ANNUAL
REVIEWS

*Annual Review of Psychology*

# Optimizing Research Output: How Can Psychological Research Methods Be Improved?

Jeff Miller[1] and Rolf Ulrich[2]

[1] Department of Psychology, University of Otago, Dunedin 9016, New Zealand;
email: miller@psy.otago.ac.nz

[2] Department of Psychology, University of Tübingen, Tübingen 72074, Germany

Annu. Rev. Psychol. 2022. 73:691–718

First published as a Review in Advance on October 6, 2021

The *Annual Review of Psychology* is online at psych.annualreviews.org

https://doi.org/10.1146/annurev-psych-020821-094927

## Keywords

research efficiency, hypothesis testing, replication, alpha level, power, false positives

## Abstract

Recent evidence suggests that research practices in psychology and many other disciplines are far less effective than previously assumed, which has led to what has been called a "crisis of confidence" in psychological research (e.g., Pashler & Wagenmakers 2012). In response to the perceived crisis, standard research practices have come under intense scrutiny, and various changes have been suggested to improve them. The burgeoning field of metascience seeks to use standard quantitative data-gathering and modeling techniques to understand the reasons for inefficiency, to assess the likely effects of suggested changes, and ultimately to tell psychologists how to do better science. We review the pros and cons of suggested changes, highlighting the many complex research trade-offs that must be addressed to identify better methods.

## Contents

## 1. INTRODUCTION

Psychological science has been increasingly in the public eye in recent years—but not in a good way. From the pages of the *New York Times* (e.g., Carey 2015) to those of *Nature* (e.g., Yong 2012), popular reports have suggested that current methods in psychological science are inefficient, untrustworthy, and wasteful of research resources. The main symptom of these problems is the poor replicability of the findings published in supposedly solid journals for scientific psychology. At least within the social and cognitive areas, it appears that fewer than 50% of the published results can be replicated (e.g., Open Sci. Collab. 2015), suggesting that many of the published findings are actually spurious, nonreplicable results. Concerns about low replicability have also been raised in other areas of science that use similar scientific methods, including neuroscience (e.g., Button et al. 2013, Hartshorne & Schachner 2012, Poldrack 2019), medicine and pharmacology (Begley & Ellis 2012, Ioannidis 2005), economics (Brown & Wood 2018), and other social sciences (Bueno de Mesquita et al. 2003, Freese & Peterson 2017) as well as traditional sciences like biology, chemistry, and physics (Baker 2016). Because the trustworthiness of research results is indispensable for scientific progress, it is understandable that this replication crisis not only alarms the scientists themselves (e.g., Pashler & Harris 2012) but also creates a lack of public trust in these fields (e.g., Begley & Ioannidis 2015, Białek 2018, Saltelli & Funtowicz 2017). Moreover, poor replicability raises suspicions that the considerable public resources devoted to research are largely wasted, with some assessments estimating the waste to be as high as 85% (Chalmers & Glasziou 2009). In response, metascientists—that is, researchers who use scientific methods and models to study science itself—are actively trying to identify the flaws in traditional practices that are responsible for poor replicability and to propose ways of improving these practices (e.g., Bero 2018, Ioannidis 2018, Schooler 2019). Collectively, the investigations and proposals associated with these efforts are sometimes referred to as the evidentiary value movement (EVM; e.g., Finkel et al. 2015).

This article provides a (mostly) nontechnical overview of this ongoing metascientific work. We review many of the problems suggested to be responsible for suboptimal scientific progress—not only in psychology but also in many other scientific fields—and consider the changes in scientific practice that have been proposed to address these problems. The issues are complex, and there is considerable debate about many of the proposed changes. Indeed, proposals range from minor extensions (e.g., Baumeister 2016), through significant modifications (e.g., Cumming 2014, Lakens & Evers 2014), to a complete overhaul (e.g., Barrett 2020, Loftus 1996, Wagenmakers et al. 2011) of current practices, and many proposals are actually incompatible with one another because of trade-offs inherent in the research process. We argue that these trade-offs doom any piecemeal approach to improving scientific methods, but that progress can be made by using quantitative models to assess the mutual influences of various research practices. Although fully optimal research practices can probably never be identified because all of the relevant variables can never be known (Simon 1947), quantitative models can increase our understanding of the trade-offs and guide the search for better practices and enhanced scientific output.

## 2. WHY IS REPLICABILITY SO POOR?

Improving scientific methods requires a clear understanding of what causes poor replicability in the first place. In some cases, failure to replicate may not reflect any problem with the original study. For example, although it is natural to be suspicious of a finding that is not successfully replicated, it is logically possible that the original study was fine and that there were problems with the replication study; perhaps the latter was too small or did not accurately recreate the original conditions. Naturally, most replicators have gone to great lengths to avoid such problems (e.g., Open Sci. Collab. 2015), but they may not have been completely successful in their efforts (e.g., Gilbert et al. 2016). Another possibility is that a nonreplicated effect was real when it was originally found, but circumstances had changed in important ways by the time the replication was attempted (McShane & Böckenholt 2014, Stroebe & Strack 2014). It is easy to imagine, for example, that certain findings in social psychology would not replicate during or immediately after the global COVID-19 pandemic affected people's attitudes toward institutions and groups (e.g., Hartman et al. 2021, Sibley et al. 2020). However, it is widely believed that procedural problems with replication attempts are responsible for only a small percentage of replication failures (e.g., Olsson-Collentine et al. 2020, Sherman & Pashler 2019, Van Bavel et al. 2016).

There is growing consensus that the main reason for low replication rates is that many original published findings are spurious. As summarized in the sidebar titled Research Outcomes, the individual studies in a research area can be broadly categorized into four types, and the probability of each outcome type can be computed based on the diagram shown in **Figure 1**. Each study can be thought of as looking for a single key effect (e.g., a difference between groups or conditions, an association between variables, etc.). The sought-after effect may be present (H1 is true) or absent (H0, the null hypothesis, is true). These possibilities are not equally likely, however; there is some base rate probability of a true effect $\pi$ that depends on the research area, as discussed below. After the data are collected and statistical testing is done, the study is regarded as having a positive result if the sought-after effect is statistically present and a negative result if it is not. Because of the random error inherent in statistical research, however, the study's results may or may not reflect the actual true state of the world. Thus, a study's positive result may be a true positive (TP), accurately revealing that the sought-after effect is actually present, or it may be a false positive (FP), with random error spuriously suggesting that an effect is present when actually it is not. Similarly, a study's negative result may be a true negative (TN), accurately revealing that the sought-after effect is absent, or it may be a false negative (FN), with random error spuriously concealing an effect that is actually present.

## RESEARCH OUTCOMES

| Label and abbreviation | Meaning |
|---|---|
| True positive (TP) | Correctly concluding that a sought-after effect is present |
| False positive (FP) | Incorrectly concluding that a sought-after effect is present (Type 1 error) |
| True negative (TN) | Correctly concluding that a sought-after effect is absent |
| False negative (FN) | Incorrectly concluding that a sought-after effect is absent (Type 2 error) |

Literature surveys indicate that almost all studies published in psychology and many other fields report positive findings (i.e., either TPs or FPs; e.g., Bakker et al. 2012, Fanelli 2012, Rosenthal 1979). In essence, then, current evidence of low replication rates tends to suggest that many published findings are FPs rather than TPs. Thus, metascientists investigating low replication rates have focused on the numerous possible causes of FPs.[1]

Some FPs may reflect outright data falsification or other scientific fraud (e.g., Stroebe et al. 2012), but evidence suggests that these are rare (e.g., Fanelli et al. 2015, Gross 2016). Other FPs may reflect honest researcher errors at many points during the research process, from setting up a

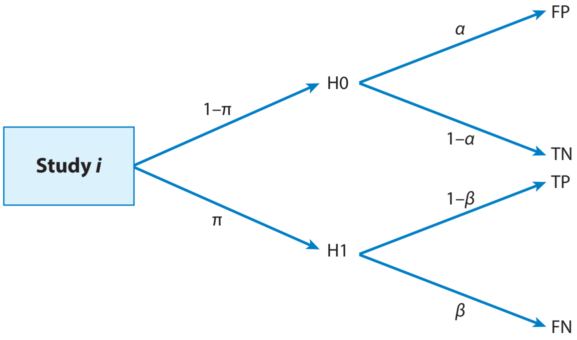


**Figure 1**

A model of the research process. In each study $i$, the null hypothesis is true (H0) or an effect is present (H1), and the base rate probability $\pi$ represents the probability that a sought-after effect is actually present across all studies within a given research area. Based on the observed data, each study's null hypothesis is rejected (positive result, P) or not (negative result, N), leading to a true (T) or false (F) conclusion. The probability of a positive result is $\alpha$ when H0 is true and $1 - \beta$ (i.e., statistical power) when the effect is present.

[1] The bias toward publication of positive results is problematic in its own right because it distorts the overall evidence patterns portrayed in the literature (e.g., Rosenthal 1979, Ulrich & Miller 2018, Ulrich et al. 2018), but it seems deeply rooted in academic incentives and has been around for many years (e.g., Sterling 1959). This publication bias is sometimes cited as one of the causes of high FP rates (e.g., Zwaan et al. 2018), but it does not affect the rate of FPs as that is often defined (see the sidebar titled Outcome Probabilities and Rate of False Positives). As a numerical example, suppose that researchers conduct studies with $\alpha = 0.05$ and power $1 - \beta = 0.8$ to test for true effects that are present with a base rate of $\pi = 0.1$. The overall probabilities of TPs and FPs are $\pi \times (1 - \beta) = 0.08$ and $(1 - \pi) \times \alpha = 0.045$, respectively, so the overall probability of an FP is $0.045/(0.045 + 0.08) = 0.36$. Now suppose that there is a strong publication bias such that the probability of a positive result being published is 0.9. In the published literature the probability of an FP will be $0.045 \times 0.9/(0.045 \times 0.9 + 0.08 \times 0.9) = 0.36$. Thus, publication bias per se does not affect the proportion of FPs relative to TPs in the literature, even though it does affect the proportion of FPs relative to FNs or TNs.

### OUTCOME PROBABILITIES AND RATE OF FALSE POSITIVES

$$\Pr(FP) = (1 - \pi) \cdot \alpha$$

$$\Pr(TP) = \pi \cdot (1 - \beta)$$

$$\Pr(TN) = (1 - \pi) \cdot (1 - \alpha)$$

$$\Pr(FN) = \pi \cdot \beta$$

$$R_{\text{FP}} = \frac{\Pr(FP)}{\Pr(FP) + \Pr(TP)}$$

where

$\alpha =$ probability of concluding a sought-after effect is present when it is not (Type 1 error rate)

$1 - \beta =$ probability of concluding a sought-after effect is present when it is (power)

$\pi =$ probability that a sought-after effect is truly present (base rate)

study's conditions, through recording and analyzing data, all the way up to interpreting the results (e.g., Leek & Peng 2015). For example, errors may arise from researchers' misunderstanding and misuse of the $p$ values produced by standard null hypothesis significance testing (NHST; e.g., Etz & Vandekerckhove 2016, McShane et al. 2019, Nuzzo 2014). To combat statistical errors, many have suggested improved statistical training and possibly the adoption of revised—and hopefully easier to understand—statistical techniques (e.g., Asendorpf et al. 2013, Barrett 2020). There is not much evidence about the extent to which these changes would be effective, but on the whole there has been little dispute about the advisability of improved statistical training.

## 3. STATISTICAL CAUSES OF FALSE POSITIVES

Random variability inevitably produces some FPs by chance, and three statistical parameters determine the frequency of such FPs. The effects of these parameters are illustrated in **Figure 2**.

First, the rate of FPs ($R_{\text{FP}}$) is smaller when researchers use a smaller value of $\alpha$. Within the standard NHST framework, using a smaller $\alpha$ effectively establishes a more stringent all-or-none criterion for concluding that an effect is present. Analogous and somewhat arbitrary cutoffs are inherent in any statistical method for making dichotomous decisions (e.g., to publish or not) about whether an effect was found (McElreath & Smaldino 2015). As shown **Figure 2**, $R_{\text{FP}}$ is always smaller with a smaller $\alpha$, regardless of the power and base rate.

Second, the rate of FPs decreases as power increases. In particular, $R_{\text{FP}}$ increases especially rapidly when power drops below about 0.5, which it commonly does with the small samples and noisy measurement techniques common in psychology and neuroscience (e.g., Button et al. 2013, Maxwell 2004).

