Social Psychological and Personality Science
1-8

DOI: 10.1177/1948550615584199
spps.sagepub.com


A Powerful Nudge? Presenting Calculable Consequences of Underpowered Research Shifts Incentives Toward Adequately Powered Designs

Will M. Gervais[1], Jennifer A. Jewell[1], Maxine B. Najle[1], and Ben K. L. Ng[1]

Abstract

If psychologists have recognized the pitfalls of underpowered research for decades, why does it persist? Incentives, perhaps: underpowered research benefits researchers individually (increased productivity), but harms science collectively (inflated Type I error rates and effect size estimates but low replication rates). Yet, researchers can selectively reward power at various scientific bottlenecks (e.g., peer review, hiring, funding, and promotion). We designed a stylized thought experiment to evaluate the degree to which researchers consider power and productivity in hiring decisions. Accomplished psychologists chose between a low sample size candidate and a high sample size candidate who were otherwise identical. We manipulated the degree to which participants received information about (1) productivity, (2) sample size, and (3) directly calculable Type I error and replication rates. Participants were intolerant of the negative consequences of low-power research, yet merely indifferent regarding the practices that logically produce those consequences, unless those consequences were made quite explicit.

Keywords

methods, power, incentives, replication, research practices

> We refuse to believe that a serious investigator will knowingly accept a .50 risk of failing to confirm a valid research hypothesis. (Tversky & Kahneman, 1971, p. 110)

Power matters. The negative consequences of conducting low-power research are well documented (e.g., Cohen, 1962, 1992a, 1992b). For example, low-power research inflates the share of Type I errors in the published literature (Overall, 1969), makes it difficult for researchers to detect genuine effects (e.g., Cohen, 1992a), and severely hampers replication efforts. Researchers' opinions are nonetheless disproportionately swayed by low power, statistically significant results (Tversky & Kahneman, 1971).

As a concept, power is neither new nor obscure. The previous paragraph cites five articles. On average, they are 36.8 years old (median = 43) and have been cumulatively cited more than 18,000 times (mean = 3,718.6, median = 947). Nonetheless, effect sizes in psychological science tend to be moderate (e.g., Button et al., 2013; Richard, Bond, & Stokes-Zoota, 2003; Sedlmeier & Gigerenzer, 1989) and samples are often too small to reliably detect even trivially obvious and strong effects (Simmons, Nelson, & Simonsohn, 2013: "men weigh more than women"). This combination yields very low power yet is persistently prevalent (e.g., Button et al., 2013; Cohen, 1962; Sedlmeier & Gigerenzer, 1989; Vankov, Bowers, & Munafò, 2014), even—and perhaps especially—in high-impact journals (Bertamini & Munafò, 2012; Fraley & Vazire, 2014).

If power is so important, why does underpowered research predominate the literature? Potentially, the proliferation of low-power research stems not from researchers' inattentiveness to or ignorance of power but rather from a strategic methodological choice to maximize productivity (e.g., Bertamini & Munafò, 2012; Vankov et al., 2014) paired with an incentive structure favorable toward underpowered designs (for discussion, see, e.g., Nosek, Spies, & Motyl, 2012). Even in the absence of questionable research practices (cf. Simmons, Nelson, & Simonsohn, 2011), a strategy of running lots of underpowered studies can result in more significant results than a strategy of running fewer adequately powered studies (Bakker, van Dijk, & Wicherts, 2012). When resources are finite, researchers must balance the number of participants they can

[1] University of Kentucky, Lexington, KY, USA

Corresponding Author:
Will M. Gervais, University of Kentucky Psychology, Kastle Hall, Lexington, KY 40506, USA.
Email: will.gervais@uky.edu

run in any given study against the overall number of studies they can run.

