PURDUE
UNIVERSITY

April 11, 2019

Dear Editor and reviewers,

We would like to thank you for your useful feedback, which we feel have helped improve the manuscript. We address the comments below in red as well as in the manuscript, and we hope that these changes will meet your requirements.

Sincerely,
Greg Francis and Evelina Thunell

Editor comments:

1. Reviewer C makes a useful point about how this is a fairly atypical paper. I think there are ways to make the work more accessible to others and useful to the field. First, I think you need to explain the TES in more concrete terms. Do not assume readers will bother to read those other papers in much detail. Walk readers through the calculations, assumptions, and interpretations.

We have elaborated on the background and explanation of the TES in the Introduction, as well as added more details about the analysis of each of the studies both in the text and in Tables 1 and 2.

Second, describe the Dong et al. (2015) work in more detail. Provide a little more about their theorizing but spend more time on each study in their package and how it was designed and tested the key ideas. Explain the experimental design and report effect sizes. Identity the key statistical result that factors into the TES. Again, do not assume readers will bother to read the Dong et al. (2015) work before reading this paper. I think these two suggestions will help readers understand the original paper and provide a tutorial for how readers might apply the TES to other papers by having a concrete example.

We have now added more details to the summary of the Dong study in our Introduction, as well as a more detailed description of the design and results of each study together with our analysis. Indeed, this is useful when explaining which tests were considered relevant for our analysis. See Tables 1 and 2. We have not included standardized effect sizes because the tests are dependent and so they are not informative for the overall success rate. We added a comment about this in the caption of Fig. 1.

2. Related to #1, I found Table 1 hard to follow without consulting the original Dong et al. (2015) paper. (I read their paper before I read your manuscript so I think it was fresher in my mind than it will be for most readers. My memory is worsening as I get older but I do think Table 1 needs to be reworked to become more informative.)

To make the information more easily accessible, we have now put the success rate calculation directly in the text, and moved some of the other details to the new Tables 1 and 2.

3. I think the sections about new studies could be expanded. For example, do you think it would be useful to talk about pre-registration? What kinds of factors do you think replicators would need to focus on to make the direct replications close enough to the original studies to be considered informative follow-ups? What experimental manipulations are critical to reproduce? What kinds of conceptual replications would prove convincing? The current section focuses on sample size and I think this is a huge consideration but there other factors that could be addressed. Expanding the scope of this section will make for a more constructive contribution.

We now elaborate more on the suggestions for future studies, and discuss the intrinsic problems related to requiring many successful outcomes. We are not specialists in conceptual metaphor theory or social psychology, so we are not in a good position to suggest exactly what are the necessary tests required for the claims about links between brightness perception and hopelessness. Instead, we now expand on the general advice and leave the details for experts in the field.

Regarding pre-registration, it can be useful for many reasons, but is not directly related to power and would not remedy any of the issues brought up in the current manuscript. Therefore, we do not see a natural way of including it in the discussion.

4. Reviewer B is the most critical and raises good points. Please take these seriously. I do think you can point to 2014 paper in Psychonomic Bulletin & Review as a place where the TES was applied to a whole universe of studies. You might have a section about the TES and how it has been used in previous studies. I do think there are cases where the TES flagged a paper and then studies failed to replicate. So that would be useful to note if possible as additional support for the method. You might also need to explain again to readers of this paper the characteristics of papers that work well for TES such as four or more studies. A few references to criticisms of the TES would seem to be appropriate.

We will address the specific concerns of reviewer B below. We did not discuss the Francis (2014) paper (analysis of articles published in Psychological Science) because we felt it was largely irrelevant to the current discussion. Likewise, we do not discuss the “four or more studies” issue because (while it made sense for the investigations in Francis (2014)), it is not a general property of the test.

We now do mention a previous application of the TES and how it led to additional studies and analyses. In particular, application of the TES to a study by Elliot et al. motivated researchers to discover that the effect size seems to be much smaller than originally reported. We use the example to make the general point about what a TES analysis can conclude.

We now also mention some criticisms of the TES method and replies to these criticisms; and the text briefly addresses one of the main concerns (publication bias for TES investigations). This *is* a topic that seems to confuse critics of the TES, so it is good to address it up front.

