

All these omissions are understandable in a field that has been focused on rewarding positive results. Why slow down your research with large samples, pre-testing, and so on? The obvious reason is to reduce the occurrence of false results, both positive and negative. But as I have argued, a way of doing research has taken hold that takes selective reporting of positive results for granted. Some even believe that finding significant results with unreliable methods means that the results are extra-strong and trustworthy, given the blurry microscope they are glimpsed through. But the concern is that the blurry microscope also runs the risk of letting us see results that aren't actually there, in the manner of a Rorschach test.

New standards?

The past few years in psychology have seen the crisis of confidence in the p -value reach a tipping point. Realizing that the p -value is not a truth-value, we need a clear criteria to replace it in judging the experimental evidence for a proposition. A number of journal editors have proposed to reduce or eliminate our reliance on p -values in favor of another kind of inferential statistics: confidence intervals, Bayesian, or none at all. These realizations have also increased the call for new criteria based on statistical power or sample size. However, in my view, the alternatives do not fully take into account the state of development of the theory and methods of social psychology.

Bayesian prior probabilities. As Ioannidis (2005) has argued, to turn the p -value into a truth value we need to also know the probability that the null hypothesis, as opposed to an alternative hypothesis, is true in the population. This is because, following Bayesian

statistical logic, the p-value only tells us about the unlikeliness of the observed result if the null hypothesis holds in reality. An interactive visualization of these arguments and formulae can be found online (Schönbrodt, 2014; <http://shinyapps.org/apps/PPV>). Playing with various values, it becomes clear that the prior probability that a hypothesis is true outstrips statistical power in determining the truth-value of a given p-value.

For example, let's assume that our hypothesis is only directional (the experimental group has a higher score than the control group in the population) and that it has a 50% "coin flip" chance of being true. At 80% power, the chance that a positive result at $\alpha = .05$ is true (positive predictive value, or PPV) is 94%. At 35% power, we have less confidence in the truth of that result; the PPV is now 87.5%, but this still means that the great majority of reported results are true. It is only when we turn to testing hypotheses with a very low likelihood of being true that we can agree with Ioannidis' titular claim that "most [positive] published research findings are false." With a hypothesis that is only 5% likely to be true, even 80% power yields a PPV just below 50%, and 99% power barely gets us to a PPV of 51%. Clearly, to be sure of the truth value of our positive findings, power alone will not save us. We have to see how likely they were in the first place. This in turn involves their proposed effect size; for example, the prediction that a tiny manipulation will lead to at least a tiny directional change is categorically more plausible than the prediction that the same manipulation will shift the outcome variable by one or more standard deviations. And of course, estimating effect size *a priori* is necessary to obtain power calculations in the first place.

But the need to consider prior probabilities seems profoundly *unfair*, especially to a discipline that has grown up around the question of social prejudice. Surely every hypothesis deserves a fair trial on the basis of the evidence, regardless of what you might think beforehand? And yet, not all prejudices are irrational (Crandall & Eshleman, 2003). Our everyday decision-making, too, takes prior knowledge into account all the time. The real problem in accepting any kind of Bayesian criterion is our inability to quantify and agree upon the likely size of a novel experimental effect a priori. The vast majority of theories in social psychology continue generate only directional hypotheses, as Meehl (1967) argued. Without knowing on theoretical grounds the size of an entirely novel effect, we cannot calculate the chance of a Type II error, or properly specify the likelihood of any given observed result. Therefore, Bayesian and power calculations are most solid when following directly on already-obtained effects, such as when replicating other labs' experiments or confirming exploratory research. At other times, they have to proceed on the basis of a "typical" effect size for social psychology, or - in the case of Bayesian analyses - draw on non-informative or weakly informative priors (Gelman, 2009).