Consider, for example, two nearly identical researchers. They study phenomena with the same underlying effect size (Cohen's $d = .4$). Their hypotheses are both correct half of the time. They have access to the same total number of participants per year. Neither employs questionable research practices. They both run simple two-group between-subject designs. The only difference between the two is that Researcher A runs experiments with 25 participants per condition, but Researcher B runs experiments with 100 participants per condition. Who wins? Despite running severely underpowered studies (power = .28, to Bs .80), A would generate 56% more significant results than B.

This is great news for Researcher A but terrible news for psychological science. Straightforward calculations (see Online Supplement, also Button et al., 2013; Colquhoun, 2014) reveal that 15% of Researcher As statistically significant results are Type I errors, and that in exact replications (identical *N*), Researcher A's results will replicate less than 25% of the time (52% at 2.5 *N*). On the other hand, Researcher B's significant effects replicate 76% of the time (94% at 2.5 *N*) and only include 6% Type I errors.

At all effect sizes, Researcher A will be more productive, but Researcher B's results will be easier to replicate (Figure 1). Running underpowered studies thus might constitute a type of *performance-enhancing design* that inflates an individual researcher's productivity while having deleterious consequences for the collective enterprise of science.

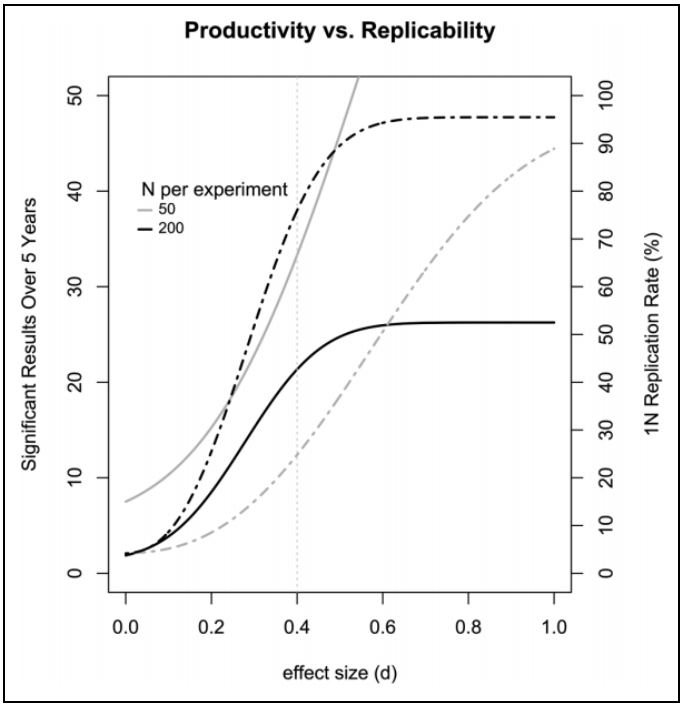

Figure 1. Productivity and replicability of two sample size strategies, given 2,000 participants per year (Researcher A in gray and Researcher B in black). Solid lines represent statistically significant findings generated over a 5-year span by each strategy (left y axis). Dashed lines represent expected exact (1 *N*) replication rates for each strategy (right y axis). Dotted gray line represents a typical effect size in social psychology (adapted from Richard et al., 2003).

Although the negative consequences of low-power research afflict everything from individual studies (yielding a low probability of supporting valid hypotheses) to interpretation of the literature as a whole (yielding overinflated effect size estimates, an undesirable number of published false positives, and undesirably low replication rates), there do exist many checkpoints at which scientists can evaluate—and hopefully reward—methodological choices that take power seriously. For example, researchers may reward power when evaluating (1) individual articles in peer review, (2) journals for the quality of evidence they tend to present (Fraley & Vazire, 2014), (3) methodological choices of job and promotion candidates, and (4) methodological choices underpinning entire programs of research. Classic (Cohen, 1962) and contemporary (Button et al., 2013; Fraley & Vazire, 2014) analyses suggest that power has perhaps been largely overlooked at many of these checkpoints for more than five decades, thus perpetuating an incentive structure rewarding productivity at the expense of power (e.g., Bakker et al., 2012).

