This manuscript is "in press" at
Psychonomic Bulletin & Review

The Frequency of Excess Success for Articles in *Psychological Science*

Gregory Francis

Department of Psychological Sciences

Purdue University

phone: 765-494-6934

gfrancis@purdue.edu

Running head: Excess Success in Psychology

Abstract

Recent controversies have questioned the quality of scientific practice in the field of psychology, but these concerns are often based on anecdotes and seemingly isolated cases. To gain a broader perspective, this article applies an objective test for excess success to a large set of articles published in the journal *Psychological Science* between 2009-2012. When empirical studies succeed at a rate much higher than is appropriate for the estimated effects and sample sizes, readers should suspect that unsuccessful findings were suppressed, the experiments or analyses were improper, or that the theory does not properly account for the data. The analyses conclude problems for 82% (36 out of 44) of the articles in *Psychological Science* that have four or more experiments and could be analyzed.

It is widely recognized that there is a bias *across* articles in the field of psychology (Fanelli, 2010; Sterling, 1959; Sterling, Rosenbaum & Weinkam, 1995). These studies noted that approximately 90% of published experiments are reported to be successful, which suggests that there must be many unsuccessful experiments that remain unpublished. However, it is not clear what such a bias means with regard to believing a specific reported experimental finding or theory. Bias across articles can arguably just reflect a desire among authors and journals to publish about topics that tend to reject the null hypothesis with typical experimental designs; and such a bias does not necessarily cast doubt on the findings or theories within any specific article. When judging the quality of scientific work, a finding of bias *within* an article is more important than bias across articles because the presence of bias within an article undermines that article's theoretical conclusions. Recent investigations (Bakker, van Dijk & Wicherts, 2012; Francis, 2012a-e, 2013a,b; Renkewitz, Fuchs & Fiedler, 2011; Schimmack, 2012) have used an objective bias analysis to indicate that some articles (or closely related sets of articles) in the field of psychological science appear to be biased. However, these individual analyses do not indicate whether the appearance of bias within an article is rare or common in psychology.

Partly to estimate the within article bias rate, I have applied the bias analysis to articles published over the last several years of the journal *Psychological Science*, which is the flagship journal of the Association for Psychological Science, has enormous reach to scientists and journalists, and presents itself as an outlet for only the very best research in the field. Although perhaps the journal is not representative of the field of psychological science in general, it would be valuable to know what proportion of findings (and which specific findings) appear to be biased in a journal that seeks to publish the field's best work.

This article summarizes the analyses of the investigated articles in the journal; article selection criteria and a full description of the analyses (and accompanying computer code) are provided in the supplemental material.[1]

In lay usage, the term "bias" means unfair prejudice, but that is not the intended meaning in this article. In this article, the term bias is used in a statistical sense, namely that the frequency of producing a significant result, or the effect's magnitude, is systematically overestimated. A prejudicial bias by authors may produce a statistical bias, but it is not necessary because statistical bias can be introduced despite good intentions from researchers. Moreover, it is not necessary to know the exact cause or source of statistical bias for a reader to be skeptical about published empirical findings and their theoretical conclusions.

Analyzing the Probability of Experimental Success

The bias analysis is based on the "test for excess significance" (TES) proposed by Ioannidis and Trikalinos (2007). The test contrasts estimates of experimental power with the reported frequency of significant findings. Suppose an article describes five experiments that are presented as empirical support for a theory. The experiments could be direct replications, conceptual replications, converging evidence, or explorations of different parts of the theory. Further suppose that every experiment produces a successful outcome (e.g., rejects the null hypothesis). Such a positive outcome might commonly be

[1] For review purposes go to http://www1.psych.purdue.edu/~gfrancis/Publications/PsychScience/ and enter "PSYCH" (all caps) and "science" to access the material.

considered vindication for the theory, but this view is unwarranted without considering the statistical power of the experiments.

Due to random sampling, even when an effect is non-zero some experiments will generate samples that produce test statistics and *p* values that do not satisfy the criterion for statistical significance. Power is the probability of rejecting the null hypothesis for a specified non-zero effect. Suppose that the power for each of five reported experiments is .6. Following Ioannidis and Trikalinos (2007), the probability that all five of such experiments would reject the null hypothesis is the product of the power values, $P_{TES} = .6^5 = .078$, which indicates that the observed pattern of consistent rejections should be rare. A common criterion in these kinds of investigations is a probability of .1 (Begg & Mazumdar, 1994; Francis, 2012a; Ioannidis & Trikalinos, 2007). A power probability below this criterion suggests that the experiment set should be considered biased, and the results and conclusions are treated with skepticism.