The third statistical factor influencing the rate of FPs is the base rate of true effects $\pi$, which is the proportion of studies in a research area seeking effects that are actually present rather than absent. For example, working backwards from observed study replication rates, Wilson & Wixted (2018) estimated that 10% of studies in social psychology test for effects that are actually present, whereas 27% of studies in cognitive psychology do so. At typical values of $\alpha$ and power, these base rates would produce FP rates of approximately 50% and 20%, respectively (see **Figure 2**).

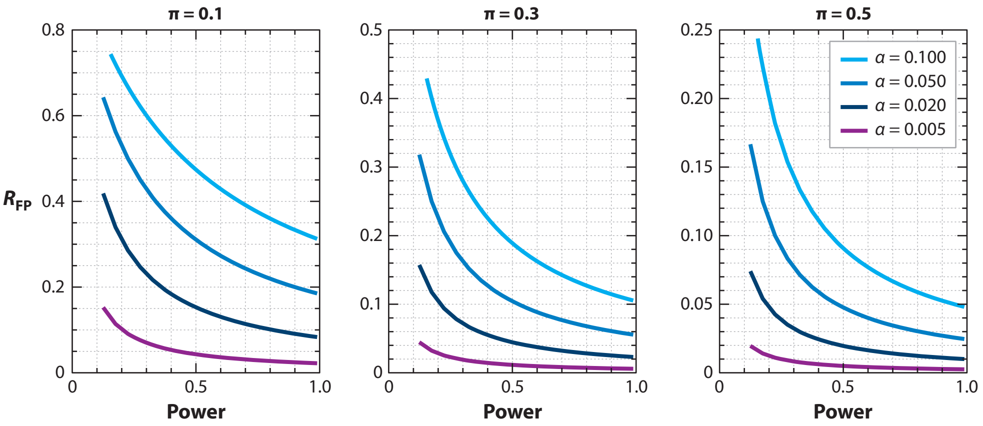


**Figure 2**

Rate of false positives ($R_{\mathrm{FP}}$) as a function of Type 1 error rate $\alpha$, power $1 - \beta$, and the base rate of true effects $\pi$.

Unlike $\alpha$ and power, base rate is a characteristic not of a particular study but of the study's broad research area. Base rate tends to be high in areas in which researchers have a solid empirical and theoretical background that allows them to predict accurately what effects should be present and how those effects can be measured (Smaldino & McElreath 2016, Wilson & Wixted 2018). For example, the base rate is presumably high for researchers investigating the effects of aging on cognition, because established knowledge tells us that many such effects are truly present for biological reasons. A research area with a high base rate might be considered low risk, in that well-designed studies would have a good chance of obtaining positive results. In contrast, the base rate is low in areas where most studies test for effects that are not actually present. These are areas with less-developed empirical and theoretical backgrounds (Smaldino & McElreath 2016, Wilson & Wixted 2018), where researchers must predict effects on intuitive or anecdotal grounds. Studies in low base rate areas would clearly be higher in risk (i.e., of searching for effects where none are present). At the extreme, the base rate in a research area might even be zero if there were no true effects at all in that area—an example might be research into extrasensory perception. With the currently available information, empirical base rates can only be estimated somewhat indirectly, but estimates in most areas of psychology suggest that base rates are 20% or less (e.g., Dreber et al. 2015, McElreath & Smaldino 2015, Miller & Ulrich 2016, Wilson & Wixted 2018).

As shown in **Figure 2**, the base rate $\pi$ has a very strong effect on the rate of FPs, with $R_{\mathrm{FP}}$ increasing systematically as $\pi$ decreases. This happens because, as the base rate decreases, fewer studies have the opportunity to produce TPs (because fewer studies are seeking effects that are actually present). Thus, more and more of the positive results are mistakes (FPs). Indeed, with typical values of $\alpha$ and power, more than half of statistically significant findings will be FPs (i.e., $R_{\mathrm{FP}} > 0.5$) when the base rate is $\pi \leq 0.1$. This happens for purely statistical reasons, even if researchers use only the most appropriate scientific methods. Obviously, replicability would then be disappointingly low (see Ulrich & Miller 2020, figure 2).

## 4. SUGGESTIONS FOR REDUCING FALSE POSITIVES

The primary suggestions for reducing the rate of FPs—and thereby increasing replicability—emerge directly from the effects of $\alpha$ level, power, and base rate shown in **Figure 2**. Although it is clear that the rate of FPs can be reduced by lowering $\alpha$, increasing power, and increasing base rate, problems associated with each of these options have led to debate about their feasibility and advisability.

## 4.1. Reduce the α Level

One certain way to reduce the rate of FPs is to reduce the $\alpha$ level (i.e., Type 1 error rate). Lowering $\alpha$ means that stronger evidence of an effect is needed before concluding that a positive result has been found (i.e., before rejecting H0), and this reduces FPs. Hence, it is not surprising that many authors (e.g., Aczel et al. 2017, Benjamin et al. 2018, Colhoun et al. 2003, Colquhoun 2014, Johnson 2013, Schimmack 2012) have advocated using $\alpha$ values below the traditional $\alpha = 0.05$ level suggested by Fisher (1925). For example, Benjamin et al. (2018) argued that researchers should reduce the criterion for statistical significance to $\alpha = 0.005$, at least for initial demonstrations of novel effects. They cited evidence that the base rate of true effects in psychology is only about 9%, and in this case $\alpha = 0.05$ produces an FP rate of at least 33% (considerably more if power is low).[2] If power is maintained at a high level, Benjamin et al. (2018, p. 7) argued that using $\alpha = 0.005$ "would reduce the false positive rate to levels we judge to be reasonable." Data yielding observed $p$ values in the range of $0.005 < p < 0.05$ could still be published, according to their proposal, but would be considered only suggestive.

## 4.2. Eliminate Questionable Research Practices

Many have argued that FPs should also be reduced by eliminating a broad category of questionable research practices (QRPs) that tend to inflate a researcher's actual Type 1 error rate above the claimed nominal $\alpha$ level (e.g., John et al. 2012, Simmons et al. 2011). These practices can be used to coax statistical significance out of initially nonsignificant results—a practice known as $p$-hacking (e.g., Head et al. 2015, Simonsohn et al. 2014), wherein a researcher capitalizes on the degrees of freedom associated with the many ways data can be analyzed (Simmons et al. 2011). For example, researchers might (*a*) perform multiple similar studies or measure multiple similar dependent variables to find one that produces a significant result; (*b*) perform multiple analyses with different statistical methods, covariates, or outlier exclusion criteria to find one that produces a significant result; or (*c*) collect further data if the initial results are nonsignificant, in hopes of obtaining a significant result with a larger sample (e.g., John et al. 2012). Such practices are tempting because they increase researchers' chances of finding statistically significant—and hence publishable—findings, but they have been heavily criticized because they can increase the Type 1 error rate far beyond the nominal $\alpha$ level (Simmons et al. 2011), thus increasing the rate of FPs just like an increase in the nominal $\alpha$ itself would. For example, a researcher with a nominal $\alpha = 0.05$ may incorrectly reject more than 25% of true null hypotheses by repeatedly checking the data among successive groups of participants and stopping if a significant result is found (e.g., Armitage et al. 1969, Strube 2006). QRPs like this are sometimes thought to be primarily responsible for the high rates of FPs that are associated with low replicability (e.g., Schimmack 2020, Simmons et al. 2011); however, the true effects of QRPs on replicability may be smaller than is generally assumed because QRPs also increase power by making it easier to reject null hypotheses that are false as well as those that are true (e.g., Ulrich & Miller 2020).

[2] With $\alpha = 0.05$ and standard two-tailed testing, 33% is the correct estimate of the FP rate only if a significant effect in either direction would be considered a positive result. If researchers have a priori expectations about the direction of the effect and would only report a positive result in the expected direction, this minimum FP rate is only 18%, as is shown in the $\alpha = 0.05$ line in **Figure 2**. Benjamin et al.'s (2018) 33% estimate corresponds to $\alpha = 0.10$ in **Figure 2**, for which there would be only a 5% Type 1 error rate in the predicted direction.

To combat QRPs and thereby hold the actual Type 1 error rate at the nominal $\alpha$ level, many have advocated increasing the transparency of research by making data and research materials publicly available (e.g., Nosek et al. 2015), which would presumably also help with the detection of honest errors and fraud (Gross 2016). Researchers could also be asked to preregister their plans for data collection and analysis, which would prevent many of the QRPs that inflate Type 1 error rates (Nosek et al. 2018). In principle, decisions about study publication could be based entirely on peer review of such plans, thereby removing even the incentive for researchers to chase after significant results (Chambers 2020). Some FPs would still occur in preregistered studies, however, for purely statistical reasons.

### 4.3. Increase Power

Another prominent recommendation for reducing the rate of FPs is to increase study power by increasing sample sizes (e.g., Asendorpf et al. 2013, Button & Munafò 2017, Button et al. 2013). Actually, there was considerable dismay over the low power of studies in psychology even before the issue of FPs gained its current prominence. Analyses of the literature in various areas suggested that actual power levels were typically below 0.20 for small effects and only about 0.50 for medium effects (e.g., Button et al. 2013, Clark-Carter 1997, Cohen 1962, Rossi 1990, Sedlmeier & Gigerenzer 1989). Cohen (1988, p. 56) recommended that researchers using $\alpha = 0.05$ should aim for a power level of at least 0.8 (i.e., $\beta = 4\alpha$), because he felt that FPs are approximately four times as bad as FNs. Following that rationale, researchers using smaller $\alpha$ levels would require even higher power (e.g., $\alpha = 0.005$ would correspond to power $1 - 4 \times 0.005 = 0.98$). Unfortunately, in practice it is very difficult to say what sample sizes are needed to attain specific target power levels, because true effect sizes are unknown. Moreover, sample sizes can be too large as well as too small, because very large samples can waste research resources and provide high power to detect effects that are too small to be of practical interest (Lenth 2001).

### 4.4. Increase the Base Rate

Although the base rate of true effects in a research area has a strong influence on the rate of FPs (e.g., Ioannidis 2005), there have been no specific suggestions about how researchers could increase their base rates in order to reduce FPs (for discussion, see McElreath & Smaldino 2015, Wilson & Wixted 2018). This is partly because researchers have little direct control over their base rates, which are determined mainly by the level of prior knowledge and theoretical development in their research areas, as discussed above. If researchers did want to consider base rates, they would have to do so when choosing a research area—or, at least, deciding what effect to seek—rather than at the later point of choosing the $\alpha$ level and sample size (i.e., power).

## 5. OBJECTIONS TO PROPOSED CHANGES

Although the above suggestions for reducing FPs all seem sensible, there are several reasonable objections to them. Despite the widespread concern about high FP rates, such objections cast doubt on the idea that minimizing FP rates would actually maximize research efficiency under any reasonable definition of that term.

First, the objection to reducing $\alpha$ is that it reduces power unless the samples sizes are substantially increased (e.g., Lakens et al. 2018), as illustrated in **Figure 3**. For any given sample size and effect size, demanding stronger evidence before declaring a positive result (i.e., reducing $\alpha$) makes it harder to obtain TPs as well as FPs. Thus, reducing $\alpha$ increases the chances of missing true effects (i.e., it increases the rate of FNs), which could be even costlier than finding FPs (e.g.,

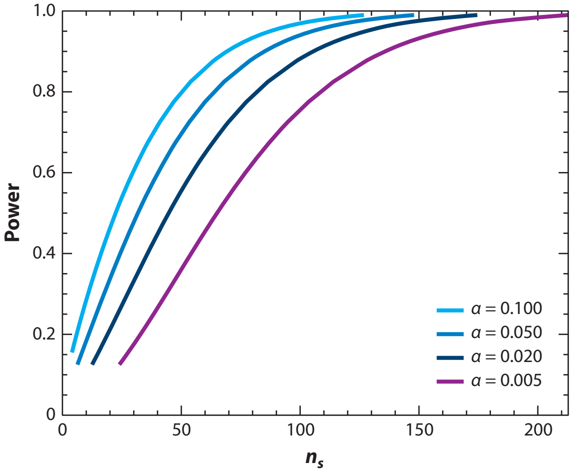


**Figure 3**

Power to detect an effect in the predicted direction as a function of the researcher's choices of sample size $n_s$ and $\alpha$ level with two-tailed testing. The computations are for a two-sample $t$-test with effect size $d = 0.5$ and $n_s$ in each sample. Plots for correlations, one-sample $t$-tests, and other $d$ values look almost identical except for the ranges of the horizontal axis.