We sought to design a simplistic and idealized thought experiment to both evaluate the degree to which researchers consider and reward power when given information they can easily access, and also test whether researchers are more likely to reward power when the negative consequences of low-power research are made plain. To do so, we—in the spirit of Tversky and Kahneman (1971)—surveyed elite social psychologists. In this thought experiment, we focused on potential hiring decisions. Hiring decisions offer an opportunity to evaluate and selectively reward the work of others, given only partial information about their research. Some aspects of research programs are fairly transparent (e.g., number of publications, typical sample size, and effect size estimates), while other aspects are quite opaque (e.g., how a priori probable their hypotheses tend to be, how many total participants they can access). Thus, we could create stylized—and admittedly unrealistic—choices between candidates, while selectively varying the amount of information participants had about each candidate.

We surveyed psychologists regarding their hiring preferences for Researcher A and Researcher B, as described previously. Across conditions, we experimentally manipulated whether participants received (1) information about productivity, (2) information about productivity as well as sample sizes (mirroring information that could reasonably be gleaned from candidate dossiers), and (3) information about productivity and sample sizes, as well as expected consequences for replication rates and publishable false positive rates, directly calculable from the information given in the second condition. This allowed us to quantitatively assess the degree to which sample size information impacts the choices researchers make when evaluating and rewarding research. More importantly, it allowed us to test the focal

Table 1. Full Summary of Stimuli Used Across Conditions.

General introduction:	Imagine two hypothetical job candidates. They're similar in a lot of ways. Both of them run 2,000 participants per year. Their hypotheses tend to be right half of the time. They use simple two-group experimental designs (between subjects), and they study similar phenomena, with similar effect sizes (Cohen's $d = .4$, which is typical for social psychology as a whole). Neither candidate employs questionable research practices
Findings:	Over the past 5 years, Job Candidate A has published 33 statistically significant experiments and over the past 5 years, Job Candidate B has published 21 statistically significant experiments
Sample size:	Job Candidate A runs experiments with 25 participants per condition (50 participants per experiment), Job Candidate B runs experiments with 100 participants per condition (200 participants per experiment), Over the past 5 years, Job Candidate A has published 33 statistically significant experiments, and Over the past 5 years, Job Candidate B has published 21 statistically significant experiments
Consequences:	Job Candidate A runs experiments with 25 participants per condition (50 participants per experiment), Job Candidate B runs experiments with 100 participants per condition (200 participants per experiment), over the past 5 years, Job Candidate A has published 33 statistically significant experiments. Of these 33 significant findings, 15% of them are false positives. In exact replication attempts, Candidate A's statistically significant experiments are successfully replicated 25% of the time, and over the past 5 years, Job Candidate B has published 21 statistically significant experiments. Of these 21 significant findings, 6% are false positives. In exact replication attempts, Candidate B's statistically significant experiments are successfully replicated 76% of the time

hypothesis that presenting explicit information about the easily calculable negative consequences of low-power research would lead researchers to more directly incentivize power.

Methods

We preregistered all methods, hypotheses, and analysis code (R Core Team, 2012) through Open Science Framework (OSF; https://osf.io/r7tk8/). All data and code will be made available on the first author's website (http://willgervais.com/journal-articles/) and through OSF (https://osf.io/m2ve3/). We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study.

Participants

In this experiment, we wanted to recruit elite practicing psychologists. Because we (the authors of the present article) are social psychologists, we decided to focus on a population of social psychologists. This enabled us to design a study using fictional job candidates who make methodological choices (between-subject design, sample sizes, and effect size) that would be plausible. To be perfectly clear, we are not trying to highlight incentives for low-power research as a strictly *social psychology* problem. It is a *psychology* problem (Button et al., 2013; Vankov et al., 2014). We just happen to be social psychologists.

