UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Francis, G. & Thunell, E. (2019). Excess Success in "Ray of hope: Hopelessness Increases Preferences for Brighter Lighting". *Collabra: Psychology,* 5(1): 22. DOI: https://doi.org/10.1525/collabra.213

ORIGINAL RESEARCH REPORT

Excess Success in "Ray of hope: Hopelessness Increases Preferences for Brighter Lighting"

Gregory Francis[*] and Evelina Thunell[*,†,‡,§]

Dong, Huang, and Zhong (2015) report five successful experiments linking brightness perception with the feeling of hopelessness. They argue that a gloomy future is psychologically represented as darkness, not just metaphorically but as an actual perceptual bias. Based on multiple results, they conclude that people who feel hopeless perceive their environment as darker and therefore prefer brighter lighting than controls. Reversely, dim lighting caused participants to feel more hopeless. However, the experiments succeed at a rate much higher than predicted by the magnitude of the reported effects. Based on the reported statistics, the estimated probability of all five experiments being fully successful, if replicated with the same sample sizes, is less than 0.016. This low rate suggests that the original findings are (perhaps unintentionally) the result of questionable research practices or publication bias. Readers should therefore be skeptical about the original results and conclusions. Finally, we discuss how to design future studies to investigate the relationship between hopelessness and brightness.

Keywords: Excess success; publication bias; brightness perception; perceptual bias; statistics

Intuitively, we might feel more convinced by a scientific finding that is directly or conceptually replicated across several experiments. In some cases this attitude is warranted, but basic probability theory indicates that there should be some failed tests if the experiments have moderate to low power. Consider for example an article reporting five independent data sets, all in support of a certain theory. Suppose the power of each study is 0.6, meaning that there is a 60% chance of successfully replicating any one of the studies with a new independent sample of the same size. We can compute the power for the set of five experiments, and because the studies are independent this joint power is the product of the individual power values, i.e. $0.6^5 = 0.08$. In other words, there is only an 8% chance of all five experiments being successful when they are repeated. The example illustrates how a set of studies that are perhaps only slightly under-powered when considered separately, can give a joint power that is too low for most scientific purposes. "Excess success" refers to the absence of failures in such under-powered conditions, and it is a sign of questionable research practices (John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011), the garden of forking paths (Gelman & Loken, 2014), or publication bias (Sterling, 1959; Sterling, Rosenbaum, & Weinkam, 1995). These practices lead to overestimated effect sizes and potentially misleading descriptions of empirical results.

The Test for Excess Success (TES) is one way of identifying excess success in multi-study publications (Francis, 2013b; Ioannidis & Trikalinos, 2007; Schimmack, 2012). Although reports from such analyses can themselves be subject to publication bias (Simonsohn, 2012, 2013), selective reporting of excess success is not a problem as long as the conclusions from the TES are restricted to the analyzed set of data. In other words, we can draw conclusions about the presence of bias in a set of studies even if other, unrelated, investigations of other studies do not indicate a presence of bias, or are not analyzed or reported. An investigation of excess success cannot, however, make inferences about other sets of studies. For example, a finding of excess success in one article does not indicate problems in other studies by the same author(s), in articles published in the same journal, or in investigations in the same field of study. Such inferences would require random selection of study sets, which is not possible when analyzing the studies in a given article. These and other concerns about the TES (Morey, 2013; Vandekerckhove, Guan, & Styrcula, 2013) have been address in Francis (2013b, 2013c).