Amrhein et al. 2019, Fiedler & Schott 2017, Fiedler et al. 2012). In some scenarios, if researchers use the same $\alpha$ level for all studies, decreasing $\alpha$ to reduce FPs can even have the unexpected effect of lowering overall replicability because of this power loss (Miller & Ulrich 2019).[3]

Second, although to our knowledge there have been no published objections to the idea that QRPs should be eliminated to hold the actual Type 1 error rate at the nominal $\alpha$ level, even this suggestion comes with a potential cost. QRPs increase power by providing multiple opportunities to reject false null hypotheses as well as true ones—a phenomenon we have previously called power inflation (Ulrich & Miller 2020). Thus, eliminating QRPs decreases power for exactly the same reason that reducing $\alpha$ does (see also Witt 2019). This is not a defense of QRPs, of course, but it does illustrate the difficulty of identifying optimal research strategies from qualitative arguments.

Third, the objection to increasing power is that, unless researchers can somehow increase effect size, this must be done by increasing sample sizes in terms of the number of participants, the number of observations per participant, or both (e.g., Baker et al. 2021), which increases research costs. Because researchers have limited resources, increasing the sample size per study implies that fewer studies can be run. This trade-off forces researchers to choose between conducting a few large, high-powered studies or a greater number of smaller, low-powered ones (e.g., Lakens & Evers 2014, Lakens et al. 2018).

[3] For example, suppose that each of two researchers tests for a true effect of $d = 0.5$ with a two-sample $t$-test, with both researchers using $N = 60$ per group, and suppose that the base rate of such true effects is $\pi = 0.3$. One researcher uses a one-tailed $\alpha = 0.05$ and consequently has power $1 - \beta = 0.86$. For this researcher the probabilities of initial TPs and FPs are $\pi \times 0.86 = 0.26$ and $(1 - \pi) \times \alpha = 0.035$, respectively. The probabilities of replicated TPs and FPs are $\pi \times 0.86^2 = 0.22$ and $(1 - \pi) \times \alpha^2 = 0.002$, so the researcher's overall replication rate is $(0.22 + 0.002)/(0.26 + 0.035) = 0.76$. The other researcher uses the smaller one-tailed $\alpha = 0.005$ and thus has less power, $1 - \beta = 0.55$. This researcher's probabilities of initial TPs and FPs are 0.16 and 0.0035, respectively, those of replicated TPs and FPs are 0.09 and 0.0002, and so the researcher's overall replication rate is only 0.54.

The trade-off between power and number of studies is further highlighted when also considering the recommendation to reduce $\alpha$, which—as already mentioned—has the unwanted consequence of reducing power. In principle, researchers could simultaneously reduce $\alpha$ and increase power by using very large samples—thus obtaining more accurate results (i.e., more TPs and fewer FPs). The price to be paid for such accuracy, in terms of a reduced number of studies, can be estimated from **Figure 3**. Consider, for example, a researcher using a two-sample $t$-test design with $\alpha = 0.05$ and testing two groups of 20 to achieve a power level of approximately $1 - \beta = 0.33$ for detecting effects of $d = 0.5$—a power value that seems fairly typical for current psychological studies (e.g., Stanley et al. 2018). If this researcher embraced the recommendations to reduce FPs by reducing $\alpha$ to 0.005 and increasing power to 80%, they would need samples of 109 per group and could then presumably only run approximately one-fifth as many studies. Moreover, if the researcher used $\alpha = 0.005$ and strove for 98% power, in accordance with Cohen's (1988) recommendation of $\beta = 4\alpha$, then samples of 191 per group would be required and only one-tenth as many studies could be run. Benjamin et al. (2018, p. 8) offered the view that "efficiency gains would far outweigh losses" as a result of such $\alpha$ decreases and power increases, but they provided no quantitative analysis in support of that view. Supporting that view would seem to require a detailed model of the research process and a clear definition of efficiency, as considered in the next section.

Finally, there are possible objections to increasing base rates, depending on exactly how that is done. In particular, it would probably not be helpful for researchers to concentrate on looking for effects that are only minor variants of previously replicated effects, even though these would presumably be present with a high base rate. It seems unlikely that science would progress very rapidly if all researchers adopted such a strategy, even though the rate of FPs would be minimal. On the contrary, unexpected findings can lead to scientific breakthroughs (e.g., Dunbar & Fugelsang 2005, Kuhn 1962, Roberts 1989), but few of these would emerge if researchers conducted only low-risk (i.e., high–base rate) studies. In addition, Popper [2002 (1963)] has argued that theories should be tested by checking their predictions that seem *least* likely to be true, because confirmations of such predictions are particularly informative. Clearly, such a research strategy would imply high-risk studies. Moreover, low-risk studies of minor variants would also tend to have little news value, making them harder to publish. Although few journals explicitly consider base rates, anecdotal evidence suggests that many prefer to publish unexpected new findings (e.g., Smaldino & McElreath 2016), and this encourages researchers to conduct the low–base rate studies that inevitably produce a large rate of FPs. This is not to say that increases in base rate are necessarily bad, of course. They may be produced by theoretical advances, which would clearly signify scientific progress.

## 6. MODELS FOR RESEARCH EFFICIENCY

Several different quantitative models have been used to try to optimize some aspects of the research process. Most commonly, these are situated entirely within a statistical framework and seek to investigate purely statistical questions, such as the sample size needed to achieve a certain level of power with a certain $\alpha$ level (e.g., Gillett 1994) or to estimate a parameter with a given reliability (e.g., Karrandinos 1976). Although they are well specified and precise, these simple models generally do not address efficiency-based questions, such as what power and $\alpha$ levels should be chosen in the first place (but see Lenth 2001).

Some more complicated models do address questions of efficiency. In the area of medical research, for example, models attempt to relate the economic costs of conducting research to the health-benefit outcomes that they might produce (e.g., Baker & Heidenberger 1989). In some cases, the models also attempt to quantify research results economically (e.g., the profitability of a new drug) in order to choose optimal research parameters (e.g., Berry & Ho 1988, Detsky 1985, Miller 1996, Mosteller & Weinstein 1985).

Because of the complex trade-offs involving $\alpha$, effect size, power, and base rate, systems-level quantitative models are needed to maximize research productivity. At a minimum, these models must provide frameworks for integrating these statistical parameters with practical constraints. One very important practical constraint is that there is a fixed total pool of research resources—conceptualized at the level of an individual researcher or a group of researchers working in a given area—that must be divided across studies. For example, the resources might be conceived of as the total number of participants $N$ that could be tested across all studies, in which case it would be possible to do $N/n_s = k$ studies with $n_s$ participants each. To be prescriptive, the models must also specify a criterion for comparing the effectiveness of different parameter combinations. For example, should the goal be to find the combinations of parameters that maximize replicability, or would it be more appropriate to use some other measure of scientific effectiveness?

Such models are useful in providing researchers with guidance about how to maximize research efficiency. As discussed earlier, for example, the $\alpha$ level could easily be altered if some other choice could be shown to be better than the current standard $\alpha = 0.05$ level. Likewise, sample sizes could also rather easily be increased or decreased relative to current levels if that could be shown to improve efficiency. The values of other parameters such as base rate and effect size are not so easily adjusted by researchers, but models can be used to gauge the importance of trying to change them. In addition, these models could illustrate how researchers' choices of optimal $\alpha$ and sample size values should depend on these more-difficult-to-control parameters even if the latter parameters cannot themselves be changed.

We are under no illusion that such models will enable researchers to make all decisions optimally. For one thing, researchers will rarely, if ever, have knowledge of "all the aspects of value, knowledge, and behavior that would be relevant" (Simon 1947, p. 108). Furthermore, the very act of defining a metric to quantify research payoff encourages researchers to "game" their way to better evaluations, with perverse consequences for the research being assessed (e.g., Edwards & Roy 2017). For example, quantifying research output as the number of publications encourages researchers to publish as many small studies as possible, whereas the interests of science might be better served by a few larger, more integrated reports (e.g., Sternberg & Sternberg 2010). Nonetheless, models for research payoffs can be helpful by clarifying the choices that researchers must make and by illustrating the quantitative consequences of these choices. A number of such models have been suggested.

## 6.1. Minimizing False Positives and Maximizing Replicability

Given the general agreement that current replication rates are too low, one proposal for improving research practices is to minimize FPs (Finkel et al. 2015). The ideal of 0% FPs and 100% replicability is of course unattainable in psychology, because there are large individual-to-individual and moment-to-moment differences that effectively constitute statistical noise (e.g., Hamann & Canli 2004, Williams et al. 2008). Nonetheless, researchers could in principle define research effectiveness in terms of the goal of minimizing the rate of FPs, which would in turn maximize replicability.

Based on the analyses in the preceding sections, it is clear how to minimize the rate of FPs: by using very small $\alpha$ levels, conducting studies with very high power, and working in areas with very high base rates. Because of resource limitations, the joint constraints of low $\alpha$ and high power imply that researchers should only conduct a few studies, each with a gigantic sample to provide high power despite the very small $\alpha$. Thus, the picture of ideal research emerging from the goal of minimizing FPs is a very cautious one in which researchers conduct a few large studies of effects that are almost sure to be present (i.e., $\pi \approx 1$). Even without delving deeply into the philosophy of science, this picture of optimal science is intuitively unattractive. As discussed above, it is debatable whether the FP-reducing benefits of low $\alpha$ and high power fully justify the required sample size increases and consequent reduction in total number of studies (see **Figure 3**), and it seems intuitively that low rather than high base rate studies would sometimes produce the most dramatic scientific advances.

## 6.2. Maximizing Benefit to the Researcher

Several theorists have addressed the research process from the point of view of individual researchers trying to advance their academic careers (e.g., Bakker et al. 2012, Smaldino & McElreath 2016). The real-world incentives for such researchers are clear: Career advancement requires publications, and publications nearly always require statistically significant results. Pragmatic researchers might reasonably ask, then, what combination of parameters would maximize their chances of academic success. In terms of **Figure 3**, for example, is it better to run many studies with small samples and low power, to run just a few studies with large samples and high power, or to find some more effective intermediate sample size?

Bakker et al. (2012) investigated this issue from the perspective of a researcher trying to obtain at least one statistically significant result when testing a fixed total number of participants, which would be a good strategy for career advancement if journals only published positive results and there were little or no career cost for published FPs. Specifically, they conducted simulations of researchers using $\alpha = 0.05$ and using the total participant pool across either one large study or five smaller ones. In addition, their simulated researchers did or did not use QRPs. Bakker et al. (2012) found that the probability of obtaining at least one significant result was much higher for researchers conducting five small studies and using QRPs than for researchers conducting a single large study and not using QRPs. Unfortunately, the strategy that maximized the probability of a significant result also produced the highest rate of FPs. Thus, they concluded that "strategic behaviors by researchers can lead literatures astray" under the assumed incentives (Bakker et al. 2012, p. 552).

**Figure 4** shows the results of an analysis extending the results of Bakker et al. (2012) in two ways. First, it examines the optimal strategy for researchers who specifically want to find TPs rather than just any significant results (i.e., TPs or FPs). Second, it considers the expected number of TPs when testing a fixed total number of participants across samples of various sizes, rather than the probability of getting at least one significant result.

As shown in **Figure 4**, the expected number of TPs depends greatly on the effect size $d$ and the base rate $\pi$, which vary across panels and are presumably in large part out of the researcher's control. Interestingly, however, the optimal strategy for the researcher depends little on the effect size or base rate. In all panels, the expected number of TPs is maximized by using a fairly low power level and a fairly large $\alpha$, reinforcing Bakker et al.'s (2012) conclusion that the most effective strategy for the researcher may involve parameter choices that lead to an overall high rate of FPs for the field as a whole (i.e., large $\alpha$ and low power).