Thus, participants were members of the Society of Experimental Social Psychology (SESP), which is an invite-only professional organization. Members of SESP must be at least 5 years post-PhD, must be nominated by a colleague, and must show evidence of substantial contribution to social psychology as an empirical science. Using SESP's online member list, we created an e-mail list consisting of 937 researchers. We sent an e-mail to this list inviting participation on June 13, 2014. On July 15, 2014, we sent a reminder to the e-mail list. Our target sample size was 200, with the qualification that our population was limited and participation was entirely voluntary. In our preregistration, we declared that if we did not have 200 participants after two e-mails, we would send a third e-mail. After sending the first e-mail in June, we decided to only send two e-mails because we did not want to unduly annoy our colleagues in SESP. We made this decision immediately—before checking any data—upon sending the first e-mail and receiving responses to it (in terms of both responses to the online survey and e-mail responses to the first author).

In total, we received 178 responses on our primary measure of interest. One response came from a participant who completed the survey twice (once after each e-mail). We omitted her second set of responses, yielding data from 177 unique participants ($M_{Age} = 52.9$, $SD_{Age} = 13.8$[1]; 42% female). Participants were from a wide range of institutions: 68% Research 1 (R1; or equivalent), 11% Research 2 (R2; or equivalent), 2% master's (or equivalent), 8% liberal arts colleges (or equivalent), and 12% were at other types of institutions. Many of the 12% who reported their institution as "other" were presumably from outside North America, where the institution ranks did not neatly map onto the provided options (e.g., R1, R2, etc.). Finally, participants represented the full range of academic rank: 8% were assistant professors, 25% were associate professors, 54% were full professors, 11% were professors emeritus, and 2% listed other ranks.

Procedure and Measures

Participants were told that the study was investigating factors that influence researchers' perceptions of potential job candidates. All participants viewed information about two hypothetical job candidates and then indicated which candidate they would prefer to hire. We randomly assigned participants to one of three experimental conditions. In the findings condition, participants were presented only with the number of statistically

significant experiments produced by each candidate over a 5-year span (see full text of all conditions in Table 1). This is comparable to the information one could glean from skimming a candidate's curriculum vitae. In the sample size condition, participants viewed the number of statistically significant experiments as well as the sample sizes employed by each candidate. This is comparable to the information that could be gleaned from reading a candidate's CV and skimming the methods sections of representative publications. This condition allowed us to empirically assess the degree to which power and productivity, respectively, shape hypothetical job hiring decisions. Finally, in the consequences condition, participants viewed the number of statistically significant experiments, the sample sizes employed by each candidate, and the expected published false positive and replication rates. Although this information is not typically available to search committee members, it can be easily calculated from the information provided in the sample size condition.[2] After being presented with the candidates, participants indicated which candidate they would prefer to hire.

Next, participants completed a brief demographic form, which asked about participant sex/gender, age, type of institution type (R1, R2, master's, SLAC, or other), and rank (assistant, associate, full, emeritus, or other). If participants selected other, they were given the opportunity to write in their institution type or rank.

In addition, participants were asked a number of questions about their own research practices for descriptive and exploratory purposes. Specifically, participants were asked to estimate the number of total participants they run per year, the typical effect size (Cohen's *d*) for their studies, and the typical sample size per condition in their studies. Finally, participants were asked to estimate how often they think their hypotheses are correct, independent of things like significance, Type I errors, and Type II errors. All exploratory analyses are in the Online Supplement. Upon completion of the survey, participants were thanked for their participation, asked not to discuss the study with colleagues, and fully debriefed on the purpose of the study.