Importantly, the appearance of excess success means that the original set of studies does not provide appropriate support for the original conclusions. The validity of these conclusions thus remains an open question. Given that excess success often seems to be driven by publication

[*] Department of Psychological Sciences, Purdue University, West Lafayette, US

[†] Centre de Recherche Cerveau et Cognition (CerCo), Université Paul Sabatier, FR

[‡] Centre National de la Recherche Scientifique (CNRS), Université Paul Sabatier, FR

[§] Department of Clinical Neuroscience, Karolinska Institute, Stockholm, SE

Corresponding author: Gregory Francis (gfrancis@purdue.edu)

bias or questionable research practices, it seems likely that even if the conclusions are true the effects are smaller than suggested by the reported empirical studies. For example, Francis (2013a) noted that a set of experiments claiming that women find men more attractive when the men are wearing red (Elliot et al. 2010) was too successful relative to the experiments' estimated power values. In response, Elliot and Maier (2013) reported a new study with a much larger sample that did not produce a significant result and had a standardized effect size that was half of that found in the original publication. In a follow-up meta-analysis that included many additional studies, Lehmann, Elliot, and Calin-Jageman (2018) in turn found a standardized effect size half as large as that reported by Elliot and Maier (2013). The first and third authors of the meta-analysis concluded that the empirical evidence suggests that the effect is very small or even non-existent, while the second author concluded that previous investigations were dramatically underpowered but held out hope that better studies might still show appropriate evidence for the effect. Both of these conclusions are consistent with the excess success analysis in Francis (2013a).

Here, we apply the Test for Excess Success to a set of studies reported by Dong, Huang and Zhong (2015). Inspired by conceptual metaphor theory, the authors hypothesized a link between brightness perception and the feeling of hopelessness. Across five successful studies, they confirm their hypotheses. After writing about a hopeless experience, participants judged their environment as darker and therefore preferred brighter ambient lighting as compared to controls. Likewise, participants who were already more hopeless to begin with showed a preference for brighter lights. Reversely, dim lighting induced a feeling of hopelessness. If this connection is valid, it could have important implications. For example, Dong et al. suggest that their conclusions could guide policies on lighting levels in public spaces. However, our TES analysis indicates that even if the effects are real, the reported studies are unsuitable to demonstrate them. We argue that policy guidelines should *not* be based on the results of Dong et al. (2015), because their findings appear to be biased.

1. Applying the Test for Excess Success

In the first study of Dong et al. (2015), the authors tested whether recalling a "hopelessness experience" could affect the perceptual judgment of light and hypothesized that "the sense of having a gloomy future, but not just feeling sad, may be psychologically represented as darkness". The experimental results showed that participants who wrote about a situation in which they had felt hopeless subsequently rated the room as more dim compared to participants who wrote about a sad, hopeful, or neutral experience. The authors concluded that hopelessness is related to perceived brightness but other emotions are not. These claims were based on seven successful statistical tests within this one experiment (**Table 1**). First, a main effect of an ANOVA indicated a difference in brightness judgments between the writing conditions. Further, three significant contrasts comparing perceived brightness for the hopeless condition against the other conditions confirmed that the hopeless condition was the odd one out. Last, three non-significant contrasts comparing the sad, hopeful and neutral conditions against each other were taken as evidence that there was no difference in brightness judgments between feelings other than hopelessness. According to the authors, all these outcomes were expected based on their hypotheses, and we therefore include them all in our TES analysis.

What is the likelihood of replicating this set of findings if the population means and standard deviations match the reported sample statistics and the experiment is repeated with the same sample sizes? Since success is defined not by one simple test, but by the specific pattern of outcomes of the seven tests, there is no simple formula for the post-hoc power of this experiment. The "success rate" of the set of tests has to be lower than the power of any individual test but higher than their product (since some tests are dependent). We estimated the success rate by analysing simulated datasets based on the reported statistics (means and standard deviations) and computing the proportion of samples that produced successful outcomes for all seven tests.[1] These Monte Carlo simulations resulted in a success rate

Table 1: Statistical properties of the tests included in the TES analysis of Study 1 in Dong et al., (2015). The hypotheses are listed to the left (μ denotes population mean). To the right, the relevant reported statistics are listed. *n* denotes sample sizes for the different induced feelings, $\bar{x}$ denotes the corresponding mean perceived brightness of the lab, and *s* indicates the standard deviation. We report means and standard deviations because some tests are dependent so joint power cannot be calculated from standardized effect sizes.