Effect size expectations. An often-cited effect size benchmark for social psychology is the conventionally small-to-medium mean effect size of $r = .21$ found by a meta-analysis of meta-analyses in the social psychology literature (Richard, Bond & Stokes-Zoota, 2003). But in these meta-analyses we can see great variety in the questions asked, some irrefutable ("impressions are based on individuals' special characteristics", $r = .69$) and others debunking popularly held myths ("subliminal advertising increases sales", $r = .00$). Few of these meta-analysis topics correspond to the vital questions increasingly asked by

experimental social psychologists over the past twenty years about the influence of mindsets, implicit processes, schemas, emotions, and goals. These topics, in fact, are often studied in idiosyncratic ways that don't lend themselves readily to meta-analysis.

Also importantly, the same question can be studied by many methods, and it is little appreciated that method can contribute greatly to observed effect sizes. For example, Abelson (1985) found that individual differences between batters contributed less than 1% of variance to the success of a single at-bat in Major League Baseball; but when studied as a more reasonable "repeated measures" design across multiple games and seasons, person factors assume their intuitively enormous role and explain nearly 49% of year-on-year variance in batting averages (Giner-Sorolla, unpublished, 2015). Richard et al. (2003) likewise found evidence that method and context factors contributed to variance in effect sizes, beyond the variance expected from mere resampling. More mature research topics paradoxically showed weaker overall effects, as experiments increasingly sought boundary conditions that would negate or reverse the finding. Should the purpose of research be to establish a method that maximizes basic effect sizes, or to find moderators and confounds that will tend to reduce them overall (and, as interaction effects, will have their own hard-to-predict effect sizes?) This open question muddies the interpretation of existing effect sizes.

Effect sizes in basic research have a more fundamental problem. They are almost always based either on abstract measures (e.g., self-report scales, relative reaction times), or on concrete measures embedded in an abstract context, such as money earned in an economic

game. Thus, effect sizes and related measures such as confidence intervals are hard to interpret. When comparing two conditions' responses on a seven-point mood scale, what does it mean to have a standardized effect size lower bound of .20, as opposed to .30? The fact that the larger lower bound is more strongly different from zero can also be conveyed by reporting the exact p-value. But comparisons to meaningful non-zero values are unlikely. The abstract nature of effect sizes in basic research makes it hard to use them in calculating cost-benefit tradeoffs, for example, or in relation to clinical criteria.

To be sure, effect size statistics are useful, and should be reported alongside significance. This makes the job of meta-analysis easier, and allows comparisons between multiple effects if they are methodologically similar. Effect sizes also are more meaningful when we measure concrete units in an ecologically valid setting, such as school test scores or clinically relevant levels of stress-related hormones. But these are not sufficient arguments to abandon p-values in favor of confidence intervals or Bayesian statistics. Considering social psychology's directional hypotheses, and our often basic and abstract measures, a precisely reported and intelligently interpreted p-value can summarize directional evidence, balancing the magnitude, variability, and sample size of an effect.

Explaining our new guidelines

With this background in place, let me now annotate the new set of editorial policies and guidelines that can be found online at <http://www.journals.elsevier.com/journal-of-experimental-social-psychology/news/jesp-editorial-guidelines/>, replacing the guidelines from 2014.

1. *All manuscripts must include a report of the following items:*

- *Explain how sample size was determined, including whether and how looking at the results influenced the collection of additional data. Reporting statistical power (together with the basis for expected effect size) is welcome, but not required.*
- *If participants were excluded from the study, report how many, how they were distributed across conditions, and the reasons for exclusion.*
- *Disclose the existence of all variables and conditions that were part of the study. These can be summarized, or put in a footnote or supplementary material in the case of large numbers of variables, but there should be enough information to judge whether they are potentially theoretically relevant or not.*
- *Affirm the above two disclosures positively (that is, with a statement such as “We report all measures, manipulations, and exclusions in these studies.”)*
- *Report procedures in sufficient detail to allow close replication by an independent lab. This may include citations of method details found in other published, peer-reviewed (or open access) sources. Materials are not required to be provided verbatim, but should be summarized with examples. To stay within the word limit of a report, necessary details can be included in Supplementary Materials.*