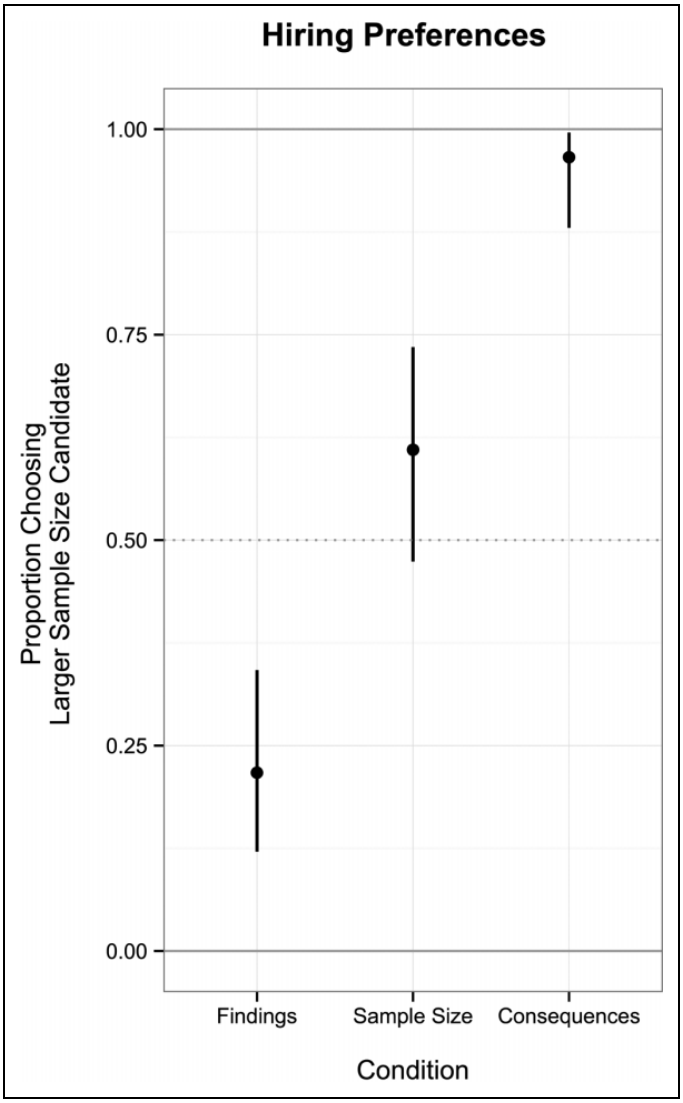

Figure 2. Participant preferences for adequately powered research (Researcher B) across the three experimental conditions. Point estimates and 95% confidence intervals (CIs) are presented. The gray dotted line represents .5, or no overall preference for either candidate.

Results

Confirmatory Analyses

We preregistered all hypotheses and primary analysis code before data collection commenced (https://osf.io/r7tk8/). We tested three primary hypotheses encompassing seven distinct statistical tests: (1) that overall preferences would differ across the three conditions, (2) that all conditions would differ from each other, such that preferences for the high sample size researcher would be lowest in the findings condition, intermediate in the sample size condition, and highest in the consequences condition, and (3) that in the findings and sample size conditions participants would prefer the low sample size candidate but preferences would flip to the high sample size candidate in the consequences condition.

Omnibus test. As hypothesized, preferences for the high sample size candidate differed across the three experimental conditions, $\chi^2(2, N = 177) = 68.64$, $p = 1 \times 10^{-15}$ (Figure 2).

Between-condition comparisons. Binary logistic regressions indicated that the specific pattern of between-condition differences was consistent with our hypotheses. Relative to the findings condition, participants in both the sample size and consequences conditions viewed the large sample size candidate more favorably, odds ratio (OR) = 5.66, 95% CI [2.58, 13.05], $p = 3 \times 10^{-5}$; $OR = 79.52$, 95% CI [23.05, 671.55], $p = 1 \times 10^{-9}$, respectively. Most importantly, participants in the consequences condition viewed the high sample size candidate more favorably than did participants in the sample size condition, $OR = 14.55$, 95% CI [4.36, 115.72], $p = .0001$.