	Statistics			
Supporting hypotheses	**Hopeless ($n = 46$)**	**Hopeful ($n = 46$)**	**Sad ($n = 46$)**	**Neutral ($n = 45$)**
Main effect of emotion				
$\mu_{hopeless} \neq \mu_{hopeful}$				
$\mu_{hopeless} \neq \mu_{sad}$	$\bar{x} = 5.38$	$\bar{x} = 6.74$	$\bar{x} = 6.22$	$\bar{x} = 6.43$
$\mu_{hopeless} \neq \mu_{neutral}$	$s = 2.24$	$s = 1.83$	$s = 1.66$	$s = 1.94$
$\mu_{hopeful} = \mu_{sad}$				
$\mu_{sad} = \mu_{neutral}$				
$\mu_{hopeful} = \mu_{neutral}$				

of 0.363. Success probabilities for the individual tests are reported at the Open Science Framework. Dong et al. reported several other successful tests to further support their conclusions, including comparisons of rated comfortableness and temperature of the room that, as predicted, were unaffected by the emotion condition. The manuscript does not contain enough details for us to incorporate these additional supporting findings into our analysis.[2] Doing so could only lower the estimated success rate, and our estimate of 0.363 should therefore be considered an upper limit.

In their supplemental material, Dong et al. (2015) reported that Study 1 had a post-hoc power of 0.82, which is notably higher than the 0.363 that we compute. It seems that Dong et al. based their calculation only on the main effect of the ANOVA, while we consider a larger set of relevant tests. The larger set is appropriate because, by itself, a significant main effect would not support the conclusions drawn by Dong et al. In particular, the authors considered it crucial that brightness ratings for participants in the hopeless condition were significantly different from brightness ratings for participants in each of the other conditions. Further, to argue for a specific effect of hopelessness on brightness perception, Dong et al. indicated that it is important that brightness judgments did *not* differ across the other emotion conditions. Ideally, all of these tests should be included when estimating the probability that a study will be successful.

In Study 2, Dong et al. (2015) hypothesized that preference for lighting should be related to hopelessness about the economy and career prospects. Here, there was no manipulation of the participants' feelings, but rather a measure of hopelessness was computed from a participant's rating of their financial situation and future job prospects. In addition, a "preference for lighting index" was calculated based on preference for pictures of desk lamps of varying brightness, the wattage that participants would prefer for their living room light bulb, and what lighting ambiance they considered ideal for their living room (on a scale from dim to bright). In agreement with their hypothesis, the authors concluded that hopeless participants preferred brighter lighting. A key finding was a mediation analysis revealing that the hopelessness index mediated an effect of a socio-economic status (SES) score on lighting preference. To estimate the success rate for a replication study, we consider the confidence interval (CI) of this mediation analysis; [0.0038, 0.2548]. We cannot quantitatively estimate the success rate because we do not have access to the raw data; but the lower limit of the reported CI is very close to zero, so a replication experiment would be expected to have a success rate close to 0.5 (a lower CI limit of zero would indicate that the CI for a replication study of the same sample size is equally likely to include or exclude zero). To ensure that we do not underestimate the success rate, we use a value of 0.6. The conclusions for Study 2 were also based on other successful tests (e.g., a non-significant reversed mediation effect, and non-significant effects of hopelessness on preferences for sofa softness and painting size), but Dong et al. did not report enough details to allow us to include them in our TES analysis. Like we argued for Study 1, including these additional successful tests can only lower the estimated success rate and our value should therefore be considered an overestimate.

