park, H., & Hullett, C. (2008). A communication researchers' guide to null hypothesis significance testing and alternatives. *Human Communication Research*, *34*(2), 188–209. doi: 10.1111/j.1468-2958.2008.00318.x

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

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806–834.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., … Reed, G. M. (2001). Psychological testing and psychological assessment. A review of evidence and issues. *American Psychologist*, *56*, 128–165.

National Science Foundation. (2002). Science and technology: Public attitudes and public understanding. Retrieved from: http://www.nsf.gov/statistics/seind02/c7/c7s5.htm

Nelson, L., Simonsohn, U., & Simmons, J. (2014). P-Curve fixes publication bias: Obtaining unbiased effect size estimates from published studies alone. *Social Science Research Network*. Retrieved from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2377290

O'Boyle, E. H., Banks, G. C., & Gonzalez-Mulé, E. (2014). The Chrysalis effect: How ugly initial results metamorphosize into beautiful articles. *Journal of Management*, 0149206314527133. doi: 10.1177/0149206314527133

Pashler, H., Coburn, N., & Harris, C. R. (2012). Priming of social distance? Failure to replicate effects on social and food judgments. *PloS ONE*, *7*(8), e42510. doi: 10.1371/journal.pone.0042510

Pratt, T., & Cullen, C. (2000). The empirical status of Gottfredson and Hirschi's general theory of crime: A meta-analysis. *Criminology*, *38*, 931–964.

Richard, F. D., Bond Jr, 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

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

Scargle, J. D. (2000). Publication bias: The "file-drawer" problem in scientific inference. *Journal of Scientific Exploration*, *14*, 91–106. Retrieved from http://www.scientificexploration.org/journal/jse_14_1_scargle.pdf

Schonemann, P. H., & Scargle, J. D. (2008). A generalized publication bias model. *Chinese Journal of Psychology*, *50*, 21–29. Retrieved from http://www.schonemann.de/pdf/91_Schonemann_Scargle.pdf

Schooler, J. (2011). Unpublished results hide the decline effect. *Nature*, *470*(7335), 437.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology. *Psychological Science*, *22*(11), 1359–1366.

Thompson, B. (2001). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. *Journal of Experimental Education*, *70*, 80–93.

Wagenmakers, E., Wetzels, R., Borsboom, D., & van der Maas, H. 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

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychological journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.Illustration of parameters affecting power. Power for population effect size $\delta = 0.7$ and sample sizes of $n_1 = n_2 = 20$. Effect of greater population effect size ($\delta = 1$) on power.