Within-condition relative preferences. We used exact binomial tests (against proportion = .5) to assess relative preferences for each candidate within each condition. As hypothesized, in the findings condition, a substantial minority (21.7%, 95% CI [12.1%, 34.2%]) of participants preferred the high sample size candidate, $p = 1 \times 10^{-5}$. Contrary to our hypotheses, participants receiving information about both productivity and sample size (sample size condition) showed, on the whole, no strong preference for either candidate (61.0% preferred the high sample size candidate, 95% CI [47.4%, 73.4%]), $p = .12$. As hypothesized, participants in the consequences condition nearly unanimously (96.6%, 95% CI [88.0%, 99.6%]) preferred the high sample size candidate, $p = 1 \times 10^{-14}$. Only two participants still preferred the low sample size researcher when presented with information about expected replication and false positive rates.

Exploratory and Descriptive Analyses

In addition to testing preregistered hypotheses, we also ran some exploratory calculations on participants' self-reported research practices. Full details appear in the Online Supplement. Naturally, these values should be treated with caution, as they are based on self-reports of things that are inherently difficult to estimate. Nonetheless, median provided values were as follows[3]: effect size = .4, sample size per condition = 50, and correct hypothesis rate = .6. Based on the values participants provided to these 3 items, we were able to calculate each participant's expected power, expected false positive rate, and expected replication rates (both at 1 *N* and at 2.5 *N* for replication attempts). Median calculated values were as follows: power = .42, false positive rate = 7%, 1 *N* expected replication rate = 40%, and 2.5 *N* expected replication rate = 73%. Among our participants, 57% reported practices that would yield power lower than .5, and 83% reported practices that would yield power less than .8. Nearly two thirds (62%) reported practices that would yield an expected exact replication rate lower than 50%. One in five (21%) reported practices that would yield an expected 2.5 *N* replication rate less than 50%. Neither reported practices nor academic demographics (rank, institution type) had any detectable predictive effect on candidate preferences (see Online Supplement).

Discussion

As a thought experiment regarding the degree to which psychologists consider power when evaluating research, we assessed psychologists' relative preferences for productivity and power in hiring decisions. Participants preferred productivity in a sample size vacuum, were ambivalent when presented with both productivity and sample size information, but overwhelmingly favored adequately powered research when expected replication and Type I error rates were made explicit. This suggests that although researchers are intolerant of the negative consequences of underpowered research (consequences condition) they are more or less indifferent regarding the practices that logically lead to those very consequences (sample size condition). However, simply highlighting the easy to calculate negative consequences of low power nearly unanimously shifted preferences toward adequately powered research. Although our participants were social psychologists, we expect similar results from other subfields (Button et al., 2013; Vankov et al., 2014).

Clarifications and Concerns

Regarding the specific domain of hiring, one might object that a researcher boasting a 25% replication rate would not be able to produce a seemingly coherent and cumulative research program. This objection similarly holds during peer review, rather than hiring: Wouldn't an article with half a dozen underpowered conceptual replications of similar phenomena nonetheless be demonstrating reliable effects? Unfortunately, a package of conceptual replications creates a veneer of reliability and coherence in underlying effects, but packages of low-power conceptual replications provide less evidentiary value than one adequately powered study (e.g., Schimmack, 2012). Thus, a researcher operating at 25% power might produce numerous multistudy packages forming a coherent program of research, without necessarily producing solid, replicable science.

Admittedly, the stylized nature of our thought experiment likely generated some experimental demand. In today's psychological climate—with intense conversations concerning replicability—many researchers in the consequences condition may have felt compelled to choose the candidate with the higher replication rate. Although this may explain the near-unanimous preference for the adequately powered researcher in this condition, it does not explain why such a near-unanimous preference was not similarly present in the sample size condition, which included all pertinent information required to calculate expected false positive and replication rates. Thus, to facilitate the easy calculation of the consequences of power, we developed an online widget (http://tinyurl.com/PowerConsequences) that quickly calculates power, false positive rates, and replication rates (1 *N* and 2.5 *N*) from user-generated input.

The Downsides of Going Big?