Smaldino & McElreath (2016) extended the analysis of Bakker et al. (2012) by modeling the consequences of traditional publication incentives from an evolutionary perspective. Unlike Bakker et al. (2012), they assumed that individual researchers did not modulate their behavior based on these incentives. Instead, there was simply natural variation in behavior from researcher to researcher or lab to lab, and this variation produced differences in academic success, as measured by the number of novel results found, the number of those results replicated—or not—by others, and the number of positive and negative replications of other researchers' results. Crucially, the

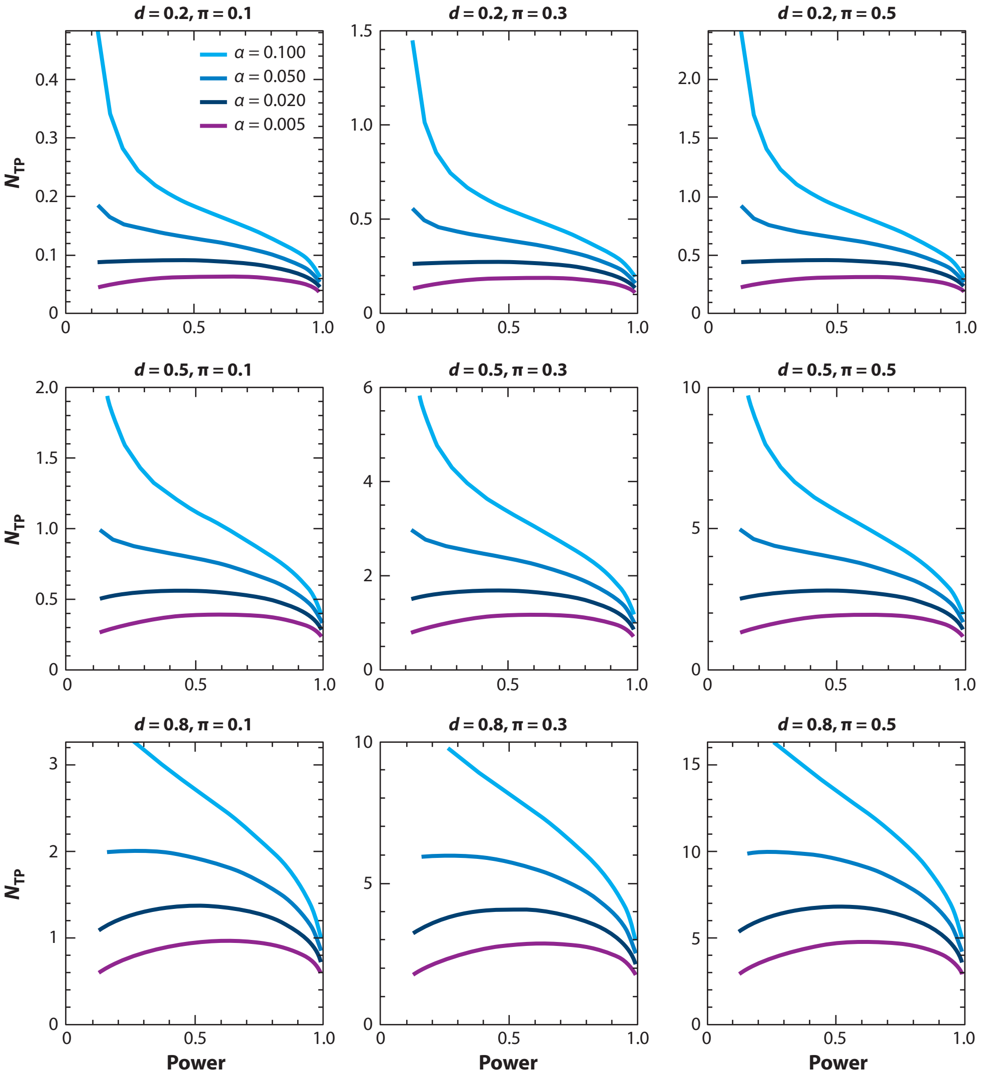


**Figure 4**

Mean number of true positive findings ($N_{\mathrm{TP}}$) per 1,000 tested participants as a function of $\alpha$ level, power, true effect size $d$, and base rate of effects $\pi$. The computations are for a two-sample $t$-test with equal group sizes and two-tailed testing. Note that the vertical axis range differs across panels.

behaviors used by the more successful researchers and labs in each generation were more likely to be passed along to the next generation of researchers. Over time, this system evolved toward smaller sample sizes and higher rates of FPs, even without any strategic individual behavior motivated by self-interest. This suggests that strong incentives to produce TPs may have certain negative consequences for science itself, even when all actors have the best intentions.

## 6.3. Optimizing Replications

With the increased emphasis on the importance of replication, several researchers have considered what researcher strategies would be most efficient from the point of view of researchers carrying out replication studies. For example, because FPs seem to be so common, LeBel et al. (2017b) argued that two high-powered replications should be conducted to confirm each novel positive result as a "true confirmed discovery" (i.e., to accept it as a TP, on the grounds that FPs would virtually never be replicated twice). Using an example with plausible assumptions about effect sizes and base rates, they showed that replication samples would be used more efficiently (i.e., lower average replication sample size per true discovery) if researchers conducted the original studies with large samples and high power rather than small samples and low power. Thus, they concluded that the benefits of increased study power outweigh the costs of a reduced number of studies, at least with respect to the resources used for the replication studies. In fact, any change in researcher strategy that reduced the rate of FPs would also reduce the average replication sample size per true discovery, so their argument also implies that researchers should use extremely small $\alpha$ levels.

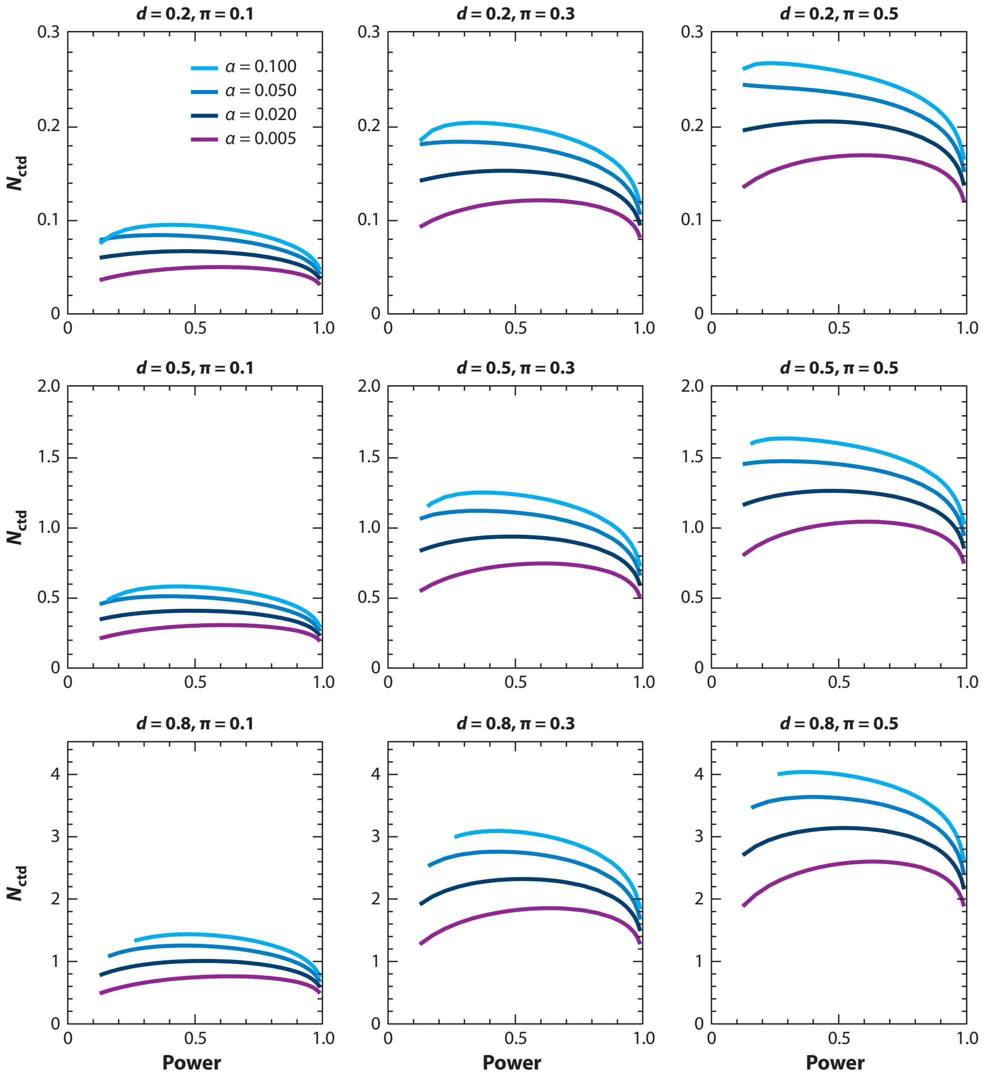
Finkel et al. (2017) questioned the arguments of LeBel et al. (2017b), noting that their analysis only considered the sample sizes used during the replication stage, and not the sample sizes used in the original studies testing for novel effects. Finkel et al.'s simulations showed that the average sample size per replicated TP—totaling the samples across original studies and replications—was minimized when the original studies had intermediate power levels in the range of approximately 0.30–0.70. When original studies were run with lower or higher power levels, the average sample size per replicated TP was actually larger, suggesting that power can be too high for maximum efficiency as well as too low (cf. Lenth 2001).

In a reply to Finkel et al. (2017), LeBel et al. (2017a) defended their previous conclusion that larger original studies lead to greater research efficiency, revising their previous model to consider the sample sizes of original studies as well as replications. Unlike Finkel et al. (2017), however, they assumed that there were separate participant pools for the original and the replication studies, and that the latter pool was too small to allow replication of all originally significant findings. With a limited number of replications, their computations indicated that the average sample size per replicated effect was smaller when the original studies had high (80%) rather than medium (50%) power, with higher original power also producing fewer false discoveries.[4] The total sample size (i.e., total of participant pools for original studies plus replications) was the same for both computations, so this simply means that there were more successful replications when the original studies had higher power, as is to be expected because the rate of FPs is lower in this case.

LeBel et al.'s (2017b) suggestion that novel findings should be replicated twice before they are accepted as confirmed true discoveries is worth considering, but it does not seem appropriate to evaluate it using distinct participant pools for original versus replication studies. If a two-replication rule was adopted, then the fixed total pool of participants should be split flexibly between original and replication studies, in the end using whatever proportions are necessary to carry out two replications of each novel finding. **Figure 5** shows how the expected number of confirmed true discoveries would vary with the $\alpha$ level and power of the original study under that procedure. Clearly, the largest numbers of true discoveries are found with low or intermediate levels of power, and therefore larger samples are not necessarily better with a flexible split of the participant pools between original and replication studies, contrary to LeBel et al.'s (2017a) findings with fixed participant pools.

Lewandowsky & Oberauer (2020) took a different approach to the question of how to use research resources optimally when the research framework involves replications. Specifically, they simulated replication regimes in which either (*a*) all novel positive findings are replicated before

[4] Several bugs were present in the R code used to prepare figure 3 in their original paper, but these do not appear to have affected the results qualitatively.



## Figure 5

Mean number of confirmed true discoveries ($N_{\text{ctd}}$) per 1,000 tested participants as a function of the original researcher's $\alpha$ level, power, true effect size $d$, and base rate of effects $\pi$. Confirmed true discoveries are effects that are replicated twice in the same direction as the original finding by replication studies with $\alpha = 0.05$ and power $1 - \beta = 0.95$ regardless of the $\alpha$ and power levels of the original studies. The 1,000 participants are divided between initial and replication studies as needed so that every initial positive result can be replicated twice. The computations are for a two-sample $t$-test with equal group sizes and two-tailed testing. Note that the vertical axis range differs across panels.

publication (what they called a "private" replication regime), or (*b*) a single novel positive finding is published immediately and replication is reserved for findings judged to be interesting by the scientific community (i.e., a "public" regime). They compared the two regimes with respect to the number of interesting findings each revealed and to the number of studies needed to reveal them. Effect size and sample size were held constant at levels producing 80% power. They found that the public replication regime was more efficient (i.e., it revealed more interesting findings with fewer total studies), essentially because the private regime wasted resources on studies replicating uninteresting results. Thus, they concluded that it could be inefficient to require replication of all novel results before publication. Note that the same conclusion would apply to the policy of requiring a small $\alpha$ level for demonstrating a novel effect, because requiring a replication is analogous to reducing $\alpha$ (e.g., requiring replication effectively reduces $\alpha$ from 0.05 to $0.05^2 = 0.0025$ with one-tailed testing).