The TES is sensitive to the distribution of reported statistics in a way that is similar to other tests for publication bias (Egger, Davy Smith, Schneider & Minder, 1997) and *p*-hacking (Simonsohn, Nelson & Simmons, in press). However, these alternative tests require a set of findings to have fixed effect sizes or independent *p* values, which is often not the case for a series of experiments in psychology. The TES's focus on experimental power allows it to consider inhomogeneous effect sizes and statistical dependencies between experimental outcomes, provided the appropriate statistical information about the dependencies is published.

The TES can be generalized to consider the probability of experimental success, including a pattern of significant and non-significant results. For the remainder of this

article, the term TES stands for the test for excess *success*, which has the same acronym and follows the same principles as the test for excess significance that was proposed by Ioannidis and Trikalinos (2007). If experimental success is defined as statistical significance, then the tests are identical. Table 1 describes the properties of a hypothetical article with five experiments that were interpreted as providing unanimous support for a theoretical conclusion. The middle two columns of Table 1 summarize the statistics and hypothesis tests that were used to support the theoretical claims. Although the statistics and analyses reported in Table 1 are artificial, they reflect the kinds of statistics, hypotheses, and analyses that are common in *Psychological Science* articles. For all of the estimated success probabilities, it was assumed that the original hypothesis tests were appropriate for the reported data (e.g., the data were randomly sampled from normal distributions with a common variance).

The first row describes the properties of Experiment 1, which was a between-subjects design having three different groups. A successful outcome was for every pair of means to be significantly different. The reported statistics satisfy this definition of success (all p's less than .05), using an ANOVA and contrasts. The last column of Table 1 provides the *post hoc* estimated probability of success for this type of experiment with these samples sizes. The probabilities are calculated by supposing that the reported statistics accurately reflect the population parameters. One hundred thousand simulated experiments with the same sample sizes were then generated by drawing random samples from the simulated populations, running the hypothesis tests, and computing the proportion of tests that satisfy the desired outcome (rejecting the null hypothesis). The ANOVA and the contrast between conditions 1 and 3 are very likely to reject the null. However, the

contrasts between conditions 1 and 2 and between conditions 2 and 3 have only modest power values. Moreover, the joint probability of all four hypothesis tests being successful is only 0.435. This low probability reflects the multiple constraints that are imposed by the definition of the experiment's success.

Table 1: Statistical properties, hypotheses, and estimated probability of success for a hypothetical set of five experiments.

	Statistics	**Hypotheses**	**Probability of success**
Exp. 1	n_1=15, n_2=16, n_3=19	ANOVA	.996
	$\overline{X_1}$=17.2, $\overline{X_2}$=15.3, $\overline{X_3}$=13.4	$\mu_1 \neq \mu_2$	.674
	s=2.15	$\mu_1 \neq \mu_3$	.999
		$\mu_2 \neq \mu_3$	.723
		Joint	*.435*
Exp. 2	n_1=23, n_2=22, n_3=20	$\mu_1 \neq \mu_2$	.934
	$\overline{X_1}$=5.5, $\overline{X_2}$=4.6, $\overline{X_3}$=4.9	$\mu_1 \neq \mu_3$	.617
	s=0.85	$\mu_2 = \mu_3$	.801
		Joint	*.532*
Exp. 3	n_1=25, n_2=25	$\mu_{X1} \neq \mu_{X2}$	.779
	$\overline{X_1}$=5.5, $\overline{X_2}$=4.3, s_X=1.5	$\mu_{Y1} \neq \mu_{Y2}$	.673
	$\overline{Y_1}$=11, $\overline{Y_2}$=8, s_Y=4.25	Joint	*.673*
Exp. 4	n_1=35, n_2=36	$\mu_{X1} \neq \mu_{X2}$	.672
	$\overline{X_1}$=24.1, $\overline{X_2}$=20.1, s_X=6.9	$\mu_{Y1} \neq \mu_{Y2}$	.639
	$\overline{Y_1}$=33.1, $\overline{Y_2}$=29.3, s_Y=6.8	Joint	*.460*
	r_{XY} = 0.23		
Exp. 5	n_{1A}=56, n_{2A}=53, n_{1B}=52, n_{2B}=51	ANOVA interaction	.615
	$\overline{X}_{1A}$=15.0, $\overline{X}_{1B}$=20.5	$\mu_{1A} \neq \mu_{1B}$	.989
	$\overline{X}_{2A}$=14.2, $\overline{X}_{2B}$=15.5	$\mu_{2A} = \mu_{2B}$	.836
	s=6.75	Joint	*.587*
P_{TES}			.042

Experiment 2 was a between-subjects design with condition 1 corresponding to an experimental condition and conditions 2 and 3 acting as controls. A successful outcome was that the experimental condition would differ from each of the controls, which would

not differ from each other. For the given statistics, all of the tests are successful. The last column indicates the probability of success (estimated with one hundred thousand simulated experiments). For the comparison of the control conditions, success is defined as producing a non-significant result. Although the probabilities are relatively high for two of the tests, the joint success probability is only slightly above one half. Again, the multiple constraints on the pattern of results, including a predicted non-significant result, reduce the probability of experiment success.