The purpose of Study 3 in Dong et al. (2015) was to show that the relationship between feelings of hopelessness and lighting preference found in Study 2 were causal. Here, in a between subjects design, participants were asked to recall and describe a situation that had evoked a hopeless, hopeful, or neutral feeling. Thus, the feeling was experimentally induced in a randomized way rather than simply assessed as in Study 2. As the authors expected, participants who recalled a hopeless situation expressed a greater desire for brighter lighting than those who recalled hopeful or neutral events (a main effect from an ANOVA and two contrasts indicated differences between the hopeless and the two other emotion conditions). In addition, four control measures (preferred sofa softness, room temperature, and painting and fish tank sizes) were not affected by emotion condition (non-significant ANOVAs).[3] As for Study 1, we simulated data sets based on the reported means and standard deviations and computed the proportion of successful outcomes (see **Table 2** for details about the included tests). The various measures are correlated and (following common practices) Dong et al. did not report the correlation. We therefore ran our simulation with many different correlations between the measures and picked the correlation ($r = -0.24$) that resulted in the highest success rate (this method was not possible for Study 1 where the authors reported fewer details about the control measures). Since the true correlation may be different than the value that produces the highest rate of success, the estimated success rate of 0.234 for Study 3 should be considered an overestimate. Success probabilities for the individual tests are reported at the Open Science Framework.

In Study 4 the authors proposed that the lighting level in the room should influence participants' hopefulness. In agreement with their hypothesis, they concluded that dim lighting can induce hopelessness: participants in a dim room reported feeling more hopeless about their future job search prospects than participants in a well-lit room. Using the *pwr* library in R (Champely et al., 2018), we computed the success rate (power) of this study based on the sample sizes and the standardized effect size of the difference between the dim and well-lit conditions ($n_{dim} = n_{well\text{-}lit} = 53$, $t(104) = 2.15$, $g = 0.41$). There were additional successful tests that Dong et al. used to support their conclusions, but without access to the raw data we cannot include them in our analysis. Thus, our computed power of 0.562 should be considered an overestimate of the success rate of the full set of experimental tests.

In their supplemental material, Dong et al. reported that Study 4 had a post-hoc power of 0.99, which is much larger than our estimated value of 0.562. It seems that Dong et al. considered only their strongest finding (a difference between brightness conditions that controlled for several demographic factors) in their power analysis, which is inappropriate because the conclusions are based not only on that finding but also on several weaker results. When a conclusion depends on multiple tests, replication success (power) depends primarily not on the strongest but on the *weakest* finding.

Table 2: Statistical properties of the tests included in the TES analysis of Study 3 in Dong et al., (2015). The different types of ratings are listed in the left column, and corresponding supporting hypotheses and z-scored results are shown in the middle and rightmost columns, respectively. μ denotes population means and *n* sample sizes for the different induced feelings. $\bar{x}$ denotes the corresponding mean rating, and *s* indicates the standard deviation. The simulated data also assumed $r = –0.24$ across ratings; this value maximized power across the tests.

		Statistics		
Rating	**Supporting hypotheses**	**Hopeless ($n = 68$)**	**Hopeful ($n = 68$)**	**Neutral ($n = 68$)**
Lighting	Main effect			
	$\mu_{hopeless} \neq \mu_{hopeful}$	$\bar{x} = 0.23$	$\bar{x} = –0.19$	$\bar{x} = –0.05$
	$\mu_{hopeless} \neq \mu_{neutral}$	$s = 0.88$	$s = 0.72$	$s = 0.75$
Sofa	No main effect	$\bar{x} = –0.02$	$\bar{x} = –0.01$	$\bar{x} = 0.04$
		$s = 1.08$	$s = 0.99$	$s = 0.94$
Painting	No main effect	$\bar{x} = 0.05$	$\bar{x} = –0.04$	$\bar{x} = –0.01$
		$s = 1.05$	$s = 0.98$	$s = 0.96$
Fish tank	No main effect	$\bar{x} = 0.15$	$\bar{x} = 0.05$	$\bar{x} = –0.17$
		$s = 1.05$	$s = 1.06$	$s = 0.87$
Temperature	No main effect	$\bar{x} = 0.17$	$\bar{x} = –0.09$	$\bar{x} = –0.05$
		$s = 0.95$	$s = 1.14$	$s = 0.91$