## 6.4. Maximizing Overall Payoff

The efficiency models just discussed make it clear that the criterion chosen to measure research efficiency is an important determinant of which research strategy appears to be the best. Yet each of the above models is focused on a single, specific criterion for good research (e.g., replicability) to the exclusion of other properties of high-quality science, such as discovery, consequentiality, cumulativeness, and several kinds of validity (Finkel et al. 2017).

In our own work (Miller & Ulrich 2016), we suggested another efficiency model that could potentially integrate multiple features of good scientific practice into an overall payoff score. The payoff for a true positive, $\mathcal{P}_{\mathrm{TP}}$, is its overall scientific value, which reflects its novelty, importance, generality, validity, and so on. The units of the payoff scale are arbitrary, so it is convenient to assign $\mathcal{P}_{\mathrm{TP}} = 1$ and scale the payoffs of the other outcomes relative to that. Presumably, the payoff associated with an FP outcome, $\mathcal{P}_{\mathrm{FP}}$, is negative, reflecting the harm associated with this incorrect decision. The payoffs for TNs and FNs, $\mathcal{P}_{\mathrm{TN}}$ and $\mathcal{P}_{\mathrm{FN}}$, are also presumably positive and negative, respectively, although the previously mentioned evidence that negative results are rarely published and have little impact on the field suggests that these have negligible payoffs (i.e., $\mathcal{P}_{\mathrm{TN}} \approx 0$ and $\mathcal{P}_{\mathrm{FN}} \approx 0$; see Rosenthal 1979). If negative results were increasingly published, though, the payoffs associated with them would increase in magnitude.

Based on the probability model in **Figure 1** and a set of four assumed outcome payoffs, it is possible to compute the expected total payoff $\mathrm{E}[\mathcal{P}_{\mathrm{T}}]$ across all studies conducted within a research area, as summarized in the sidebar titled Example Payoff Computation. The expected payoff from a single study $i$, $\mathrm{E}[\mathcal{P}_i]$, is the sum of the possible payoffs weighted by their probabilities, $\Pr(TP) \times \mathcal{P}_{\mathrm{TP}} + \Pr(FP) \times \mathcal{P}_{\mathrm{FP}} + \Pr(TN) \times \mathcal{P}_{\mathrm{TN}} + \Pr(FN) \times \mathcal{P}_{\mathrm{FN}}$. Thus, the expected total payoff across all studies is this value multiplied by the number of studies $k$—that is, $\mathrm{E}[\mathcal{P}_{\mathrm{T}}] = k \times \mathrm{E}[\mathcal{P}_i]$ (Miller & Ulrich 2016, equation 7). Of course, the number of studies depends on the total pool of resources available for research and on how the researcher chooses to divide those resources across studies. Thus, the model can be used to compare the relative payoffs of conducting many smaller studies versus fewer larger ones. With several reasonable combinations of effect sizes, base rates, and payoff values, for example, we found that payoff was actually maximized by conducting many small studies, each having power of only approximately 50% or less (Miller & Ulrich 2016). Similarly, we also used the model to compare the payoffs obtained with various $\alpha$ levels, as this is relevant to the debate about whether $\alpha = 0.05$ is too large (Miller & Ulrich 2019). We found that the $\alpha$ level producing the highest payoff depends very strongly on the base rate of true effects and can easily be substantially larger than 0.05 (cf. Michaels 2017).

Unfortunately, determining the payoff values is, as Simon (1947, p. 188) commented in a more general context, "a research task of the first magnitude"; but the model makes it clear that there is no completely rational way to maximize the overall payoff of scientific research in a field without knowing these values. Moreover, payoff magnitudes could well vary across different types of research. For example, the payoff of a true negative $\mathcal{P}_{\mathrm{TN}}$ might be small in an empirically driven search for effective medical treatments, whereas it might be large in a theoretically driven model

### EXAMPLE PAYOFF COMPUTATION

- Consider a researcher who conducts studies with $\alpha = 0.05$ and 50 participants per study. With the available resources, the researcher can test a total of 500 participants, and therefore 10 studies can be conducted.
- Assume that true effects are present with a base rate of 0.2 and that the power to detect these effects is 0.8 with the chosen $\alpha$ and sample size.
- The probabilities of the possible outcomes are thus as follows:

$$\Pr(TP) = 0.2 \times 0.8 = 0.16$$

$$\Pr(FP) = (1 - 0.2) \times \alpha = 0.04$$

$$\Pr(TN) = (1 - 0.2) \times (1 - \alpha) = 0.76$$

$$\Pr(FN) = 0.2 \times (1 - 0.8) = 0.04$$

- Suppose the payoffs for these outcomes are

$$\mathcal{P}_{\text{TP}} = 1, \mathcal{P}_{\text{FP}} = -2, \mathcal{P}_{\text{TN}} = 0.2, \text{and } \mathcal{P}_{\text{FN}} = -0.4.$$

- The expected payoff per study is thus

$$1 \times 0.16 - 2 \times 0.04 + 0.2 \times 0.76 - 0.4 \times 0.04 = 0.216.$$

- The researcher's expected total payoff across all studies is $0.216 \times 10 = 2.16$.

comparison. Fortunately, the model can be useful even when the payoff values are not known exactly, because it allows metascientists to explore the patterns of overall payoffs that would be obtained under various plausible ball-park assumptions about these unknown values. This is particularly true because only relative—not absolute—payoff values are needed to decide among the various researcher options (e.g., $\alpha$ level; see Simon 1947).

## 6.5. Limited Researcher Options and Trade-Offs

When contemplating the variety of proposals for optimizing scientific research (e.g., minimizing FPs or sample size per discovery, or maximizing replicability, number of true discoveries, and payoff), it is easy to lose sight of the fact that researchers have rather limited control over the research process. Effect sizes, base rates, and payoffs are largely determined by the research area being investigated, and there are practical or institutional constraints on the researcher's available resources. Researchers—perhaps in conjunction with journal editors and reviewers—have immediate control only over $\alpha$ levels and sample sizes. Thus, any proposal for optimizing research outcomes must ultimately be considered in terms of its implications for how researchers should set these two parameters. All other summary measures of research efficiency (e.g., power, rate of FPs, probability of replication) are downstream consequences of the $\alpha$ and sample size choices, together with the uncontrollable aspects of the research scenario, and the idea of manipulating these other consequences directly is to some extent a mirage. These limitations on researcher control imply that researchers cannot optimize all outcome measures simultaneously but must contend with trade-offs among the desired outcomes, in addition to the inherent trade-offs among other desirable research characteristics (e.g., generalizability, precision, and realism; see McGrath 1981).

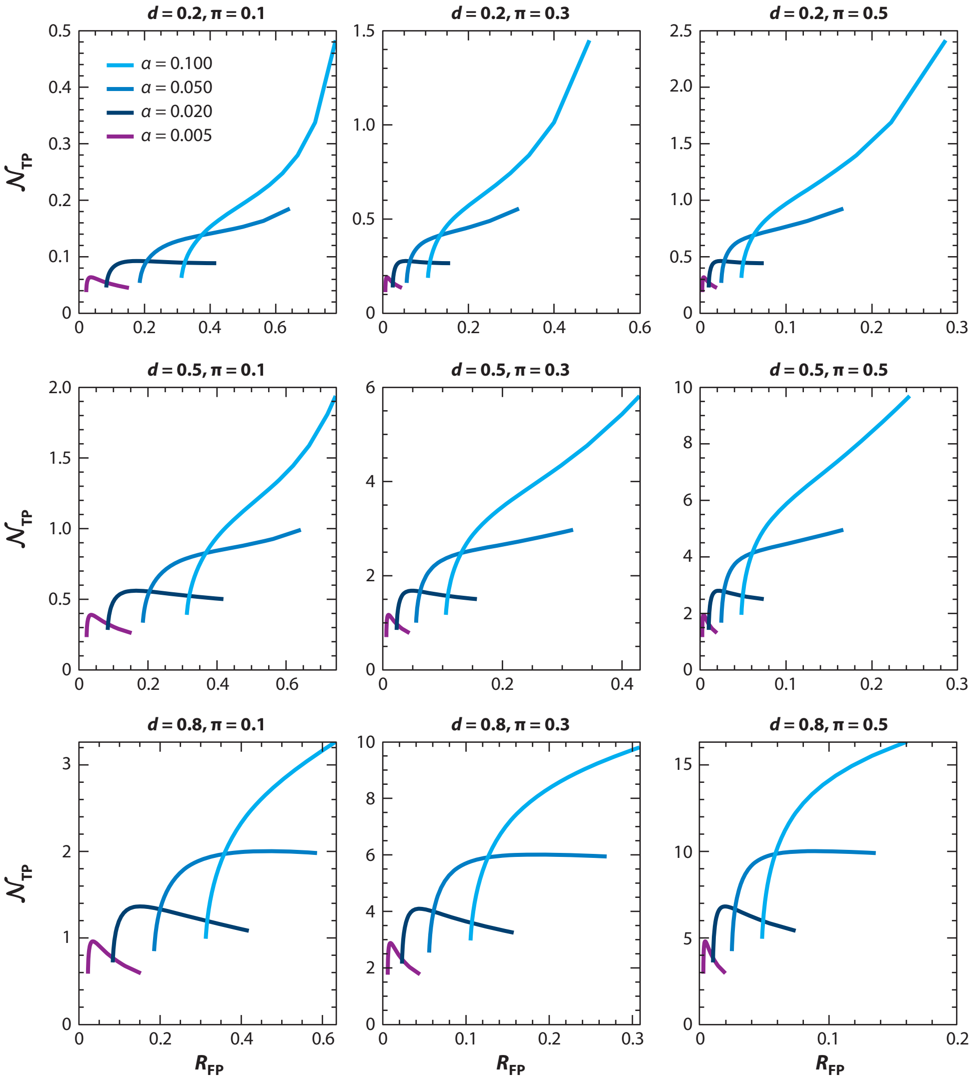


**Figure 6**

Parametric plot of the relationship between number of true positives ($\mathcal{N}_{\mathrm{TP}}$) and rate of false positives ($R_{\mathrm{FP}}$) as a function of the $\alpha$ level, true effect size $d$, and base rate of effects $\pi$. Each curve was traced out by varying the per-study sample size and computing the corresponding expected values of $\mathcal{N}_{\mathrm{TP}}$ and $R_{\mathrm{FP}}$. The computations are for a two-sample $t$-test with two-tailed testing and a total sample size across studies of 1,000. Note that both horizontal and vertical axis ranges differ across panels.

For example, **Figure 6** is a parametric plot illustrating the trade-off between the number of TPs ($\mathcal{N}_{\mathrm{TP}}$) and the rate of FPs ($R_{\mathrm{FP}}$). Obviously, researchers would like to obtain as many TPs as possible while keeping a low rate of FPs, but the positive relationships evident in these panels reveal the incompatibility of these two goals. Each curve depicts the trade-off between $\mathcal{N}_{\mathrm{TP}}$ and $R_{\mathrm{FP}}$ for a given indicated combination of base rate, effect size, and $\alpha$ level. Each curve is traced out by varying sample size and computing the resulting $\mathcal{N}_{\mathrm{TP}}$ and $R_{\mathrm{FP}}$ values for the indicated combination. Larger samples always produce smaller rates of FPs but generally also produce fewer

TPs, so the numbers of FPs and TPs tend to be positively related. Furthermore, comparison of the curves for different $\alpha$ levels reveals that decreasing $\alpha$ also reduces both the rate of FPs and the number of TPs. Thus, there is an inescapable trade-off between the number of TPs, which might be regarded as a measure of research productivity, and the rate of FPs, which might be regarded as a measure of research reliability.

## 7. USING RESOURCES FOR REPLICATIONS

In view of the disappointingly low replicability of findings in the psychological literature, many have argued that more replication studies should be conducted (e.g., Zwaan et al. 2018). To encourage that, registered replication reports are now being considered by many journals, with researchers having the potential to publish replication studies regardless of their outcomes (e.g., Nosek et al. 2018). Replications could add value to the field by providing additional information about the presence and size of effects of particular theoretical importance.