These are the same requirements as before, with one change: papers will now have to include a positive affirmation that they have followed the previous requirements. This fits the goal to reduce the selective analysis of results and increase acceptance of a realistic

picture of evidence. Authors should also note that the requirement to justify sample size does not rule out using fully reported, statistically corrected methods of sequential sampling (Lakens, 2014). This appears to be a promising way to establish a first estimate of an initially unknown effect size, avoiding much of the uncertainty around a priori power analysis.

2. All manuscripts should report complete statistics relevant to the analyses at hand, using supplementary materials if needed:

- *cell means, SD and n for experimental designs*
- *correlations between variables for multivariate designs including regression and repeated-measures*
- *inferential statistics with exact p-values and effect sizes regardless of significance level.*
- *If figures use error bars, these should be explained in a caption (e.g., standard error, 95% confidence interval, etc.)*
- *If meeting any of these requirements proves impractical, the authors should explain why.*

These requirements are expanded from the previous requirement to report effect size, because basic descriptive statistics are often critical to evaluating results (for example, examining multicollinearity in multivariate analyses, or floor or ceiling effects in variables). The importance of exact *p*-value has already been explained, while error bars are often misunderstood or underreported (Belia, Fidler, Williams & Cumming, 2005). Often excluded from manuscripts in the interest of brevity, basic descriptive statistics can be

included in Supplemental Materials, and will also be useful going forward in informing meta-analyses and other curations of the empirical literature.

3. *FlashReports will now follow similar standards of evidence and potential topics as the other forms of articles, but will still be distinguished by a brief 2500 word limit and streamlined review process. The short format encourages research reports with background ideas and methods that do not require a great deal of explanation, but it does not mean that papers should be “short” on theoretical relevance or amount of evidence. Method and analysis details not crucial to the narrative of the paper should still be reported in Supplementary Materials.*

4. *Soliciting reviews of revised manuscripts is now the exception, not the rule. Associate Editors will typically come to a decision on a revised manuscript based on its capacity to address the reviewers’ comments, unless expert opinions on technical matters are required. We expect this will make the review process faster and more efficient.*

As explained, both these policies are part of the move to make the review process faster without compromising on quality.

5. *Our standards for articles rest on a tripod of strong theory, strong methods, and strong statistical evidence for the points being made. Deficiencies in one of these areas can, to some extent, be compensated by strengths in the other two.*

Strong theories and strong methods - in terms of statistical power, consistency between studies, and clear distinction between exploratory and confirmatory tests - can improve the plausibility of reported results. A theory that convincingly explains unexpected findings, in fact, actually increases their subjective truth value; findings that are published merely because they are surprising, because they start from a low prior probability, are intrinsically less convincing. It is understood that a good paper for JESP will provide evidence not only that something happens, but also make at least a first attempt at explaining why. Strong theoretical relevance -- for example, a series of studies systematically pitting one theory's explanation of an effect against another -- is a highly desirable trait as well. If competing theories are well characterized and implemented, aggregate results need not be as strongly positive as when an author makes a one-sided theoretical argument.

The following guidelines (a-d) further explain how qualities of reported results can lead to a conclusion of strong statistical evidence.

- a. *In particular, our view of strong statistical evidence departs from the previous unwritten standard of "multiple studies, each with a significant effect." Instead, strong statistical evidence for a central positive claim now rests more upon showing a plausible pattern of results, usually from multiple comparable studies, whose total joint probability under the null would be about $p = .01$ or less (e.g., using Fisher's method to aggregate p values from results in the same direction, or meta-analysis otherwise). However, we emphasize that this view **is not a strict criterion**. It does not prohibit publication of less strong results if theory and methods are both strong, nor*

is it a guarantee of publication if the article has other limitations. For example, although we continue to discourage submission of single-study articles because they often do not provide enough development and evidence for an idea, we are willing to consider them if based on good theory and exceptionally strong methods (possibly including: high power to detect a small-to-medium effect size, pre-registered methods and analyses, unusual or highly representative samples).