An additional study (here referred to as Study 5) is described in the supplemental material of Dong et al. Since this study is described as corroborating Studies 2 and 3 and is described as being in the supplementary material only due to space constraints, we include it in our analysis. Here, the authors report significant correlations between hopelessness (measured as lack of preparedness for an upcoming exam) and several related variables: ratings of the brightness of lamp pictures, liking of bright lamps, and intention to purchase bright lamps. In contrast, the correlations of hopelessness with height, aesthetics and sturdiness of the lamps were non-significant. Further, chronic optimism, importance of academic goal, general preference for light, and general preference for bright colors did not influence participants' liking or purchase intention of the lamps. Using the *pwr* library in R (Champely et al., 2018), we calculated a success rate of 0.551 for this study, based only on the correlation between preparedness for the exams and rated brightness of the lamps ($n = 59$, $r = –0.27$). There were multiple other successful outcomes, so this value is likely an overestimate of the success rate for the full set of tests. Including the additional tests in our analysis would require additional information (or raw scores). Note that there was a failure in Study 5: A manipulation to induce a sense of being prepared or unprepared for exams did not have the intended effect. However, this manipulation is not critical for any of the conclusions drawn by Dong et al. since they instead considered only the reported preparedness for the exam in their correlation analyses.

We have thus far estimated the success rates for the five experiments separately. These estimates suppose that the experimental effects are real and of the magnitude reported by the samples. Since the five experiments are based on independent samples, the probability that a set of five experiments like these would be successfully replicated is the product of the individual success rates, i.e. 0.363 × 0.6 × 0.234 × 0.562 × 0.551 = 0.0157. Such a small value indicates that some failures should be highly likely and should make readers wonder how the results in Dong et al. could be uniformly successful. Perhaps the researchers were extremely lucky to have all five experiments produce exactly the pattern of significance needed to support the authors' conclusions, but accepting luck as an explanation for the findings is tantamount to rejecting the foundations of null hypothesis testing (Francis, 2012a, 2012b; Ioannidis & Trikalinos, 2007).

A more plausible interpretation of the reported findings is that the studies are biased: either additional experiments were run but not reported (publication bias), the experiments were not run properly and so rejected the null hypothesis more frequently than they should have, or the results were analyzed inappropriately. Any of these interpretations should make readers skeptical about the conclusions that Dong et al. make with regard to their results. We cannot infer that their conclusions are necessarily wrong, but it seems likely that their data (at best) overestimates the magnitude of the effects of interest.

Thus, we advise researchers who are interested in investigating the relation between hopelessness and brightness perception to treat the findings in Dong et al. (2015) as anecdotal. New studies will be needed to determine whether the hypothesized relation actually exists.

2. Designing new studies

A tempting approach for new studies might be to use the methods and analyses described in Dong et al. (2015) with larger sample sizes. Such direct replication attempts are becoming common in psychology (e.g., Alogna et al., 2014; Galak, Leboeuf, Nelson, & Simmons, 2012; Open Science Collaboration, 2015). Although the findings reported by Dong et al. probably overestimate effect sizes, it might be worthwhile to use them as a starting point to identify appropriate methods and sample sizes for replication attempts. As we will see, the nature of Dong et al.'s analyses makes it impossible for some of the experiments to have a high success rate.

We consider two general types of replications. First, we aim for the traditional power (success rate) of 0.8 for an individual study, which might be appropriate for a scientist planning to replicate a single study. However, if five studies are individually powered to 0.8, the chance of all five being successful is only $0.8^5 = 0.33$. We therefore also consider a situation where a researcher requires a success rate of 0.8 across the entire set of experiments. In principle, there are multiple combinations of probabilities across the experiments that could produce the desired success rate, but for simplicity we restrict ourselves to the situation where each study has the same success rate: $0.8^{1/5} = 0.96$. Note that such a high value is often difficult to achieve (sometimes even impossible depending on the design). This difficulty reflects an intrinsic problem of designs that include many different tests and where success is defined as confirming results throughout (Westermann & Hager, 1986). For all sample size analyses, we take into account only the tests that we used for the excess success analysis above (again, additional information would be needed to include all relevant tests). Note that by using the effect sizes reported by Dong et al, we are likely overestimating effects and therefore underestimating the required sample sizes. The R code used for generating all calculations is available at the Open Science Framework.