Experiment 3 was a between-subjects design that measured two variables, *X* and *Y*, for an experimental group and a control group. The predicted outcome was that both variables would produce a significant difference between groups. The power value varies across the two tests, and with the given statistics, it is not possible to directly estimate the joint power. If the two measured variables were independent, the joint power could be estimated by multiplying the powers of the separate variables. However, because the *X* and *Y* scores are correlated (they come from a common set of participants), this product will likely underestimate the true power. A conservative approach, which likely overestimates the true power, is to take the lower of the two power values as an upper limit on the joint power.

Experiment 4 had a design similar to Experiment 3, in that it measured two variables for both an experimental and control group and in that the predicted outcome was for both variables to reject the null. Unlike Experiment 3, the reported statistics include the correlation between the two measures, and knowing the correlation enables the joint power to consider the probability of both studies producing a significant outcome. The estimated

joint power, derived from simulated experiments, is smaller than either of the individual powers but larger than their product.

Experiment 5 had a two by two between-subjects design and the expected outcome was a significant interaction coupled with a significant difference between scores in condition 1 but a non-significant difference between scores in condition 2. The probability of success is mostly dominated by the probability of the significant interaction, with the other tests only modestly reducing the success probability.

The probability of five experiments like these producing uniformly successful outcomes is the product of the joint success probabilities, which is 0.042. This probability is so low that readers of such an article should be skeptical that the experiments were run properly, analyzed properly, and interpreted correctly relative to the theoretical ideas.

The logic of the TES analysis is similar to traditional hypothesis testing, where the null hypothesis is that the experiments were run properly and without bias. With that premise, one estimates the probability of producing successful experimental outcomes that are equivalent, or more extreme, than the observed outcomes. If this probability is low, then the analysis suggests that the null is not viable: the set of experiments appears to be biased or flawed in some way. As Francis (2013b) noted, the analysis checks for consistency of outcomes across a set of experiments. If experiments are run properly, analyzed properly, and published fully, then the rate of success should be consistent with the estimated probabilities of experimental success. It is also appropriate to think of P_{TES} as the estimated probability that a direct replication of a set of experiments with the same sample sizes would produce results at least as successful as those that were published.

The TES analysis deviates from traditional hypothesis testing in that the .1 criterion for P_{TES} does not define the frequency of rejecting the null hypothesis when it is really true. Precise control of the Type I error rate for a judgment of bias requires knowing the true probability of experimental success, but almost every practical application of the TES analysis must estimate experimental success using the reported data. Francis (2013b) showed that for two-sample *t* tests, using the $P_{TES} \leq .1$ criterion usually produces a Type I error rate of around .01, so the test is conservative.

Applying the TES Analysis to Articles in *Psychological Science*

I downloaded all 951 articles published in *Psychological Science* during years 2009-2012. Articles were considered for the TES analysis only if the success probability could be estimated for at least four experiments, because the analysis needs at least that many experiments in order to have much chance of detecting any type of bias (see simulations in Francis, 2012a, 2013b). The count of experiments included subdivisions (e.g., Study 1a, 1b) and occasionally included an experiment that was described outside of a formal header (e.g., an experiment summarized in the conclusion section as a follow-up). There were a total of 44 articles with four or more experiments that could provide success probability estimates. The supplemental material further discusses article selection and describes the success probability calculations for each article. The supplemental material also includes source code for the simulation-based probability calculations.

Table 2 lists the P_{TES} value for each of the 44 analyzed articles. The most striking property is that 36 out of the 44 articles have a P_{TES} value smaller than the .1 criterion. Despite the conservative nature of the TES analysis, bias appears to be present for 82% of the articles in *Psychological Science* with four or more experiments having designs and

reported statistics that enable success probability calculations. This high rate is troubling because biased articles do not provide appropriate scientific arguments for their derived theories.

Table 2: Results of the TES analysis for each of forty-four articles in *Psychological Science*.