5. Reviewer A was the most positive reviewer. This reviewer noted that the Test of Insufficient Variance also flags the Dong et al. (2015) work. This could be useful to include and might help mitigate some of the concerns raised by Reviewer B.

These results are reassuring. However, we are not sure whether the test of insufficient variance is applicable to the data set in Dong et al. Given that there are multiple relevant tests in several experiments, we are not sure how to pick one for the TIVA analysis and (as far as we can tell) it is not appropriate to include more than one test from an experiment. Thus, we are reluctant to include a TIVA analysis in our manuscript.

6. Here are some random quibbles and things I noted in the margins while reading the paper.
a. I don't love the phrase "too good to be true" in the Abstract. I think it is possible to phrase that idea more neutrally and probabilistically – as in "The likelihood of this pattern of results is quite low in light of the set of assumptions underlying the test for excess success"

We have reformulated this phrase throughout the manuscript.

b. When you use the term biased on page 9, I would explain to readers a bit more about the meaning of such a claim.

We tried to explain what we meant in the subsequent text, and have now slightly rephrased to make this more obvious to the reader.

c. When talking about the requirements of setting beta = .04 or power to .96, it might be useful to evaluate if such a suggestion is reasonable in light of the design demands of the particular studies. It isn't incredibly hard to get large samples of college students for researchers at large research schools but other kinds of samples are much harder to recruit. So I think critically evaluating whether setting power to .96 is general feasible would be a worthwhile addition.

We agree that this is an important issue, but it seems a bit tangential to the current situation (where subjects are pretty easy to recruit). We have expanded the section where we address this specific issue, and now discuss the intrinsic problems related to requiring many successful outcomes. Since it is indeed often not reasonable to have such high power as 0.96, we suggest, as a general rule, to always simplify/limit the design and required patterns of results as much as possible.

Reviewer A:

1) General comments and summary of recommendation
Describe your overall impressions and your recommendation, including changes or revisions. Please note that you should pay attention to scientific, methodological, and ethical soundness only, not novelty, topicality, or scope. A checklist of things to you may want to consider is below:
- Are the methodologies used appropriate?
- Are any methodological weaknesses addressed?
- Is all statistical analysis sound?
- Does the conclusion (if present) reflect the argument, is it supported by data/facts?
- Is the article logically structured, succinct, and does the argument flow coherently?
- Are the references adequate and appropriate?:
Yes, everything is ok

2) Figures/tables/data availability:
Please comment on the author's use of tables, charts, figures, ifrelevant. Please acknowledge that adequate underlying data is available to ensure reproducibility (see open

data policies per discipline of Collabra here).:
yes

3) Ethical approval:
If humans or animals have been used as research subjects, and/or tissue or field sampling, are the necessary statements of ethical approval by a relevant authority present? Where humans have participated in research, informed consent should also be declared.
If not, please detail where you think a further ethics approval/statement/follow-up is required.:
NA

4) Language:
Is the text well written and jargon free? Please comment on the quality of English and any need for improvement beyond the scope of this process.:
Yes

Reviewer B:

1) General comments and summary of recommendation
Describe your overall impressions and your recommendation, including changes or revisions. Please note that you should pay attention to scientific, methodological, and ethical soundness only, not novelty, topicality, or scope. A checklist of things to you may want to consider is below:
- Are the methodologies used appropriate?
- Are any methodological weaknesses addressed?
- Is all statistical analysis sound?
- Does the conclusion (if present) reflect the argument, is it supported by data/facts?
- Is the article logically structured, succinct, and does the argument flow coherently?
- Are the references adequate and appropriate?:
This article format by Greg Francis is well-known, and the flaws against publishing criticisms on single articles have been well-articulated elsewhere. Furthermore, performing analyses of bias in sets of studies is very common (although the only interesting papers use clear unbiased inclusion criteria, such as all papers selected for a meta-analysis). Bias tests like this one have lead to the widespread insight there is a real problem with bias in the literature, and the current manuscript does not explain why we need another demonstration. The most important take home message in this article for me is that many authors do not listen to criticisms. In this case, both Dong, Huang, and Zhong (2015) as the current authors ignore the criticisms in the literature.