The present study found that researchers were fairly intolerant of low-power research when the negative collective consequences of low-power research—driven by sample size decisions—were made clear. However, little consideration was given to potential negative side effects of running larger studies. We view potential downsides to going big in three primary domains: (1) the negative consequences of big studies will disproportionately be borne by researchers who access smaller populations, (2) running big studies may incentivize "easy" studies, and (3) requiring big studies might lead researchers to prioritize "safe" research ideas.

I. Big samples, small pools. First, placing emphasis on larger and larger studies carries obvious negative consequences for researchers with access to fewer participants overall. In order to maintain productivity while running 100 participants per experimental condition requires access to lots of participants. This concern is much more serious for researchers at institutions with smaller subject pools, researchers studying special populations, and researchers without resources to recruit larger samples. Plausibly, this could have long-lasting and deleterious consequences for researchers operating outside of large universities and researchers whose designs are not amenable for experimentation on undergrads (who thus do not have access to large pools), and for early career researchers (who thus probably have less financial support for participant recruitment). Of the three objections considered in this article, this one offers the fewest ready solutions.

II. Big samples, small methods. Requiring larger samples will provide an immediate disincentive to researchers seeking to use complex, involved, or expensive methods. Within social psychology, this may create an even more intense push away from studying actual behavior and toward the use of massive online labor markets (e.g., MTurk) as participant pools. Recruiting larger samples, in other words, may lead researchers to turn away from rich methods that provide key insights into the human condition in favor of "self-reports and finger movements" (see, e.g., Baumeister, Vohs, & Funder, 2007) in massive online markets.

III. Big samples, small advances. Given finite resources, running larger studies will necessarily entail running fewer studies. At the individual level, this means taking a hit to productivity. However, at the collective level this means that fewer and fewer new ideas will be tested. As a field, this may result in less new knowledge. In the present study, Researcher A generated not only more false positives than Researcher B; Researcher A would generate about 40% more *true* positives than Researcher B. Thus, a field consisting exclusively of Researcher B-minded individuals may be less progressive than a field consisting only of Researcher A types. To further compound things, as individual studies get progressively more challenging and time-consuming to run, researchers may begin testing "safer," more a priori probable hypotheses. Spending 50 participants on a risky new idea might not be painful, but spending 400 participants on a risky new idea is much more daunting.

Solutions? To partially address all three of the aforementioned challenges, it is worth noting that power does not solely depend on sample size, and many options are available for researchers aiming to increase power without running gigantic studies. Increased reliability of measures, stronger manipulations, and (especially) the incorporation of within-subject elements can yield much more powerful designs. For example, if Candidate A had used identical sample sizes but utilized within-subject designs, he or she would have garnered more than 40 published findings while maintaining low false positive rates (6%) and high replication rates (75% and 94% at 1 *N* and 2.5 *N*, respectively). In principle, it is possible to adapt many research paradigms typically dominated by between-subject research for within-subject designs (e.g., Francis, Milyavskaya, & Inzlicht, 2015).

The latter two challenges may in part be addressed by consciously and explicitly treating individual underpowered studies as merely suggestive and exploratory, but pairing them in multi-study packages with adequately powered replications. That is, a researcher may conduct a number of underpowered studies to test novel ideas or to use costly methods. Of those that produce significant results, some could then be replicated (directly when cost allows, conceptually for difficult methods) in adequately powered designs. While many multi-study packages superficially follow this template, often the single well-powered study is missing, and the package relies on number of studies, rather than strength of any individual study, to perhaps falsely argue for robustness (Schimmack, 2012).

Finally, it is important that researchers balance novelty with rigor. A psychological science consisting solely of large, well-powered, incremental extensions of well-trodden literatures is a somewhat dreary prospect. Science progresses not only through minimizing Type I errors and increasing replicability but also by the generation and testing of bold, novel hypotheses. However, these two components of a healthy science face some inherent tension. Bold, novel hypotheses are a priori less likely to be true than are more incremental extensions (else they would by definition not be bold, novel, or surprising). And, as the calculations we present make plain, a priori less probable hypotheses are also more likely to produce false positives. Thus, although it is tempting to use small sample studies to test lots of novel ideas (Objection III), these are precisely the ideas that demand the strongest supporting evidence. Adequate power should not be seen as an impediment to bold new research programs, but rather a necessary cost individual researchers must pay before others take their bold ideas seriously. Moving forward, if a phenomenon is worth claiming, it is also worth supporting with adequately power.