Year	Authors	Short title	P_{TES}
2012	Anderson, Kraus, Galinsky & Keltner	Sociometric Status and Subjective Well-Being	.167
2012	Bauer, Wilkie, Kim & Bodenhausen	Cuing Consumerism	.062
2012	Birtel & Crisp	Treating Prejudice	.133
2012	Converse & Fishbach	Instrumentality Boosts Appreciation	.110
2012	Converse, Risen & Carter	Karmic Investment	.043
2012	Keysar, Hayakawa & An	Foreign-Language Effect	.091
2012	Leung, Kim, Polman, Ong, Qiu, Goncalo & Sanchez-Burks	Embodied Metaphors and Creative "Acts"	.076
2012	Rounding, Lee, Jacobson & Ji	Religion and Self-Control	.036
2012	Savani & Rattan	Choice and Inequality	.064
2012	van Boxtel & Koch	Visual Rivalry Without Spatial Conflict	.071
2011	Evans, Horowitz & Wolfe	Weighting of Evidence in Rapid Scene Perception	.426
2011	Inesi, Botti, Dubois, Rucker & Galinsky	Power and Choice	.026
2011	Nordgren, Morris McDonnell, & Lowenstein	What Constitutes Torture?	.090
2011	Savani, Stephens & Markus	Interpersonal and Societal Consequences of Choice	.063
2011	Todd, Hanko, Galinsky & Mussweiler	Difference Mind-Set and Perspective Taking	.043
2011	Tuk, Trampe & Warlop	Inhibitory Spillover	.092
2010	Balcetis & Dunning	Wishful Seeing	.076
2010	Bowles & Gelfand	Status and Workplace Deviance	.057
2010	Damisch, Stoberock & Mussweiler	How Superstition Improves Performance	.057
2010	de Hevia & Spelke	Number-Spacing Mapping in Human Infants	.070
2010	Ersner-Hershfield, Galinsky, Kray & King	Counterfactual Reflection	.073
2010	Gao, McCarthy & Scholl	The Wolfpack Effect	.115

2010	Lammers, Stapel & Galinsky	Power and Hypocrisy	.024
2010	Li, Wei & Soman	Physical Enclosure and Psychological Closure	.079
2010	Maddux, Yang, Falk, Adam, Adair, Endo, Carmon & Heine	Culture and the Endowment Effect	.014
2010	McGraw & Warren	Benign Violations	.081
2010	Sackett, Meyvis, Nelson, Converse & Sackett	When Time Flies	.033
2010	Savani, Markus, Naidu, Kumar & Berlia	What Counts as a Choice?	.058
2010	Senay, Albarracín & Noguchi	Interrogative Self-Talk and Intention	.090
2010	West, Anderson, Bedwell & Pratt	Red Diffuse Light Suppresses Fear Prioritization	.157
2009	Alter & Oppenheimer	Fluency and Self-Disclosure	.071
2009	Ashton-James, Maddux, Galinsky & Chartrand	Affect and Culture	.035
2009	Fast & Chen	Power, Incompetence, and Aggression	.072
2009	Fast, Gruenfeld, Sivanathan & Galinsky	Power and Illusory Control	.069
2009	Garcia & Tor	The *N*-Effect	.089
2009	González & McLennan	Hemispheric Differences in Sound Recognition	.139
2009	Hahn, Close & Graf	Transformation Direction	.348
2009	Hart & Albarracín	Describing Actions	.035
2009	Janssen & Caramazza	Phonology and Grammatical Encoding	.083
2009	Jostmann, Lakens & Schubert	Weight and Importance	.090
2009	Labroo, Lambotte & Zhang	The Name-Ease Effect and Importance Judgments	.008
2009	Nordgren, van Harreveld & van der Pligt	Restraint Bias	.0998
2009	Wakslak & Trope	Construal Level and Subjective Probability	.061
2009	Zhou, Vohs & Baumeister	Symbolic Power of Money	.041

The TES analysis cannot identify the source of bias for an article that appears to have too much success, but publication bias and questionable research practices (John, Lowenstein & Prelec, 2012; Simmons, Nelson & Simonsohn, 2011) are plausible explanations. For an apparently biased experiment set, it is possible that the reported experiments are part of a larger set containing unsuccessful studies that are relevant to the

theoretical ideas but are not reported. It is also possible that the experiments utilized inappropriate data collection or analysis methods to produce *p* values below the .05 criterion. Such manipulations tend to overestimate two variables that scientists care about: effect size and replication rate (Lane & Dunlap, 1978; Francis, 2012d,e, 2013b,c). As a result of these biases, readers cannot estimate the true effect sizes with the provided data (they might be zero) or predict the outcome of future studies. It is important to note that the presence of bias should not be taken as an indication that the theoretical claims are necessarily wrong. The proper interpretation is that the presented justification for the theoretical claims is invalid. In some cases, it may be possible to interpret some subset of the reported findings as being valid (Francis, 2013c), but such a re-analysis requires subject matter expertise and possibly a new theoretical interpretation of the findings.