We disagree with some of the reviewer's characterizations of what we have done, criticisms of this kind of work, and the nature of scientific discourse. We address specifics below, where the reviewer returns to each of these points.

The current authors keep using scientifically imprecise and meaningless statements (most notably, "However, the results seem too good to be true." in the abstract) and

(See also reply above)
We have reformulated this phrase throughout the manuscript.

do not cite any of the published criticisms on their approach. The most important problem I personally see is Francis himself commits publication bias – I have not seen him publish a paper analysing a single article where the Test for Excessive Significance was not statistically significant, even though he must have a huge file-drawer of papers where he tried, but did not find a significant effect. I think it should be discussed in this manuscript.

Criticism regarding the existence of a file drawer of TES analyses is based on a misunderstanding of the interpretation of that file drawer. One of us (Francis) has responded to this criticism multiple times in published papers. Nevertheless, as this might be an issue for readers who are not familiar with this work, we added a paragraph discussing this issue to the introduction.

The reason this should be discussed is that the authors argue that the authors could be lucky, but that it is more plausible that there is publication bias. I disagree. This point depends completely on the file-drawer of Greg Francis. If he tried 10 papers from SPPS and this was the only one that was flagged by TES (which has a 10% error rate) than it is actually more plausible that the authors were just lucky (I am ignoring priors here, since TES is a frequentist method – arguably if we included priors this logic would shift). So, it is difficult for me to believe the authors because they do not share their own file drawer.

I personally think that the current article can not be published unless it is accompanied by a complete overview of all TES analyses performed by Greg Francis, including a careful analysis of the decision rule that is used to select studies. However, even if this happens, it is difficult to know if the reported file-drawer and selection plan are accurate. TES analyses of single papers should be pre-registered to be believable. In this sense, I think (Francis, 2014) was probably the only good TES manuscript published – it has a clear decision rule, and no file drawer. Single study versions of TES papers can not be interpreted without a complete overview of the file-drawer and an unbiased decision rule used to select articles. This is obviously all explained in detail in the articles criticizing the TES approach that the authors chose not to cite. We need a file-drawer and selection rule to see if this study is truly as surprising as the p-value suggests. We all know (among others based on the work by Francis himself) that p-values from a biased set of analyses can not be trusted.

We believe that there is a major error in the reviewer's reasoning. Namely, a file drawer is a problem only when *relevant* studies/results are not reported. A file drawer (no matter how big) does not affect the interpretation of *unrelated* sets of studies. It is for precisely this reason that we focus on studies in a single article that were identified by the original authors as being relevant for their conclusions. While excess success in one paper about brightness perception cannot be used as evidence for excess success in the whole field of brightness perception, it *can* be used as evidence for excess success in that particular paper because the paper is supposed to report all relevant conducted experiments. In other words, we are not investigating this one paper in order to make inferences about a population of other publications. Rather, the experiments included in the Dong article constitute the full population of interest and accordingly we draw conclusions about that set. For example, if a multi-study paper about afterimages passed the test for excess success, that would not in any way change the interpretation of the Dong et al. analysis.

The file drawer *does* restrict what kinds of conclusions we can draw from the TES analysis. For example, if we had concluded that Dong et al. generally practice QRPs, then it would indeed be problematic had we analyzed more of their articles and put in a file drawer any TES analyses that did not indicate excess success. In that setting, the file drawer would contain information that is relevant to the conclusion. However, our conclusions are only about the specific set of studies we analyzed. We cannot (and do not) draw conclusions about other studies or about authors more generally. The conclusion that we can draw is

interesting because it is about exactly the set of studies used by the original authors. We discuss all of this in the revised text.

As noted also above, we now do cite criticisms of the TES analysis together with replies and also directly address the main concern about publication bias.

Dong et al (2015) also ignore the literature, and continue to commit publication bias (based on a recent retraction of another paper by the original research team, the problems might very well be more substantial).

This is true, but we do not feel it is directly relevant to the current analysis.