The value of replications must be weighed against their costs, however (cf. Coles et al. 2018, Miller & Ulrich 2016, Zwaan et al. 2018). Each replication study uses limited research resources that could be used instead for original studies investigating novel phenomena. Replication studies often use even more resources than original studies because they typically use larger samples (e.g., Open Sci. Collab. 2015). Thus, to optimize scientific payoffs, the benefits of replications should be weighed against the costs of reducing the number of original studies that can be carried out with the available resources.

Some of the models for research efficiency considered earlier do have implications about what proportion of research resources should be allocated to replications. If efficiency is optimized by minimizing FPs, for example, presumably few replications would be needed because there would be few FPs in the first place. Likewise, to maximize benefit to individual researchers, few replications should be conducted because these are—at least currently—usually regarded as less helpful than novel findings for career enhancement (e.g., Nosek et al. 2012). In contrast, LeBel et al.'s (2017b) proposal that every new finding should be replicated twice before acceptance implies that a large share of research resources should be devoted to replications.

The question of what proportion of research resources should be allocated to replication efforts can also be examined quantitatively within the payoff model of Miller & Ulrich (2016). Assume that a total pool of $N$ participants can be divided among all studies, including both original studies and replications. Let $\mathcal{R}$ be the proportion of the $N$ participants allocated to replication studies ($0 \leq \mathcal{R} < 1$). The question is, what value of $\mathcal{R}$ would optimize the total payoff summed across original and replication studies?

First, let $\mathcal{P}_{\text{orig}}|N$ be the expected payoff for original studies using all $N$ participants; that is, this is the expected payoff when $\mathcal{R} = 0$. There are $k_{\text{orig}} = N/n_{\text{s}}$ such studies, where $n_{\text{s}}$ is the sample size in each one. Using the same technique as before, we can compute $\mathcal{P}_{\text{orig}}|N$ from $N$, base rate, effect size, $\alpha$, $\mathcal{P}_{\text{TP}}$, and so on. We obtain

$$\mathcal{P}_{\text{orig}}|N = k_{\text{orig}} \times [\Pr(TP) \times \mathcal{P}_{\text{TP}} + \Pr(FP) \times \mathcal{P}_{\text{FP}} + \Pr(TN) \times \mathcal{P}_{\text{TN}} + \Pr(FN) \times \mathcal{P}_{\text{FN}}].$$

Second, let $\mathcal{P}_{\text{rep}}|N$ be the expected payoff that would be obtained if all $N$ participants were used for replication studies. That is, this is the expected payoff when $\mathcal{R} = 1$. This is only a hypothetical quantity, as there would be no original findings to replicate if all participants were used for replication studies, but it is nonetheless convenient to compute its value for comparison purposes. There are $k_{\text{rep}} = N/n_{\text{sr}}$ such studies, where $n_{\text{sr}}$ is the sample size in each replication study. $\mathcal{P}_{\text{rep}}$ can also be computed using the same general payoff formula,

$$\mathcal{P}_{\text{rep}}|N = k_{\text{rep}} \times \left[\Pr(TP_{\text{r}}) \times \mathcal{P}_{\text{TP,r}} + \Pr(FP_{\text{r}}) \times \mathcal{P}_{\text{FP,r}} + \Pr(TN_{\text{r}}) \times \mathcal{P}_{\text{TN,r}} + \Pr(FN_{\text{r}}) \times \mathcal{P}_{\text{FN,r}}\right],$$

where $\Pr(TP_{\text{r}})$ is the probability of a TP in a replication study and $\mathcal{P}_{\text{TP,r}}$ is its payoff (and analogously for the other outcomes).

Note that the values of the replication parameters [$\Pr(TP_{\text{r}})$, $\mathcal{P}_{\text{TP,r}}$, and so on] will most likely be different for replication studies and original studies. For example, assuming that the replications are only carried out for original studies that produced significant positive results, the base rate of true effects will be higher for replications than it was in the original studies, which will affect $\Pr(TP_{\text{r}})$ and the other probabilities. The payoffs (e.g., $\mathcal{P}_{\text{TP,r}}$) would probably also be different for the replication studies, because these studies do not provide novel findings but instead serve to confirm or—perhaps more importantly—disconfirm original positive results. In particular, the payoff for a TN in a replication study, $\mathcal{P}_{\text{TN,r}}$, seems particularly high, because a negative replication result would alert researchers to the possibility that the earlier positive result was an FP.

Finally, using the values of $\mathcal{P}_{\text{orig}}|N$ and $\mathcal{P}_{\text{rep}}|N$, we can compute the overall total payoff, $\mathcal{P}_{\text{total}}$, as a function of $\mathcal{R}$. This total payoff is simply the sum of the payoffs of the original and replication studies, weighing each type of study by the proportion of participants allocated to it, that is,

$$\begin{aligned}\mathcal{P}_{\text{total}}(\mathcal{R}) &= (1 - \mathcal{R}) \times \mathcal{P}_{\text{orig}}|N + \mathcal{R} \times \mathcal{P}_{\text{rep}}|N \\ &= \mathcal{P}_{\text{orig}}|N + \mathcal{R} \times (\mathcal{P}_{\text{rep}}|N - \mathcal{P}_{\text{orig}}|N).\end{aligned} \qquad 1.$$

Equation 1 is quite informative even if none of the quantities needed for computing $\mathcal{P}_{\text{orig}}|N$ and $\mathcal{P}_{\text{rep}}|N$ are known (e.g., base rate, $\mathcal{P}_{\text{TP}}$, etc). In particular, on a first look it implies the surprising conclusion that researchers should either replicate all of their studies or none of them, even though this might sound absurd to many readers. Specifically, if $\mathcal{P}_{\text{rep}}|N > \mathcal{P}_{\text{orig}}|N$, then $\mathcal{P}_{\text{total}}$ is maximized when $\mathcal{R}$ is as large as possible, so researchers should replicate all studies. Alternatively, if $\mathcal{P}_{\text{rep}}|N < \mathcal{P}_{\text{orig}}|N$, then $\mathcal{P}_{\text{total}}$ is maximized when $\mathcal{R} = 0$, which implies that no study should be replicated. Intriguingly, the all-or-none conclusion from this simple model is directly contrary to the common-sense assertion by Coles et al. (2018, p. 16) that "it is unlikely that directly replicating every study, or never directly replicating any study, is optimally efficient."

**Figure 7** depicts examples of the computations needed to compare the payoffs of original versus replication studies for two-sample *t*-test designs with an $\alpha$ level in the original study of either 0.005 or 0.05. As in the computations shown in the sidebar titled Example Payoff Computations, the payoffs for a TP, FP, TN, and FN in the original studies were assumed to be 1, −2, 0.2, and −0.4, respectively. It was assumed that replication studies would only be conducted to replicate original studies that had produced significant findings, so the base rate of true effects in the replication studies was naturally higher than it was in the original studies, and the probabilities of the four replication study outcomes were computed accordingly. The payoffs for a TP, FP, TN, and FN in the replication studies were assumed to be 0.5, −4, 4, and −1, respectively, based on the following principles: (*a*) A replication TP is not as useful as the original TP because it simply reinforces previous evidence; (*b*) a replication FP is especially damaging because it strongly reinforces the original FP; (*c*) a replication TN is especially beneficial because it raises doubt about the original FP; and (*d*) a replication FN is also rather damaging—at least worse than an original FN—because it casts doubt on the original TP.

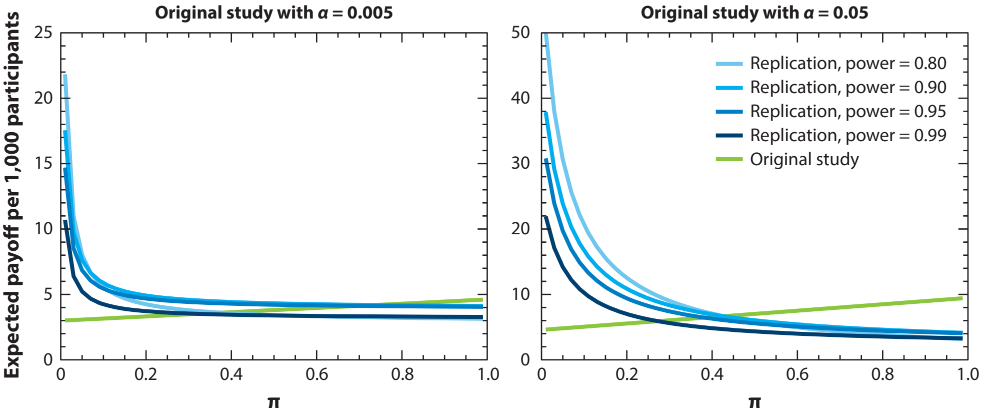


**Figure 7**

Expected payoff per 1,000 participants for original and replication studies as a function of the $\alpha$ level of the original study, the base rate $\pi$ of true effects, and the power of the replication study. The computations are for two-sample *t*-tests with two-tailed testing and an effect size of $d = 0.5$. The sample sizes of the original studies were adjusted to produce an original study power of 0.5 with the indicated original $\alpha$. Replication study sample sizes were adjusted to obtain the indicated power levels. Note that the vertical axis range differs across panels.

As shown in **Figure 7**, the original study's $\alpha$ level and base rate of true effects $\pi$ have substantial influences on the relative payoffs of original versus replication studies. As in most of the scenarios considered by Miller & Ulrich (2019), overall payoffs are generally lower for original studies with $\alpha = 0.005$ compared to studies with $\alpha = 0.05$, suggesting that $\alpha = 0.005$ would not be the optimal $\alpha$ for original studies in the first place. Regardless of the original $\alpha$, however, there is a clear crossover of payoffs for the original and replication studies depending on the base rate and the replication study's power. Specifically, payoffs are higher for replication studies when the base rate of true effects is less than approximately $\pi = 0.4$, but payoffs are higher for original studies than for replications when the base rate is larger than that. In areas with relatively low base rates, using the limited participant-testing resources to replicate all significant original findings would therefore be a worthwhile investment. In areas with relatively high base rates, however, overall payoffs would be greater if all resources were devoted to new original studies. Naturally, the exact crossover base rate would depend on the specific research scenario (e.g., statistical test, effect size, $\alpha$ level) and payoffs, but it seems intuitively quite reasonable for replications to be especially worthwhile when base rates are low.

**Figure 7** also shows that with these parameters, payoffs tend to be slightly higher for replication studies with only moderate power (i.e., smaller sample sizes), contrary to the common practice of designing replication studies to have extremely high power (e.g., Open Sci. Collab. 2015). Interestingly, however, the crossover point with the original study's payoffs depends little on replication power, at least across the range of powers considered here.

The patterns depicted in **Figure 7** come with at least three caveats. First, the plotted expected payoff values depend on the payoffs assumed for the individual outcomes (original FNs, replication TNs, etc.). Such individual payoff values are difficult to estimate and probably vary between areas, as mentioned above, and it is possible that the patterns would change qualitatively under different assumptions about the individual outcome payoffs. Second, replication studies may have value not only because they provide information with respect to an individual phenomenon under investigation, as assumed for the computations in **Figure 7**, but also because they provide general information about a research field. For example, the replications reported by Open Sci. Collab. (2015) are informative not only about the replicability of the 100 specific effects studied but also, to some extent, about the replicability of the wider class of effects from which these 100 were selected (e.g., Kuehberger & Schulte-Mecklenbeck 2018). The value of this more general information about effects reported in psychological journals may not be adequately reflected in the assumed individual replication outcome payoffs that underlie **Figure 7**. Third, the proportion of research resources allocated to replication studies is ultimately decided by the individual researchers' choices of whether to do an original study or a replication. In the long run, it may be necessary to modify the incentives for individual researchers to help ensure that individual researchers make these choices in ways that would maximize the overall scientific payoff within the field (e.g., Zwaan et al. 2018). Thus, although quantitative models are helpful in principle for analyzing the resource trade-off between original studies and replications, changes in scientific culture may be necessary to shape actual practice in a manner that would be optimal for the field as a whole (e.g., Lilienfeld 2017, Nosek et al. 2012, Zwaan et al. 2018).

## 8. SEQUENTIAL SAMPLING METHODS

Most of the discussion about research optimization has focused on research designs with fixed sample sizes (e.g., on what sample sizes should be used), presumably because the vast majority of studies use such designs. Within the statistical literature, however, it has long been known that designs with sequential sampling are more efficient in some situations, often providing statistical power equal to that of fixed-sample designs with 20–30% smaller sample sizes, on average (Wald 1947). Sequential designs can thus increase research efficiency by allowing a greater number of studies to be performed with the same research resources (e.g., Lakens 2014).