A skeptical attitude toward the validity of a biased experiment set mirrors how researchers typically think about bias within the context of a single experiment. Scientists would not accept a conclusion based on data from an experiment where a researcher excluded participants who did not show a desired result or insisted that each participant perform a task until showing a desired result. Similar kinds of biases exist at the level of experiment sets, and they can create experiment sets where the reported rate of experiment success is incongruent with computations of experimental success probability. The TES analysis detects this discrepancy.

To what extent the rate of apparent bias in Table 2 generalizes beyond the immediate sample depends on whether one interprets the sample as being representative of a given population. The majority of articles in *Psychological Science* include fewer than four experiments, and such articles may have different rates of bias (lower or higher).

Moreover, the findings may not generalize beyond *Psychological Science.* The journal deliberately publishes articles of broad interest with innovative and ground-breaking findings, and it could be that such an emphasis tends to attract articles that appear to be biased. Other journals with different publishing goals may be less prone to include articles that appear to be biased. Regardless of whether the findings in Table 2 generalize, the high rate of apparent bias is a serious concern for *Psychological Science*, and it raises concerns about the quality of findings in other journals.

Unless there is a flaw in the TES analysis (for debates see Francis, 2013b, c; Johnson, 2013; Johnson & Yuan, 2007; Morey, 2013; Simonsohn, 2012, 2013), there are two broad explanations for how there could be such a high rate of apparent bias among the articles in *Psychological Science*: malfeasance or ignorance. The former interpretation supposes that researchers deliberately introduce bias into their findings, which is essentially fraud. Although there may be a few unscrupulous researchers who generate flawed investigations with a full understanding that what they are doing misleads the field, the TES analysis does not necessarily lead to such a conclusion. An alternative explanation is that these authors were unaware that some of the methods used to gather data, analyse data, or theorize would introduce a bias. Such a charge may seem nearly as disparaging as an accusation of fraud, but there are no pleasant choices here (Gelman, 2013). Ignorance is a plausible explanation for the findings in Table 2 because it is easy to introduce bias even when a researcher attempts to run proper studies (Francis, 2012b, 2013b,c; Gelman & Loken, 2013; John *et al.*, 2012; Simmons *et al.*, 2011). The TES analysis may cast doubt on the validity of a report, but it should not, by itself, be used to denigrate the ethics of any author whose work is listed in Table 2.

There is a fourth possible explanation for the appearance of bias for any specific article: chance. It sometimes happens that an experiment set that is properly sampled, analysed, and reported produces results that are uncommonly successful relative to the estimated effects. This possibility is an additional reason why a conclusion of apparent bias from a TES analysis should not, by itself, be used to denigrate an author's ethics. However, such caution does not mean that scientists should ignore the TES conclusion; skepticism about the validity of the reported experiments is appropriate even though there is a chance of a false alarm. To behave otherwise dismisses the principles of hypothesis testing, which formed the foundation for all of the articles in Table 2.

Conclusions

Table 2 provides the first estimate of the frequency of biased articles for an important psychology journal. The finding that 82% of the analysed articles in *Psychological Science* appear to be biased indicates that the problems within the field are severe. Many readers, editors, and reviewers accepted the articles listed in Table 2 as containing innovative and ground-breaking findings in the field, but the TES analyses suggest that scientists should actually be skeptical about the validity of many of those articles. The frequency of such flawed articles implies fundamental misunderstandings about how to generate and identify good scientific arguments from multiple experiments, which suggests a need for radical changes in statistical analysis, experimental design, and theory development.

The findings reported here validate the general recognition that something is wrong with common scientific practice in psychology. This concern has prompted many proposals for reform, including increased emphasis on replication (Koole & Lakens, 2012; Roediger,

2012), enhanced disclosure of experimental methods (Eich, 2014; LeBel, Borsboom, Giner-Sorolla, Hasselman, Peters, Ratliff & Smith, 2013), improved access to experimental data (Nosek, Spies & Motyl, 2012), focusing on confidence intervals and meta-analysis rather than hypothesis testing (Cumming, 2014), adoption of Bayesian data analysis methods (Kruschke, 2010; Rouder, Speckman, Sun, Morey & Iverson, 2009; Wagenmakers, 2007), pre-registration of hypotheses and methods (Chambers, 2013; Wagenmakers, Wetzels, Borsboom, van der Maas & Kievit, 2012; Wolfe, 2013), and badges for articles that follow some of these reforms (Eich, 2014). These changes to data analysis and publishing will hopefully reduce the frequency of articles with excess success, and it would be valuable for future TES analyses to compare the frequencies of articles with excess success before and after a journal institutes such reforms.