The section 2 on Designing New Studies should probably be deleted. It is well known that power analysis should not be based on pilot studies (for a recent explanation see Albers & Lakens, 2018 and the cited references therein). If it is attempted, bias corrected effect sizes should be used, but smallest effect sizes of interest are a better approach. The suggestion to use Bayesian analyses or equivalence tests is of no help – if sample sizes are too small, these will also be uninformative. This whole section can basically be summarized as 'perform a power analysis' but then how the authors do this is actually not good advice on determining sample sizes for future research in this manuscript.

We tend to disagree with the reviewer. First, in this section we are being proactive to help scientists design better experiments, and we think our advice is not widely known. Second, we are well aware of the dangers of using pilot studies for power analyses, but surely (when properly interpreted) such approaches are better than nothing at all. Third, we don't know of any way to compute "bias corrected effect sizes" here. Likewise, Dong et al. did not report "smallest effect sizes of interest", so that does not seem to be an option for sample size planning. Fourth, one of our main points is that telling researchers to "perform a power analysis" does not really work well because researchers misunderstand what needs to be done. When multiple tests are involved, the power analysis is rather complicated. We wanted to step through the whole process so that readers could see what is involved. Fifth, the recommendation for using Bayesian or equivalence tests was to deal with finding support for the null. The reviewer is correct that small samples will not help here; but we do not believe that these methods are "of no help". Without those methods there is no sample size that will produce high power; with those methods it is at least possible to have high power for large sample sizes.

Minor points:

An interesting question worth discussing is if findings reported in supplementary materials should be included in the TES analysis (Study 5)?

As Study 5 is placed in the supplementary material only due to space constraints, and is said to corroborate studies 2 and 3 we believe it should be included in the TES analysis. We have added a note about this in the manuscript.

The authors do not discuss whether they approached the authors and ask them for the data, to resolve issues in for example footnote 2. Did they ask for the data? That seems required to attempt to make the most informed argument in this manuscript.

We contacted both the corresponding and first author after receiving this recommendation from the reviewer. After more than a month, we have not received a reply. We mention this in footnote 2. As the text notes, having that data would only further reduce the success probability for the study.

Signed,
Daniel Lakens

Albers, C., & Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. Journal of Experimental Social Psychology, 74, 187–195. https://doi.org/10.1016/j.jesp.2017.09.004
Francis, G. (2014). The frequency of excess success for articles in Psychological Science. Psychonomic Bulletin & Review, 21(5), 1180–1187. https://doi.org/10.3758/s13423-014-0601-x

2) Figures/tables/data availability:
Please comment on the author's use of tables, charts, figures, ifrelevant. Please acknowledge that adequate underlying data is available to ensure reproducibility (see open data policies per discipline of Collabra here).:
not applicable

3) Ethical approval:
If humans or animals have been used as research subjects, and/or tissue or field sampling, are the necessary statements of ethical approval by a relevant authority present? Where humans have participated in research, informed consent should also be declared.
If not, please detail where you think a further ethics approval/statement/follow-up is required.:
NA

4) Language:
Is the text well written and jargon free? Please comment on the quality of English and any need for improvement beyond the scope of this process.:
It is well written

Reviewer C:

1) General comments and summary of recommendation
Describe your overall impressions and your recommendation, including changes or revisions. Please note that you should pay attention to scientific, methodological, and ethical soundness only, not novelty, topicality, or scope. A checklist of things to you may want to consider is below:
- Are the methodologies used appropriate?
- Are any methodological weaknesses addressed?
- Is all statistical analysis sound?
- Does the conclusion (if present) reflect the argument, is it supported by data/facts?
- Is the article logically structured, succinct, and does the argument flow coherently?
- Are the references adequate and appropriate?:
The present manuscript uses simulation data to examine the plausibility of the claims reported in Dong, Huang, and Zhong (2015). These simulation data show that the probability

of finding the results reported in Dong et al. (2015), assuming the effect size estimates are accurate and the same sample sizes are used, is .016. I felt that this paper had a lot of strengths: the simulation approach was clearly explained, the assumptions were conservative (which I think is important when the evidentiary value of a set of studies is being questioned), and the conclusions seem reasonable and well-calibrated. The comments I make below are somewhat subjective reactions I had while reading the paper that, if addressed, could make the paper stronger in my opinion.