- b. Within an article, individual studies with good theoretical and methodological connections to the others need not be statistically significant if they contribute to a strong overall pattern, a precise estimate, and/or a complete and open reporting of a program of research. Article-wise meta-analytic estimates are encouraged to aggregate the findings of multiple studies.*

Following from the discussion of the nature and likely distribution of p-values, these items describe a new way to think about significance that roughly matches the previous implicit standard. For example, using Fisher's method, two independent results in the same direction at $p = .04$ have a joint value about $p = .011$. However, aggregate results that fall short of this standard can still be considered if they are theoretically well-grounded and arise from strong and systematic methodology. I encourage authors to take a cumulative approach to summarizing their central findings across studies while also addressing their consistency, rather than feeling the need to only report results that look significant.

c. Bayesian analyses with well-justified and well-explained priors can be used to assess evidence for reasonable null and alternative hypotheses.

Interesting answers to many questions can come as non-effects (such as “no gender difference in this experimental effect” or “no effect of subliminal advertising on consumer behavior”) but the technique of arguing such a case statistically is not always known. Confidence in methodology that could have detected any reasonable effect “out there” should be high. Also, the question of which range of effect sizes should be treated as null needs to be considered as well. I believe that with these prior conditions, Bayesian analyses offer one way to balance evidence for a null and alternative hypothesis against each other, although the alternative hypothesis also needs to be specified with a reasonable range; overly strong alternative effect sizes tend to bias conclusions in favor of the null.

d. A pattern of multiple central results that are all below, but close to, $p = .05$, although they might have arisen by chance, also can be said to show low statistical robustness. Ideally, confidence in such a set of results can be bolstered by pre-registered studies or other methodological assurances against selective reporting.

Again, this is not a strict requirement but one of many considerations. I do realize that at a certain level of selective reporting, even an aggregate estimate becomes unreliable.

While it has not been worked out yet in all its details, the p-curve and similar methodologies offer a simple insight: given the usual variability of p-values in a true effect, it is unlikely that a series of independent tests specified a priori would yield

results all slightly under $p = .05$. Such results may be given greater scrutiny in terms of methodology and reporting. But rather than apply a hard-and-fast forensic rule, or require a lab-wise disclosure statement that would run into many problems defining the scope of a research line, I choose to start by encouraging authors to report consistent analyses and methods that support a larger picture, even if some individual results are not strictly significant. To this end, I also want to encourage authors to submit research articles based in part or fully on pre-registered studies; and to submit methods articles that develop, test, validate or criticize techniques for making decisions about the plausibility and robustness of a set of reported findings.

e. As before, we welcome rigorously conducted replication articles that meet the criteria described in Brandt, Ijzerman, et al. (2014).

This guideline has not changed. If I could offer one further piece of advice for replication attempts, it is to consider carefully the pretesting, calibration, and validation of stimuli and contexts for the new sample, rather than just assuming that a duplication of the original procedure will stand as valid (cf. Fabrigar & Wegener, in press). This is consistent with the reasons why original research should also pay more attention to calibration and validation.

6. Authors are encouraged to present exploratory work openly. It is deceptive to present hypotheses as perfectly precognitive when they aren't (i.e., avoid HARKing; Kerr, 1998). It is OK for authors to admit they entertained multiple hypotheses, based on

multiple theoretical perspectives, without coming down on the side of any one of them – or even to admit that they came down on the side of the wrong one! Put another way: an initial study, with an uncertain or novel prediction, should be treated as exploratory regardless of what story the authors choose to tell. Confidence in the results depends on sound methods, sample size, and consistent replication. Honest distinctions between exploratory and confirmatory work can be reported concisely without turning it into a long-winded “intellectual odyssey.”