While these reforms aim for a better future, it is also important to understand and interpret the problems of the past. It is not enough to simply recognize that a high proportion of past studies appear biased. To be able to plan future studies and develop theories, scientists need to know which past studies contain flawed findings or theories. If a study appears to have an excess of success, then scientists know to not trust the results and they can run new experiments to check on the theoretical ideas. By noting past problems and motivating verification checks, TES analyses complement reform efforts and improve both scientific practice and the accumulation of scientific knowledge.

References

Alter, A. L., & Oppenheimer, D. M. (2009). Suppressing secrecy through metacognitive ease: Cognitive fluency encourages self-disclosure. *Psychological Science*, *20*(11), 1414-1420.

Anderson, C., Kraus, M. W., Galinsky, A. D., & Keltner, D. (2012). The local-ladder effect: Social status and subjective well-being. *Psychological Science*, *23*(7), 764-771.

Ashton-James, C. E., Maddux, W. W., Galinsky, A. D., & Chartrand, T. L. (2009). Who I am depends on how I feel: The role of affect in the expression of culture. *Psychological Science*, *20*(3), 340-346.

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

Balcetis, E., & Dunning, D. (2010). Wishful seeing: More desired objects are seen as closer. *Psychological Science*, *21*(1), 147-152.

Bauer, M. A., Wilkie, J. E. B., Kim, J. K., & Bodenhausen, G. V. (2012). Cuing consumerism: Situational materialism undermines personal and social well-being. *Psychological Science*, *23*(5), 517-523.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics, 50*, 1088-1101.

Birtel, M. D., & Crisp, R. J. (2012). "Treating" prejudice: An exposure-therapy approach to reducing negative reactions toward stigmatized groups. *Psychological Science*, *23*(11), 1379-1386.

Bowles, H. R., & Gelfand, M. (2010). Status and the evaluation of workplace deviance. *Psychological Science*, *21*(1), 49-54.

Chambers, C. D. (2013). Registered reports: A new publishing initiative at *Cortex*. *Cortex*, *49*, 609-610.

Converse, B. A., & Fishbach, A. (2012). Instrumentality boosts appreciation: Helpers are more appreciated while they are useful. *Psychological Science*, *23*(6), 560-566.

Converse, B. A., Risen, J. L., & Carter, T. J. (2012). Investing in karma: When wanting promotes helping. *Psychological Science*, *23*(8), 923-930.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7-29.

Damisch, L., Stoberock, B. & Mussweiler, T. (2010). Keep your fingers crossed! How superstition improves performance. *Psychological Science*, *21*(7), 1014-1020.

de Hevia, M. D., & Spelke, E. S. (2010). Number-space mapping in human infants. *Psychological Science*, *21*(5), 653-660.

Egger, M., Davey Smith, G, Scheneider, M. & Minder, C. E. (1997). Bias in a meta-analysis detected by a simple, graphical test. *British Medial Journal*, *315*, 639-634.

Eich, E. (2014). Business not as usual. *Psychological Science*, *25*(1), 3-6.

Ersner-Hershfield, H., Galinsky, A. D., Kray, L. J., & King, B. G. (2010). Company, country, connections: Counterfactual origins increase organizational commitment, patriotism, and social investment. *Psychological Science*, *21*(10), 1479-1486.

Evans, K. K., Horowitz, T. S., & Wolfe, J. M. (2011). When categories collide: Accumulation of information about multiple categories in rapid scene perception. *Psychological Science*, *22*(6), 739-746.

Fanelli, D. (2010) "Positive" results increase down the hierarchy of the sciences. *PLoS ONE*, *5*(4): e10068. doi:10.1371/journal.pone.0010068

Fast, N. J., & Chen, S. (2009). When the boss feels inadequate: Power, incompetence, and aggression. *Psychological Science*, *20*(11), 1406-1413.

Fast, N. J., Gruenfeld, D. H., Sivanathan, N., & Galinsky, A. D. (2009). Illusory control: A generative force behind power's far-reaching effects. *Psychological Science*, *20*(4), 502-508.

Francis, G. (2012a). Too good to be true: Publication bias in two prominent studies from experimental psychology, *Psychonomic Bulletin & Review*, *19*, 151-156.

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

Francis, G. (2012c). Evidence that publication bias contaminated studies relating social class and unethical behavior. *Proceedings of the National Academy of Sciences*. *109*:E1587.

Francis, G. (2012d). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review*, *19*(6), 975-991.

Francis, G. (2012e). The psychology of replication and replication in psychology. *Perspectives on Psychological Science*, *7*(6), 580-589.