Specific points

1. This type of paper is somewhat unusual in my experience. I've seen papers that propose general techniques for determining the evidentiary value in a set of studies (e.g., the p-curve papers, the excess success tests). I've also seen papers that challenge the robustness of a particular finding, but usually they involve new data (e.g., a replication study). Finally, I've seen data forensic techniques applied to specific findings, but usually I see these in the form of blog posts (e.g., the Data Colada post on power posing). In principle, I don't see any reason why the current type of paper shouldn't be published as a journal article, but I am still left wanting to know more about the contribution of this work (e.g., Why was it important to re-examine this particular finding? Are there more general lessons to be learned from the present article that couldn't be learned in other articles describing excess success tests?)
(See reply also above)

Like the reviewer, we believe that it is indeed useful for other scientists interested in the same topic to point out apparent flaws in studies, and that it is educational for the field to see how questionable research practices can influence the credibility of the results. We also believe that every research paper should potentially be open to post-publication evaluation. We do not really have a good answer for "why this particular finding?". In our view, every paper can be re-examined; that's part of science.

As for general lessons, we feel that our manuscript provides some new insights into how power analyses should be done (considering all the relevant tests can greatly reduce estimated power of an experiment).

2. This point is more subjective than the first, thought the two are related. I have the impression that the manuscript was carefully written to avoid accusations directed at the original authors. Still, there are some suggestions that those authors must have done their research poorly or engaged in questionable research practices. I don't necessarily think those conclusions are unjustified, but I do think that rhetorically these kinds of targeted criticisms have more impact when they err on the side of being (perhaps overly) diplomatic.

We have removed the phrase "good to be true". Indeed, when publishing this type of criticism, it is important to not be rude. We feel that we have been as diplomatic as possible without undermining our strong claims.

2) Figures/tables/data availability:
Please comment on the author's use of tables, charts, figures, ifrelevant. Please acknowledge that adequate underlying data is available to ensure reproducibility (see open data policies per discipline of Collabra here).:
Use of tables and figures is appropriate.

3) Ethical approval:
If humans or animals have been used as research subjects, and/or tissue or field sampling, are the necessary statements of ethical approval by a relevant authority present? Where

humans have participated in research, informed consent should also be declared.
If not, please detail where you think a further ethics approval/statement/follow-up is required.:
No human subjects data.

4) Language:
Is the text well written and jargon free? Please comment on the quality of English and any need for improvement beyond the scope of this process.:
The paper is well-written with minimal jargon.

Collabra
http://www.collabra.org/
@collabraoa

Dr. Thunell-

I am sending along a signed comment from one of the reviewers. I just issued the revise decision but this may not have been appended. Here it is.

Sincerely,

-brent donnellan

This article reports the results of a bias analysis of an original article on hopelessness and dim lighting.

I conducted an alternative bias analysis with the Test of Insufficient Variance and arrived at the same conclusion

S1: t(179) = 2.62
S2a: r(143) = .22; t(141)=2.68
S2b: r(57) = .24; t(55) = 1.83
S3: t(201) = 2.07
S4: t(104) ¼ 2.15
S5: r(59) = .27, t(57)=2.12

==========

```
df = c(179,141,55,201,104,57)
t = c(2.62,2.68,1.83,2.07,2.15,2.12)
z = qnorm(pt(t,df))
z
var.z = var(z)
var.z
pchisq(var.z*5,5)
> pchisq(var.z*5,5)
[1] 0.01004873
```

I leave it to the authors whether they want to include this confirmatory analysis or not.

I think reporting the results of this bias analysis is important and makes a valuable contribution to the literature.

https://replicationindex.wordpress.com/2014/12/30/the-test-of-insufficient-variance-tiva-a-new-tool-for-the-detection-of-questionable-research-practices/

Ulrich Schimmack

(See reply also above)
These results are reassuring. However, we are not sure whether the test of insufficient variance is applicable to the data set in Dong et al. Given that there are multiple relevant tests in several experiments, we are not sure how to pick one for the TIVA analysis and (as far as we can tell) it is not appropriate to include more than one test from an experiment. Thus, we are reluctant to include a TIVA analysis in our manuscript.