We start with two easy cases and one case where we cannot estimate sample sizes. Studies 4 and 5 involve only a single hypothesis test, so the required sample sizes can be calculated using the "pwr" library in R (Champely et al., 2018). As **Table 3** indicates, Study 4 would require 93 participants (in each of the two groups) to reach a power of 0.8, and 162 participants to reach a power of 0.96. For Study 5, the corresponding numbers are 105 and 182 participants. In contrast, we do not have enough information to compute an appropriate sample size for the mediation confidence interval in Study 2.

Study 1 involves multiple tests and both significant and non-significant outcomes. For a given sample size, we analyzed 10,000 simulated samples using the means and standard deviations reported by Dong et al. (2015). We set anticipated null effects to have equal mean values, thereby maximizing the probability of producing the desired non-significant outcomes. Such simulations were generated for each possible sample size between 50 and 250 participants per group. For each sample size we computed the proportion of simulated samples that fully satisfied all the hypothesis tests listed in **Table 1**. We then identified the smallest sample size that produced a success proportion of 0.8 or 0.96. The simulations indicate that 86 participants are required in each group for a success probability of 0.8. However, there is no sample size that can produce a success probability of 0.96 for the design and analyses of Study 1 because the analyses depend on multiple non-significant findings. With a significance criterion of 0.05, a hypothesis test can only produce a successful non-significant result with a maximum probability of 0.95. Thus, strictly following the analysis in Dong et al. makes it impossible to achieve a 0.96 probability of success for Study 1. Since success for this study requires three non-significant results, the highest success probability produced by the simulation is 0.88, which is found for large samples where the tests required to show significant effects have power close to 1.0.

We treated Study 3 in a similar way as Study 1, using the reported means and standard deviations from Dong et al., but setting population means for anticipated null results to have equal values, and using the correlation that maximized power for the excess success analysis ($r = -0.24$). We analyzed 10,000 simulated data sets for each possible sample size between 70 and 350. The smallest sample size that produced a success rate of 0.8 for the full set of analyses in **Table 2** was 237 participants in each group. Because Dong et al.'s analysis involves four non-significant outcomes (comparing the effect of emotion on the control measures), it is not possible for any experiment to achieve a success probability of 0.96. The highest success probability in our simulations was 0.82, and nearly all sample sizes above 250 (per group) produced a similar success rate.

The low maximum success probabilities for studies 1 and 3 mean that the analyses used by Dong et al. (2015) are fundamentally unable to have a high success rate across all five experiments. Even if studies 2, 4, and 5 have power values close to 1.0, the full set can have a success rate no higher than the product of the success probabilities for studies 1 and 3: $0.88 \times 0.82 = 0.72$. There is a general lesson here. Although it is common in psychological research to use many different tests to draw specific conclusions about a set of data, such an approach tends to dramatically reduce experimental power. In fact, when the conclusions depend on many successful test outcomes within and across experiments, the researcher has a poor chance to confirm the alternative hypothesis even when it is true. In order to preserve power, a general rule is that the design should be kept as simple as possible

Table 3: Minimum sample sizes for replication studies that use the same methods and analysis plan as Dong et al. (2015).

Study	Power = 0.8	Power = 0.96
1	$n_{hopeless} = n_{hopeful} = n_{sad} = n_{neutral} = 86$	Not possible
2	Cannot compute	Cannot compute
3	$n_{hopeless} = n_{hopeful} = n_{neutral} = 237$	Not possible
4	$n_{dim} = n_{well\text{-}lit} = 93$	$n_{dim} = n_{well\text{-}lit} =162$
5	$n = 105$	$n = 182$

and the conclusions should be based on as few statistical comparisons as possible. We suggest that new studies on the relationship between hopelessness and brightness perception should be designed with these principles in mind. Regarding expected null effects, researchers might for example use Bayesian approaches (Kruschke, 2010; McElreath, 2016; Schonbrodt & Wagenmakers, 2018) or equivalence tests (Lakens, 2017), although such tests tend to require large sample sizes.