Francis, G. (2013a). Publication bias in "Red, Rank, and Romance in Women Viewing Men" by Elliot et al. (2010). *Journal of Experimental Psychology: General*, *142*, 292-296.

Francis, G. (2013b). Replication, statistical consistency, and publication bias. *Journal of Mathematical Psychology*, *57*, 153-169.

Francis, G. (2013c). We should focus on the biases that matter: A reply to commentaries. *Journal of Mathematical Psychology*, *57*, 190-195.

Gao, T., McCarthy, G., & Scholl, B. J. (2010). The wolfpack effect: Perception of animacy irresistibly influences interactive behavior. *Psychological Science*, *21*(12), 1845-1853.

Garcia, S. M., & Tor, A. (2009). The *N*-effect: More competitors, less competition. *Psychological Science*, *20*(7), 871-877.

Gelman, A. (2013). Is it possible to be an ethicist without being mean to people? *Chance*, *26*(4), 51-53.

Gelman, A. & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Downloaded January 30, 2014 from http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

González, J., & McLennan, C. T. (2009). Hemispheric differences in the recognition of environmental sounds. *Psychological Science*, *20*(7), 887-894.

Hahn, U., Close, J., & Graf, M. (2009). Transformation direction influences shape-similarity judgments. *Psychological Science*, *20*(4), 447-454.

Hart, W., & Albarracín, D. (2009). What I was doing versus what I did: Verb aspect influences memory and future actions. *Psychological Science*, *20*(2), 238-244.

Inesi, M. E., Botti, S., Dubois, D., Rucker, D. D., & Galinsky, A. D. (2011). Power and choice: Their dynamic interplay in quenching the thirst for personal control. *Psychological Science*, *22*(8), 1042-1048.

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

Janssen, N., & Caramazza, A. (2009). Grammatical and phonological influences on word order. *Psychological Science*, *20*(10), 1262-1268.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science*, *23*, 524-532.

Johnson, V. (2013). On biases in assessing replicability, statistical consistency and publication bias. *Journal of Mathematical Psychology, 57*(5)*,* 177-179.

Johnson, V. & Yuan, Y. (2007). Comments on "An exploratory test for an excess of significant findings" by JPA loannidis and TA Trikalinos. *Clinical Trials, 4,* 254-255.

Jostmann, N. B., Lakens, D., & Schubert, T. W. (2009). Weight as an embodiment of importance. *Psychological Science*, *20*(9), 1169-1174.

Keysar, B., Hayakawa, S. L., & An, S. G. (2012). The foreign-language effect: Thinking in a foreign tongue reduces decision biases. *Psychological Science*, *23*(8), 661-668.

Koole, S. L. & Lakens, D. (2012). Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science*, *7*(6), 608-614.

Kruschke, J. K. (2010). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(5), 658-676. doi:10.1002/wcs.72

Labroo, A. A., Lambotte, S., & Zhang, Y. (2009). The "name-ease" effect and its dual impact on importance judgments. *Psychological Science*, *20*(12), 1516-1522.

Lammers, J., Stapel, D. A., & Galinsky, A. D. (2010). Power increases hypocrisy: Moralizing in reasoning, immorality in behavior. *Psychological Science*, *21*(5), 737-744.

Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. Br*itish Journal of Mathematical and Statistical, Psychology*, *31*, 107-112.

LeBel, E. P., Borsboom, D., Giner-Sorolla, R., Hasselman, F., Peters, K. R., Ratliff, K. A. & Smith, C. T. (2013). PsychDisclosure.org: Grassroots Support for Reforming Reporting Standards in Psychology. *Perspectives on Psychological Science*, *8*(4), 424-432.

Leung, A. K., Kim, S., Polman, E., Ong, L. S., Qiu, L., Goncalo, J. A., & Sanchez-Burks, J. (2012). Embodied metaphors and creative "acts". *Psychological Science*, *23*(5), 502-509.

Li, X., Wei, L., & Soman, D. (2010). Sealing the emotions genie: The effects of physical enclosure on psychological closure. *Psychological Science*, *21*(8), 1047-1050.

Maddux, W. W., Yang, H., Falk, C., Adam, H, Adair, W., Endo, Y., Carmon, Z. & Heine, S. J. (2010). For whom is parting with possessions more painful?: Cultural differences in the endowment effect. *Psychological Science*, *21*(12), 1910-1917.

McGraw, A. P., & Warren, C. (2010). Benign violations: Making immoral behavior funny. *Psychological Science*, *21*(8), 1141-1149.

Morey, R. D. (2013). The consistency test does not–and cannot–deliver what is advertised: A comment on Francis (2013). *Journal of Mathematical Psychology, 57*(5)*,* 180-183.