Research with strong and well-tested methods, like high-powered research, gives more credible positive findings. It also allows for evaluation of “failed” studies in a principled way, rather than assuming that their methods were faulty merely because they failed to support the preferred hypothesis (LeBel & Peters, 2011). To this end, I encourage brief reporting (perhaps in footnotes or supplementary material) of interpretable failures or minor procedural tests along the way. Also, pre-registration of hypotheses and analyses offers one way to distinguish confirmatory from exploratory work. I urge those interested to watch for the upcoming special issue of JESP on Confirmatory Research for more examples and commentary on this practice.

7. *Although mediation analyses are used in many, if not most, recent JESP articles, we urge greater caution in using and interpreting this technique (cf. Fiedler, Schott & Meiser, 2011; Kline, 2015; Spencer, Zanna, & Fong, 2005). As before, we see little value in mediation models in which the mediator is conceptually very similar to either the predictor or outcome. Additionally, good mediation models should have a*

methodological basis for the causal assumptions in each step; for example, when the predictor is a manipulation, the mediator a self-reported mental state, and the outcome a subsequent decision or observed behavior. Designs that do not meet these assumptions can still give valuable information about potential processes through correlation, partial correlation, and regression, but should not use causal language and should interpret indirect paths with caution. We reiterate that mediation is not the only way to address issues of process in an experimental design.

I have withdrawn the previous policy's suggestion, which was to test mediation models without clear causal direction by "reversing" the mediator and outcome to see which model seems stronger. Since then, it has been persuasively argued that this technique is influenced by artefacts such as measure reliability and does not necessarily indicate true underlying patterns of data (Lemmer & Gollwitzer, 2016; Thoemmes, 2015). Many authors apparently think that mediation is a necessary requirement to be published. We want to dispel this misconception, and encourage alternate ways to show the role of third variables in experimental effects. A significant indirect path, where all three variables are measured at the same time and have no clear causal logic to them, can merely tell us that the "mediator" is not irrelevant to the "predictor," at the same time that it is not entirely redundant with the "predictor" in its relationship to the "outcome."

8. *Arbitrary use of covariates can be used to engineer significant results. Therefore, covariates need to be justified as a response to a clear threat to validity. Reporting the uncovariates analysis can help clear this up (e.g., in a Footnote).*

9. *Interaction effects on their own are not sufficient; they must be decomposed with simple effects analysis (Aiken & West, 1991) or other means. At the same time, the direction and significance of simple effects are influenced both by interactions and main effects; therefore, it is not always necessary to “validate” an interaction by showing that both simple effects are significant (Petty, Fabrigar, Wegener, & Priester, 1996).*

These two points are unchanged from the 2014 guidelines.

What do the new standards look like in action? Perhaps one paper that I handled in my last year as Associate Editor can stand as a good example. Subscribers to the journal will be able to find Bostyn and Roets (in press), available in an online preprint version prior to the release of the March 2016 issue. In its final form, this article presents three studies, with somewhat different individual results, that give a strong and theoretically well-supported “big picture” of an asymmetry in the action-omission effect for positive and negative moral judgments. Pre-registered internal replication, aggregate meta-analysis across studies, and a Bayesian analysis using well-reasoned priors to support negative results in a particular condition, all form part of the new look of evidence. Going forward, I hope to see other manuscripts that take this approach to multi-study research.

In conclusion, I am sure that this Editorial will not be the last word during my term at JESP. I have learned so much from my own experience and the arguments of others already, and am prepared to learn more over the next three years. As always, before more change happens, I will endeavor to make sure it is understood and approved by the whole editorial team; and when it happens, I will be sure to let you know, online or in print.

Roger Giner-Sorolla

January, 2016

References

Abelson, R. P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin*, 97(1), 129–133.

American Psychological Association (2010). *Publication manual of the American Psychological Association* (6th ed.) Washington, DC: American Psychological Association.