3. Conclusions

Our TES analysis indicates that the experiments in Dong et al. (2015) show success at a substantially higher rate than is to be expected if the true effects are of a magnitude similar to what is reported. With multiple experiments and multiple tests within each experiment, random sampling should have produced some test failures. The absence of such failures indicates that something has gone wrong in data collection, data reporting, analysis, or theorizing. Our analysis does not prove that there is no relationship between hopelessness and brightness perception. Rather, we can merely conclude that the findings reported by Dong et al. do not adequately support their conclusions and that if there is a relationship between hopelessness and brightness, it is probably weaker than suggested by the reported data. Further, we demonstrate that by interpreting non-significant outcomes as support for null results, replication studies that use the analysis strategy of Dong et al. only have a moderate chance of providing full support for their claims, even for studies with large samples. We advise researchers who would like to replicate any of the studies in Dong et al. to completely re-design the experiments and to plan appropriate sample sizes for the new analyses.

Data Accessibility Statement

R scripts are available at the Open Science Framework https://osf.io/ayme7/.

Notes

[1] All calculations were performed using the programming language R (R core team, 2017); the code for all simulations is available at the Open Science Framework https://osf.io/ayme7/.

[2] Dong et al. also reported a marginal interaction as support for their claims. However, the reported degrees of freedom, [3, 179], do not correspond to an interaction test. For a mixed ANOVA with $N = 183$ subjects, $C = 3$ within-subjects conditions, and $K = 4$ between-subjects conditions, the degrees of freedom should be $[(C-1) \times (K-1), (N-C) \times (K-1)] = [6, 540]$. Possibly, the reported test is actually for a main effect. We do not include this test in our analysis. We asked the authors about this issue but did not receive a reply.

[3] Similar to Study 1, Dong et al. report a marginally significant interaction with degrees of freedom [2, 201], but these values do not correspond to an interaction test. The correct degrees of freedom should be [8, 804]. Again, we suspect that the reported test is actually a main effect. We do not include this test in our estimation of success probability.

Competing Interests

The authors have no competing interests to declare.

Author Contributions

GF and ET contributed to conception and interpretation. GF developed the analyses and wrote the code. GF and ET wrote the article and approved the submitted version for publication.

References

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, Š., Birch, S., Birt, A. R., Was, C. A., et al. (2014). Registered replication report: Schooler and Engstler-Schooler (1990). *Perspectives on Psychological Science*, *9*(5), 556–578. DOI: https://doi.org/10.1177/1745691614545653

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Ford, C., & **Volcic, R.** (2018). Package 'pwr': Basic functions for power analysis.

Dong, P., Huang, X. I., & **Zhong, C. B.** (2015). Ray of hope: Hopelessness increases preferences for brighter lighting. *Social Psychological and Personality Science*, *6*(1), 84–91. DOI: https://doi.org/10.1177/1948550614542344

Elliot, A. J., & **Maier, M. A.** (2013). The red-attractiveness effect, applying the Ioannidis and Trikalinos (2007b) test, and the broader scientific context: A reply to Francis (2013). *Journal of Experimental Psychology: General*, *142*(1), 297–300. DOI: https://doi.org/10.1037/a0029592

Elliot, A. J., Niesta Kayser, D., Greitemeyer, T., Lichtenfeld, S., Gramzow, R. H., Maier, M. A., & **Liu, H.** (2010). Red, Rank, and Romance in women viewing men. *Journal of Experimental Psychology: General*, *139*(3), 399–417. DOI: https://doi.org/10.1037/a0019689

Francis, G. (2012a). The same old new look: Publication bias in a study of wishful seeing. *I-Perception*, *3*(3), 176–178. DOI: https://doi.org/10.1068/i0519ic