Nordgren, L. F., Morris McDonnell, M-H., & Lowenstein, G. (2011). What constitutes torture?: Psychological impediments to an objective evaluation of enhanced interrogation tactics. *Psychological Science*, *22*(5), 689-694.

Nordgren, L. F., van Harreveld, F., & van der Pligt, J. (2009). The restraint bias: How the illusion of self-restraint promotes impulsive behavior. *Psychological Science*, *20*(12), 1523-1528.

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*(6), 615-631.

Renkewitz, F., Fuchs H. M., & Fiedler, S. (2011). Is there evidence of publication biases in JDM research? *Judgment and Decision Making*, *6*, 870-881.

Roediger, H. L., III. (2012). Psychology's woes and a partial cure: The value of replication. *APS Observer*, *25*(2) http://www.psychologicalscience.org/index.php/publications/observer/2012/february-11-2012-observer-publications/psychology's-woes-and-a-partial-cure-the-value-of-replication.html

Rouder, J.N., Speckman, P.L., Sun, D., Morey, R.D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, **16**, 225–237.

Rounding, K., Lee, A., Jacobson, J. A., & Ji, L-J. (2012). Religion replenishes self-control. *Psychological Science*, 23(6), 635-642.

Sackett, A. M., Meyvis, T., Nelson, L. D., Converse, B. A. & Sackett, A. L. (2010). You're having fun when time flies: The hedonic consequences of subjective time progression. *Psychological Science*, *21*(1), 111-117.

Savani, K., Markus, H. R., Naidu, N. V. R., Kumar, S., & Berlia, N. (2010). What counts as a choice?: U.S. Americans are more likely than Indians to construe actions as choices. *Psychological Science*, *21*(3), 391-398.

Savani, K., & Rattan, A. (2012). A choice mind-set increases the acceptance and maintenance of wealth inequality. *Psychological Science*, *23*(7), 796-804.

Savani, K., Stephens, N. M., & Markus, H. R. (2011). The unanticipated interpersonal and societal consequences of choice: Victim blaming and reduced support for the public good. *Psychological Science*, *22*(6), 795-802.

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple study articles. *Psychological Methods*, *17*(4), 551-566.

Senay, I., Albarracín, D., & Noguchi, K. (2010). Motivating goal-directed behavior through introspective self-talk: The role of the interrogative form of simple future tense. *Psychological Science*, *21*(4), 499-504.

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.

Simonsohn, U. (2012). It does not follow evaluating the one-off publication bias critiques by Francis (2012a, 2012b, 2012c, 2012d, 2012e, in Press). *Perspectives on Psychological Science, 7*(6)*,* 597-599.

Simonsohn, U. (2013). It really just does not follow, comments on Francis (2013). *Journal of Mathematical Psychology, 57*(5)*,* 174-176.

Simonsohn, U. Nelson, L. D., & Simmons, J. P. (in press). P-curve: A key to the file drawer. *Journal of Experimental Psychology: General*.

Sterling, T. D. (1959). Publication decisions and the possible effects on inferences drawn from test of significance—or vice versa. *Journal of the American Statistical Association, 54*, 30-34.

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician*, *49*, 108-112.

Todd, A. R., Hanko, K., Galinsky, A. D., Mussweiler, T. (2011). When focusing on differences leads to similar perspectives. *Psychological Science*, *22*(1), 134-141.

Tuk, M. A., Trampe, D. & Warlop, L. (2011). Inhibitory spillover: Increased urination urgency facilitates impulse control in unrelated domains. *Psychological Science*, *22*(5), 627-633.

van Boxtel, J. J. A., & Koch, C. (2012). Visual rivalry without spatial conflict. *Psychological Science*, *23*(4), 410-418.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p-values. *Psychonomic Bulletin & Review*, *14*, 779–804.

Wagenmakers, E.-J. Wetzels, R., Borsboom, D., van der Maas, H. L. J. & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632–638.

Wakslak, C., & Trope, Y. (2009). The effect of construal level on subjective probability estimates. *Psychological Science*, *20*(1), 52-58.

West, G. L., Anderson, A. K., Bedwell, J. S., & Pratt, J. (2010). Red diffuse light suppresses the accelerated perception of fear. *Psychological Science*, *21*(7), 992-999.

Wolfe, J. (2013). Registered Reports and Replications in Attention, Perception, & Psychophysics. *Attention, Perception, & Psychophysics*, *75*(5), 781-783.

Zhou, X., Vohs, K. D., & Baumeister, R. F. (2009). The symbolic power of money: Reminders of money alter social distress and physical pain. *Psychological Science*, *20*(6), 700-706.