Coda

Much like increased transparency in research (e.g., Miguel et al., 2014; Simmons et al., 2011), increased power can also strengthen psychological science as a cumulative, repeatable enterprise (e.g., Asendorpf et al., 2013). But these changes require changes to existing incentive structures. Power needs to be rewarded at various levels. Academia consists of a series of bottlenecks (publication, employment, and tenure). Savvy researchers can adopt research strategies they view as likely to

promote advancement through these bottlenecks (e.g., Bakker et al., 2012; Vankov et al., 2014). This potentially creates an incentive mismatch between what is good for individual researchers and what is good for psychological science collectively (e.g., Nosek et al., 2012). Tversky and Kahneman (1971) were skeptical that any researcher would knowingly operate at 50% power. However, operating at 50%—or even much lower—power may be a rational strategy for a researcher seeking to maximize productivity. The present research suggests that this strategy will no longer be incentivized if psychologists carefully consider easily calculable consequences of power when deciding who and what gets through the various bottlenecks of academic success.

Acknowledgments

WMG developed the study concept, ran primary analyses, and wrote the first draft of the manuscript. JAJ took lead on generating our participant list, contributed to the study design, and contributed revisions. MBN and BKLN contributed to study design, implementation, analyses, and revisions. All authors approved the final article. We thank Dr. Catherine Rawn for advice on the design and final article.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Three participants provided implausible ages (0, 0, 11), and their age values were treated as missing data.
2. Acknowledging, of course, that committee members would have no way to assess the a priori probability of the candidate's hypotheses and could only guess underlying effect sizes from the published estimates in the candidate's articles.
3. Many participants provided a range (e.g., *N* from 30 to 50). All values reported here are derived from the *highest* value provided in the range.

Supplementary Materials

The online data supplements are available at http://spps.sagepub.com/supplemental.

References

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., . . . Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*, 108–119.

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

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

Bertamini, M., & Munafò, M. R. (2012). Bite-size science and its undesired side effects. *Perspectives on Psychological Science*, *7*, 67–71.

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

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145–153.

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

Cohen, J. (1992b). Statistical power analysis. *Current Directions in Psychological Science*, *1*, 98–101.

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science*, *1*, 140216.

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PloS One*, *9*, e109019.

Francis, Z., Milyavskaya, M., & Inzlicht, M. (2015). *Flipping the self-control switch: A novel within-subject paradigm to test the effects of ego depletion*. Poster presented at the 2015 SPSP conference, Long Beach, California.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., . . . Van der Laan, M. (2014). Promoting transparency in social science research. *Science*, *343*, 30–31.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*, 615–631.

Overall, J. E. (1969). Classical statistical hypothesis testing within the context of Bayesian theory. *Psychological Bulletin*, *71*, 285–292.

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*, 331–363.

R Core Team. (2012). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0*. Retrieved from http://www.R-project.org/

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

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2013). *Life after p-hacking*. Meeting of the Society for Personality and Social Psychology, New Orleans, LA, 17-19 January 2013. Retrieved from SSRN: http://ssrn.com/abstract=2205186 or http://dx.doi.org/10.2139/ssrn.220518

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*, 309.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*, 105–110.

Vankov, I., Bowers, J., & Munafò, M. R. (2014). On the persistence of low power in psychological science. *The Quarterly Journal of Experimental Psychology*, *67*, 1037–1040.

Author Biographies

Will M. Gervais is an assistant professor at the University of Kentucky.

Jennifer A. Jewell is a graduate student at the University of Kentucky.

Maxine B. Najle is a graduate student at the University of Kentucky.

Ben K. L. Ng is a graduate student at the University of Kentucky.