Francis, G. (2012b). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review*, *19*(2), 151–156. DOI: https://doi.org/10.3758/s13423-012-0227-9

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*(1), 292–296. DOI: https://doi.org/10.1037/a0027923

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

Francis, G. (2013c). We should focus on the biases that matter: A reply to commentaries. *Journal of Mathematical Psychology*, *57*(5), 190–195. DOI: https://doi.org/10.1016/j.jmp.2013.06.001

Galak, J., Leboeuf, R. A., Nelson, L. D., & **Simmons, J. P.** (2012). Correcting the past: Failures to replicate psi.

Journal of Personality and Social Psychology, *103*(6), 933–948. DOI: https://doi.org/10.1037/a0029709

Gelman, A., & **Loken, E.** (2014). The statistical crisis in science. *American Scientist*, *102*(6), 460–465. DOI: https://doi.org/10.1511/2014.111.460

Ioannidis, J. P. A., & **Trikalinos, T. A.** (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, *4*(3), 245–253. DOI: https://doi.org/10.1177/1740774507079441

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

Kruschke, J. K. (2010). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Academic Press/Elsevier Science.

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. DOI: https://doi.org/10.1177/1948550617697177

Lehmann, G. K., Elliot, A. J., & **Calin-Jageman, R. J.** (2018). Meta-analysis of the effect of red on perceived attractiveness. *Evolutionary Psychology*, *16*(4), 1–27. DOI: https://doi.org/10.1177/1474704918802412

McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC Press.

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. DOI: https://doi.org/10.1016/j.jmp.2013.03.004

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). DOI: https://doi.org/10.1126/science.aac4716

R core team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/.

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, *17*(4), 551–566. DOI: https://doi.org/10.1037/a0029487

Schonbrodt, F. D., & **Wagenmakers, E.-J.** (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*, 128–142. DOI: https://doi.org/10.3758/s13423-017-1230-y

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*(11), 1359–1366. DOI: https://doi.org/10.1177/0956797611417632

Simonsohn, U. (2012). It does not follow: Evaluating the one-off publication bias critiques by Francis. *Perspectives on Psychological Science*, *7*(6), 597–599. DOI: https://doi.org/10.1177/1745691612463399

Simonsohn, U. (2013). It really just does not follow, comments on Francis (2013). *Journal of Mathematical Psychology*, *57*(5), 174–176. DOI: https://doi.org/10.1016/j.jmp.2013.03.006

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance – or vice versa. *Journal of the American Statistical Association*, *54*, 30–34. DOI: https://doi.org/10.2307/2282137

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. *American Statistician*, *49*(1), 108–112. DOI: https://doi.org/10.1080/00031305.1995.10476125

Vandekerckhove, J., Guan, M., & **Styrcula, S. A.** (2013). The consistency test may be too weak to be useful: Its systematic application would not improve effect size estimation in meta-analyses. *Journal of Mathematical Psychology*. DOI: https://doi.org/10.1016/j.jmp.2013.03.007

Westermann, R., & **Hager, W.** (1986). Error Probabilities in Educational and Psychological Research. *Journal of Educational Statistics*, *11*(2), 117–146. DOI: https://doi.org/10.3102/10769986011002117

Peer review comments
The author(s) of this paper chose the Open Review option, and the peer review comments are available at: http://doi.org/10.1525/collabra.213.pr

How to cite this article: Francis, G. & Thunell, E. (2019). Excess Success in "Ray of hope: Hopelessness Increases Preferences for Brighter Lighting". *Collabra: Psychology*, 5(1): 22. DOI: https://doi.org/10.1525/collabra.213

Senior Editor: M. Brent Donnellan

Editor: M. Brent Donnellan

Submitted: 10 December 2018 **Accepted:** 21 April 2019 **Published:** 08 May 2019

UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Collabra: Psychology is a peer-reviewed open access journal published by University of California Press.

OPEN ACCESS