With sequential designs, researchers collect and analyze the data incrementally, stopping when the data provide evidence one way or another regarding the hypothesis under test. Although sequential data collection and analysis can inflate Type 1 error rates when done improperly, accepted sequential methods hold Type 1 error rates constant at the researcher's chosen nominal $\alpha$ level (e.g., Albers 2019). Unfortunately, these methods are computationally complex, which may explain why they are not widely used in psychology (Albers 2019, Lakens 2014).

Although it is beyond the scope of this article to describe a wide range of sequential methods (for a recent review, see Schnuerch & Erdfelder 2020), we cannot resist the temptation to mention the particularly simple independent-segments procedure that we recently proposed (Miller & Ulrich 2021). With this procedure, the researcher collects and independently analyzes data in at most $k_{\text{max}}$ separate subsamples or segments, possibly stopping after each segment based on the data from that segment. In particular, to achieve any given overall $\alpha$ level, the researcher stops and fails to reject the null hypothesis after any segment if the data yield an observed $p$ value that is greater than $\alpha^{1/k_{\text{max}}}$. If the null hypothesis is true, the probability of carrying out all $k_{\text{max}}$ segments and rejecting the null hypothesis is $\left(\alpha^{1/k_{\text{max}}}\right)^{k_{\text{max}}} = \alpha$, as desired.

As a simple example, suppose a fixed-sample researcher plans a study that will be analyzed using a two-group $t$-test and wants to have a power of $1 - \beta = 0.8$ to detect an effect of $d = 0.5$ with a nominal $\alpha = 0.01$, one-tailed. Assuming equal group sizes, standard power calculations indicate that the researcher needs a total sample size of 162.

In contrast, an independent-segments researcher planning the analogous study might use three segments, intending to stop if the observed $p$ in any segment exceeds $0.01^{1/3} = 0.21544$. Calculations specific to this method indicate that the researcher needs samples of 82 in each segment to have a power of $1 - \beta = 0.8$ with an effect of $d = 0.5$.[5] In essence, the different segments can be thought of as mini-replications of an initial effect that was weakly significant (i.e., $p < 0.21544$). If the null hypothesis is true, there is only a probability of $\alpha = 0.21544^3 = 0.01$ that all three segments would produce $p$ values lower than this weak cutoff and that the null hypothesis would therefore be rejected.

Which researcher needs to test more participants? Of course, the fixed-sample researcher always tests 162 participants, whether the effect is present or not. In contrast, it can be shown that the independent-segments researcher will test, on average, 229.1 participants when a true effect is present but only 103.5 participants when it is not. If expected effects are present 20% of the time (i.e., $\pi = 0.2$)—a reasonable estimate of the base rate of true effects in at least some areas of psychology (e.g., Miller & Ulrich 2016)—then on average the independent-segments researcher only tests $229.1 \times 0.2 + 103.5 \times 0.8 = 128.6$ participants, leading to an average savings of more than 33 participants per study. The savings is even larger if the base rate is lower, as many have suggested may be the case in some areas (e.g., Dreber et al. 2015, McElreath & Smaldino 2015, Wilson & Wixted 2018). Moreover, the independent-segments procedure can be extended to allow for the possibility of stopping early if very strong evidence of an effect is found in any segment (e.g., $p < 0.005$), while still maintaining the desired Type 1 error probability. This extension can further reduce the expected sample size, so that in many cases the procedure requires smaller average samples than the fixed-sample approach not only when the null hypothesis is true but also when it is false (for details, see Miller & Ulrich 2021).

## 9. CONCLUSIONS

Quantitative analyses of scientific publications have provided strong evidence that current scientific methods are suboptimal, not only in psychology but also in many other fields with similar statistical variability. Given this evidence, the current ongoing discussion about how to improve these methods is timely (Freese & Peterson 2017), and its wide-ranging nature offers promise that better methods will be found. Nevertheless, as we hope this review has illustrated, identifying these methods will not be easy, because the research enterprise involves complex trade-offs.

Because of these trade-offs, it is unlikely that optimal research methods can be identified by focusing narrowly on any isolated measure of research efficiency. For example, as discussed above, much recent discussion has regarded the low replication rates as indicating a crisis within psychology and other fields (e.g., Pashler & Wagenmakers 2012), and researchers have been encouraged to increase replicability by reducing $\alpha$ and increasing sample sizes. Even granting the importance of replicability (but see, e.g., Francis 2013 for an alternative viewpoint), it must be acknowledged that the proposed changes in $\alpha$ and sample size tend to decrease study power and increase study costs—neither of which is desirable. In a research scenario with a 20% base rate of small effects (i.e., $d = 0.2$), for example, a researcher would have the choice between either running a certain number of large studies with $\alpha = 0.005$ and 80% power, obtaining results that are 97.5% replicable, or running six times as many small studies with $\alpha = 0.05$ and 40% power, obtaining results that are 67% replicable. It is debatable whether choosing the option producing higher replicability would necessarily result in the fastest scientific progress.

[5] For computational formulas, readers are referred to Miller & Ulrich (2021), or they might just use the toolbox at **https://phaden.shinyapps.io/seght_shiny** for calculations.

To make such choices rationally will require a complex model of the entire scientific process. This model must take into account statistical factors like $\alpha$ level, base rate, effect size, and power, as well as practical factors like resource constraints and researcher incentives. The preliminary outlines of such a model and estimates of its parameters are starting to emerge from ongoing metascientific work, but much more work will be needed to produce and refine a model of science that can allow researchers to choose optimal research designs. It seems especially critical to obtain good estimates of base rates, for example, because this parameter has such a large influence on the rates of FPs (e.g., Ioannidis 2005, Ulrich & Miller 2020) and on the relative value of replications (e.g., **Figure 7**). Of course, the development of such a model and the estimation of its parameters must necessarily involve simplifications and approximations, because such things are—as was said about research problems more generally—"simply an inevitable feature of the way science works" (Chalmers et al. 2014, p. 156). There are good reasons to be optimistic about the development of useful metascientific models, though. The skills of reasoning, model building, data collection, and data analysis that researchers acquire within their specific areas appear quite general (e.g., Schunn & Anderson 1999), and they should thus also be effective in studying the scientific methods themselves.

## DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

## ACKNOWLEDGMENTS

The authors thank Edgar Erdfelder for helpful discussions concerning power, Derek Berger for assistance in understanding the R code associated with LeBel et al. (2017a), Jackie Hunter for pointers to research on COVID-19, and Patricia Haden and Eli Finkel for comments on earlier versions of this article.

## LITERATURE CITED

Aczel B, Palfi B, Szaszi B. 2017. Estimating the evidential value of significant results in psychological science. *PLOS ONE* 12(8):e0182651

Albers C. 2019. The problem with unadjusted multiple and sequential statistical testing. *Nat. Commun.* 10:1921

Amrhein V, Greenland S, McShane B. 2019. Retire statistical significance. *Nature* 567:305–7

Armitage P, McPherson CK, Rowe BC. 1969. Repeated significance tests on accumulating data. *J. R. Stat. Soc. A* 132(2):235–44

Asendorpf JB, Conner M, De Fruyt F, De Houwer J, Denissen JJA, et al. 2013. Recommendations for increasing replicability in psychology. *Eur. J. Pers.* 27(2):108–19

Baker DH, Vilidaite G, Lygo FA, Smith AK, Flack TR, et al. 2021. Power contours: optimising sample size and precision in experimental psychology and human neuroscience. *Psychol. Methods* 26(3):295–314

Baker M. 2016. Is there a reproducibility crisis? *Nature* 533:452–54

Baker SG, Heidenberger K. 1989. Choosing sample sizes to maximize expected health benefits subject to a constraint on total trial costs. *Med. Decis. Mak.* 9(1):14–25

Bakker M, Van Dijk A, Wicherts JM. 2012. The rules of the game called psychological science. *Perspect. Psychol. Sci.* 7(6):543–54

Barrett LF. 2020. Forward into the past. *Observer* 33(3):5–7

Baumeister RF. 2016. Charting the future of social psychology on stormy seas: winners, losers, and recommendations. *J. Exp. Soc. Psychol.* 66:153–58

Begley CG, Ellis LM. 2012. Drug development: raise standards for preclinical cancer research. *Nature* 483(7391):531–33

Begley CG, Ioannidis JPA. 2015. Reproducibility in science: improving the standard for basic and preclinical research. *Circ. Res.* 116(1):116–26

Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, et al. 2018. Redefine statistical significance. *Nat. Hum. Behav.* 2:6–10

Bero L. 2018. Meta-research matters: meta-spin cycles, the blindness of bias, and rebuilding trust. *PLOS Biol.* 16(4):e2005972

Berry DA, Ho CH. 1988. One-sided sequential stopping boundaries for clinical trials: a decision-theoretic approach. *Biometrics* 44(1):219–27

Białek M. 2018. Replications can cause distorted belief in scientific progress. *Behav. Brain Sci.* 41:e122

Brown AN, Wood BDK. 2018. Replication studies of development impact evaluations. *J. Dev. Stud.* 55(5):917–25

Bueno de Mesquita B, Gleditsch NP, James P, King G, Metelits C, et al. 2003. Symposium on replication in international studies research. *Int. Stud. Perspect.* 4(1):72–107

Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, et al. 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* 14(5):365–76

Button KS, Munafò MR. 2017. Powering reproducible research. In *Psychological Science Under Scrutiny: Recent Challenges and Proposed Remedies*, ed. SO Lilienfeld, ID Waldman, pp. 22–33. New York: Wiley

Carey B. 2015. Many psychology findings not as strong as claimed, study says. *New York Times*, Aug. 27

Chalmers I, Bracken MB, Djulbegovic B, Garattini S, Grant J, et al. 2014. How to increase value and reduce waste when research priorities are set. *Lancet* 383(9912):156–65

Chalmers I, Glasziou P. 2009. Avoidable waste in the production and reporting of research evidence. *Lancet* 374(9683):86–89

Chambers CD. 2020. Frontloading selectivity: a third way in scientific publishing? *PLOS Biol.* 18(3):e3000693

Clark-Carter D. 1997. The account taken of statistical power in research published in the *British Journal of Psychology*. *Br. J. Psychol.* 88:71–83

Cohen J. 1962. The statistical power of abnormal-social psychological research: a review. *J. Abnorm. Soc. Psychol.* 65:145–53

Cohen J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum. 2nd ed.