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

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, 10(4), 389–396.

Bishop, D. V., & Thompson, P. A. (2015). Problems in using p-curve analysis and text-mining to detect rate of p-hacking (No. e1956). *PeerJ PrePrints*. Retrieved from <https://peerj.com/preprints/1266>

Bostyn, D. H., & Roets, A. (2016). The morality of action: The asymmetry between judgments of praise and blame in the action–omission effect. *Journal of Experimental Social Psychology*, 63, 19–25. Retrieved from <http://www.sciencedirect.com/science/article/pii/S0022103115300172>

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., ... Van't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, 50, 217–224.

Coulson, M., Healey, M., Fidler, F., & Cumming, G. (2010). Confidence intervals permit, but do not guarantee, better inference than statistical significance testing. *Quantitative Psychology and Measurement*, 1, 26.

Crandall, C. S., & Eshleman, A. (2003). A justification-suppression model of the expression and experience of prejudice. *Psychological Bulletin*, 129(3), 414–446.

Fabrigar, L. R., & Wegener, D. T. (in press). Conceptualizing and evaluating the replication of research results. *Journal of Experimental Social Psychology*.

Gabriel, K. R. (1978). A simple method of multiple comparisons of means. *Journal of the American Statistical Association*, 73(364), 724–729.

Gelman, A. (2009). Bayes, Jeffreys, prior distributions and the philosophy of statistics. *Statistical Science*, 24(2), 176–178.

- Giner-Sorolla, R. (2015). Effect size depends on methods. Unpublished manuscript, University of Kent. Retrieved from <https://osf.io/w4enh/>.
- Hewstone, M., & Stroebe, W. (2008). Moving at snail's pace: Some observations on the publication process in social and personality psychology. *SPSP Dialogue*, 23(1), 17–24.
- Hung, H. M. J., O'Neill, R. T., Bauer, P., & Kohne, K. (1997). The behavior of the p -value when the alternative hypothesis is true. *Biometrics*, 53(1), 11–22.
- Hurlbert, S. H., & Lombardi, C. M. (2009). Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neo-Fisherian. *Annales Zoologici Fennici*, 46, 311–349.
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124.
- Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, 103(1), 54–69.
- Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, 44(7), 701–710.
- Ledgerwood, A., & Sherman, J. W. (2012). Short, sweet, and problematic? The rise of the short report in psychological science. *Perspectives on Psychological Science*, 7(1), 60–66.
- Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association*, 88(424), 1242–1249.
- Lemmer, G. & Gollwitzer, M. (2016). The “true” indirect effect won't (always) stand up: When and why reverse mediation testing fails. Unpublished manuscript, Philipps-University Marburg.
- Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 34, 103–115.
- Murayama, K., Pekrun, R., & Fiedler, K. (2013). Research practices that can prevent an inflation of false-positive rates. *Personality and Social Psychology Review*, 18(2), 107–118.
- Richard, F. D., Bond Jr, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, 7(4), 331–363.
- Schönbrodt, F. (2014). When does a significant p -value indicate a true effect? Web page, retrieved from <http://shinyapps.org/apps/PPV>.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, 143(2), 534–547.

Spellman, B. A. (2015). A short (personal) future history of Revolution 2.0. *Perspectives on Psychological Science*, 10(6), 886–899.

Stanley, D. J., & Spence, J. R. (2014). Expectations for replications: Are yours realistic? *Perspectives on Psychological Science*, 9(3), 305–318.

Thoemmes, F. (2015). Reversing arrows in mediation models does not distinguish plausible models. *Basic and Applied Social Psychology*, 37(4), 226–234.

Ulrich, R., & Miller, J. (2015). p-hacking by post hoc selection with multiple opportunities: Detectability by skewness test?: Comment on Simonsohn, Nelson, and Simmons (2014). *Journal of Experimental Psychology: General*, 144(6), 1137–1145.