Coles NA, Tiokhin L, Scheel AM, Isager PM, Lakens D. 2018. The costs and benefits of replication studies. *Behav. Brain Sci.* 41:e124

Colhoun HM, McKeigue PM, Smith GD. 2003. Problems of reporting genetic associations with complex outcomes. *Lancet* 361(9360):865–72

Colquhoun D. 2014. An investigation of the false discovery rate and the misinterpretation of *p*-values. *R. Soc. Open Sci.* 1(3):140216

Cumming G. 2014. The new statistics: why and how. *Psychol. Sci.* 25(1):7–29

Detsky AS. 1985. Using economic analysis to determine the resource consequences of choices made in planning clinical trials. *J. Chronic Dis.* 38(9):753–65

Dreber A, Pfeiffer T, Almenberg J, Isaksson S, Wilson B, et al. 2015. Using prediction markets to estimate the reproducibility of scientific research. *PNAS* 112(50):15343–47

Dunbar KN, Fugelsang JA. 2005. Causal thinking in science: how scientists and students interpret the unexpected. In *Scientific and Technological Thinking*, ed. ME Gorman, RD Tweney, DC Gooding, AP Kincannon, pp. 57–79. Mahwah, NJ: Lawrence Erlbaum

Edwards MA, Roy S. 2017. Academic research in the 21st century: maintaining scientific integrity in a climate of perverse incentives and hypercompetition. *Environ. Eng. Sci.* 34(1):51–61

Etz A, Vandekerckhove J. 2016. A Bayesian perspective on the reproducibility project: psychology. *PLOS ONE* 11(2):e0149794

Fanelli D. 2012. Negative results are disappearing from most disciplines and countries. *Scientometrics* 90(3):891–904

Fanelli D, Costas R, Larivière V. 2015. Misconduct policies, academic culture and career stage, not gender or pressures to publish, affect scientific integrity. *PLOS ONE* 10(6):e0127556

Fiedler K, Kutzner F, Krueger JI. 2012. The long way from $\alpha$-error control to validity proper: problems with a short-sighted false-positive debate. *Perspect. Psychol. Sci.* 7(6):661–69

Fiedler K, Schott M. 2017. False negatives. In *Psychological Science Under Scrutiny: Recent Challenges and Proposed Remedies*, ed. SO Lilienfeld, ID Waldman, pp. 53–72. New York: Wiley

Finkel EJ, Eastwick PW, Reis HT. 2015. Best research practices in psychology: illustrating epistemological and pragmatic considerations with the case of relationship science. *J. Pers. Soc. Psychol.* 108(2):275–97

Finkel EJ, Eastwick PW, Reis HT. 2017. Replicability and other features of a high-quality science: toward a balanced and empirical approach. *J. Pers. Soc. Psychol.* 113(2):244–53

Fisher RA. 1925. *Statistical Methods for Research Workers*. Edinburgh, UK: Oliver & Boyd

Francis G. 2013. We don't need replication, but we do need more data. *Eur. J. Pers.* 27(2):125–26

Freese J, Peterson D. 2017. Replication in social science. *Annu. Rev. Sociol.* 43:147–65

Gilbert DT, King G, Pettigrew S, Wilson TD. 2016. Comment on "Estimating the reproducibility of psychological science." *Science* 351(6277):1037–37

Gillett R. 1994. The average power criterion for sample size estimation. *Statistician* 43:389–94

Gross C. 2016. Scientific misconduct. *Annu. Rev. Psychol.* 67:693–711

Hamann S, Canli T. 2004. Individual differences in emotion processing. *Curr. Opin. Neurobiol.* 14(2):233–38

Hartman TK, Stocks TVA, McKay R, Gibson-Miller J, Levita L, et al. 2021. The authoritarian dynamic during the COVID-19 pandemic: effects on nationalism and anti-immigrant sentiment. *Soc. Psychol. Pers. Sci.* 12(7):1274–85

Hartshorne JK, Schachner A. 2012. Tracking replicability as a method of post-publication open evaluation. *Front. Comput. Neurosci.* 6:8

Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. 2015. The extent and consequences of *p*-hacking in science. *PLOS Biol.* 13(3):e1002106

Ioannidis JPA. 2005. Why most published research findings are false. *PLOS Med.* 2(8):e124

Ioannidis JPA. 2018. Meta-research: why research on research matters. *PLOS Biol.* 16(3):e2005468

John LK, Loewenstein G, Prelec D. 2012. Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychol. Sci.* 23:524–32

Johnson VE. 2013. Revised standards for statistical evidence. *PNAS* 110(48):19313–17

Karrandinos MG. 1976. Optimum sample size and comments on some published formulae. *Bull. Entomol. Soc. Am.* 22(4):417–21

Kuehberger A, Schulte-Mecklenbeck M. 2018. Selecting target papers for replication. *Behav. Brain Sci.* 41:e139

Kuhn TS. 1962. *The Structure of Scientific Revolutions*. Chicago: Univ. Chicago Press

Lakens D. 2014. Performing high-powered studies efficiently with sequential analyses. *Eur. J. Soc. Psychol.* 44(7):701–10

Lakens D, Adolfi FG, Albers CJ, Anvari F, Apps MAJ, et al. 2018. Justify your alpha: a response to "Redefine statistical significance." *Nat. Hum. Behav.* 2:168–71

Lakens D, Evers ERK. 2014. Sailing from the seas of chaos into the corridor of stability: practical recommendations to increase the informational value of studies. *Perspect. Psychol. Sci.* 9(3):278–92

LeBel EP, Berger D, Campbell L, Loving TJ. 2017a. Falsifiability is not optional. *J. Pers. Soc. Psychol.* 113(2):254–61

LeBel EP, Campbell L, Loving TJ. 2017b. Benefits of open and high-powered research outweigh costs. *J. Pers. Soc. Psychol.* 113(2):230–43

Leek JT, Peng RD. 2015. Statistics: *P* values are just the tip of the iceberg. *Nature* 520:612

Lenth RV. 2001. Some practical guidelines for effective sample size determination. *Am. Stat.* 55(3):187–93

Lewandowsky S, Oberauer K. 2020. Low replicability can support robust and efficient science. *Nat. Commun.* 11:358

Lilienfeld SO. 2017. Psychology's replication crisis and the grant culture: righting the ship. *Perspect. Psychol. Sci.* 12(4):660–64

Loftus GR. 1996. Psychology will be a much better science when we change the way we analyze data. *Curr. Direct. Psychol. Sci.* 5:161–71

Maxwell SE. 2004. The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychol. Methods* 9:147–63

McElreath R, Smaldino PE. 2015. Replication, communication, and the population dynamics of scientific discovery. *PLOS ONE* 10(8):e0136088

McGrath JE. 1981. Dilemmatics: the study of research choices and dilemmas. *Am. Behav. Sci.* 25(2):179–210

McShane BB, Böckenholt U. 2014. You cannot step into the same river twice: when power analyses are optimistic. *Perspect. Psychol. Sci.* 9(6):612–25

McShane BB, Gal D, Gelman A, Robert C, Tackett JL. 2019. Abandon statistical significance. *Am. Stat.* 73:235–45

Michaels R. 2017. Confidence in courts: a delicate balance. *Science* 357(6353):764

Miller JO, Ulrich R. 2016. Optimizing research payoff. *Perspect. Psychol. Sci.* 11(5):664–91

Miller JO, Ulrich R. 2019. The quest for an optimal alpha. *PLOS ONE* 14(1):e0208631

Miller JO, Ulrich R. 2021. A simple, general, and efficient method for sequential hypothesis testing: the independent segments procedure. *Psychol. Methods* 26(4):486–97

Miller MG. 1996. *Optimal allocation of resources to clinical trials*. PhD Thesis, Sloan Sch. Manag., Mass. Inst. Technol., Cambridge

Mosteller F, Weinstein M. 1985. Toward evaluating the cost-effectiveness of medical and social experiments. In *Social Experimentation*, ed. JA Hausman, DA Wise, pp. 221–50. Chicago: Univ. Chicago Press

Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, et al. 2015. Promoting an open research culture. *Science* 348(6242):1422–25

Nosek BA, Ebersole CR, DeHaven AC, Mellor DT. 2018. The preregistration revolution. *PNAS* 115(11):2600–6

Nosek BA, Spies JR, Motyl M. 2012. Scientific utopia II: restructuring incentives and practices to promote truth over publishability. *Perspect. Psychol. Sci.* 7(6):615–31

Nuzzo R. 2014. Scientific method: statistical errors. *Nature* 506(7487):150–52

Olsson-Collentine A, Wicherts JM, van Assen MALM. 2020. Heterogeneity in direct replications in psychology and its association with effect size. *Psychol. Bull.* 146(10):922–40

Open Sci. Collab. 2015. Estimating the reproducibility of psychological science. *Science* 349(6251):aac4716

Pashler HE, Harris C. 2012. Is the replicability crisis overblown? Three arguments examined. *Perspect. Psychol. Sci.* 7(6):531–36

Pashler HE, Wagenmakers E. 2012. Editors' introduction to the special section on replicability in psychological science: a crisis of confidence? *Perspect. Psychol. Sci.* 7(6):528–30

Poldrack RA. 2019. The costs of reproducibility. *NeuroView* 10(1):11–14

Popper KR. 2002 (1963). *Conjectures and Refutations: The Growth of Scientific Knowledge*. London: Taylor & Francis

Roberts RM. 1989. *Serendipity: Accidental Discoveries in Science*. New York: Wiley

Rosenthal R. 1979. The "file drawer problem" and tolerance for null results. *Psychol. Bull.* 86:638–41

Rossi JS. 1990. Statistical power of psychological research: What have we gained in 20 years? *J. Consult. Clin. Psychol.* 58(5):646–56

Saltelli A, Funtowicz S. 2017. What is science's crisis really about? *Futures* 91:5–11

Schimmack U. 2012. The ironic effect of significant results on the credibility of multiple-study articles. *Psychol. Methods* 17(4):551–66

Schimmack U. 2020. A meta-psychological perspective on the decade of replication failures in social psychology. *Can. Psychol. Psychol. Can.* 61(4):364–76

Schnuerch M, Erdfelder E. 2020. Controlling decision errors with minimal costs: the sequential probability ratio *t* test. *Psychol. Methods* 25(2):206–26

Schooler J. 2019. Metascience: the science of doing science. *Observer* 32(9):26–29

Schunn CD, Anderson JR. 1999. The generality/specificity of expertise in scientific reasoning. *Cogn. Sci.* 23(3):337–70

Sedlmeier P, Gigerenzer G. 1989. Do studies of statistical power have an effect on the power of studies? *Psychol. Bull.* 105(2):309–16

Sherman RA, Pashler H. 2019. Powerful moderator variables in behavioral science? Don't bet on them (version 3). PsyArxiv, May 24. **https://doi.org/10.31234/osf.io/c65wm**

Sibley CG, Greaves LM, Satherley N, Wilson MS, Overall NC, et al. 2020. Effects of the COVID-19 pandemic and nationwide lockdown on trust, attitudes towards government, and well-being. *Am. Psychol.* 75(5):618–30

Simmons JP, Nelson LD, Simonsohn U. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* 22(11):1359–66

Simon H. 1947. *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization*. New York: Free Press. 2nd ed.

Simonsohn U, Nelson LD, Simmons JP. 2014. *P*-curve: a key to the file-drawer. *J. Exp. Psychol. Gen.* 143(2):534–47

Smaldino PE, McElreath R. 2016. The natural selection of bad science. *R. Soc. Open Sci.* 3(9):160384

Stanley TD, Carter EC, Doucouliagos H. 2018. What meta-analyses reveal about the replicability of psychological research. *Psychol. Bull.* 144(12):1325–46

Sterling TD. 1959. Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *J. Am. Stat. Assoc.* 54(285):30–34

Sternberg RJ, Sternberg K. 2010. *The Psychologist's Companion: A Guide to Writing Scientific Papers for Students and Researchers*. New York: Cambridge Univ. Press

Stroebe W, Postmes T, Spears R. 2012. Scientific misconduct and the myth of self-correction in science. *Perspect. Psychol. Sci.* 7(6):670–88

Stroebe W, Strack F. 2014. The alleged crisis and the illusion of exact replication. *Perspect. Psychol. Sci.* 9(1):59–71

Strube MJ. 2006. SNOOP: a program for demonstrating the consequences of premature and repeated null hypothesis testing. *Behav. Res. Methods* 38(1):24–27

Ulrich R, Miller JO. 2018. Some properties of *p*-curves, with an application to gradual publication bias. *Psychol. Methods* 23(3):546–60

Ulrich R, Miller JO. 2020. Meta-research: Questionable research practices may have little effect on replicability. *eLife* 9:e58237

Ulrich R, Miller JO, Erdfelder E. 2018. Effect size estimation from *t*-statistics in the presence of publication bias: a brief review of existing approaches with some extensions. *Z. Psychol.* 226(1):56–80

Van Bavel JJ, Mende-Siedlecki P, Brady WJ, Reinero DA. 2016. Contextual sensitivity in scientific reproducibility. *PNAS* 113(23):6454–59

Wagenmakers EJ, Wetzels R, Borsboom D, Van Der Maas HLJ. 2011. Why psychologists must change the way they analyze their data: the case of psi: comment on Bem (2011). *J. Pers. Soc. Psychol.* 100(3):426–32

Wald A. 1947. *Sequential Analysis*. New York: Wiley

Williams B, Myerson J, Hale S. 2008. Individual differences, intelligence, and behavior analysis. *J. Exp. Anal. Behav.* 90(2):219–31

Wilson BM, Wixted JT. 2018. The prior odds of testing a true effect in cognitive and social psychology. *Adv. Methods Pract. Psychol. Sci.* 1(2):186–97

Witt JK. 2019. Insights into criteria for statistical significance from signal detection analysis. *Meta-Psychology* 3. **https://doi.org/10.15626/MP.2018.871**

Yong E. 2012. Replication studies: bad copy. *Nature* 485:298–300

Zwaan RA, Etz A, Lucas RE, Donnellan MB. 2018. Making replication mainstream. *Behav. Brain Sci.* 41:e120