Decision letter

Date: 14 Apr 2015 23:03:10 +0100
From: Journal Cognition <cognition@elsevier.com>
To: wickelmaier@web.de
Subject: Your Submission COGNIT-D-15-00115

Ms. Ref. No.: COGNIT-D-15-00115
Title: Better think thrice: A comment on Costa, Foucart, Arnon, Aparici, and Apesteguia (2014)
Cognition

Dear Dr. Florian Wickelmaier,

Reviewers' comments on your work have now been received. Neither reviewer recommends publication. This is a hard call. I do not agree with Reviewer 1 that your argument is completely off target. One can imagine a case in which an effect appears in one condition with p = .049 but "disappears" under a different treatment with p = .051. In such a case, I would entirely agree with you that focusing on the difference of differences is necessary. But I do think the reviewer makes other valid points such as the importance of the theoretical motivation. And the question about what constitutes a failure to replicate is an important one. What is clear from Reviewer 1's discussion is that there are deeper issues at stake.

I think Reviewer 2 addresses your criticism in a reasonable way. Meta-analysis seems an appropriate tool for addressing your question. You may be thinking that the meta-analysis should have appeared in the original article, and I would agree. That is as much the journal's fault as the authors'. The reviewer does refer to a variety of data that convinces me the Costa et al. effect is real.

In the end, given that I'm convinced the effect is real, I'm just not sure that the community would benefit from this interchange. I think there are deep issues of statistical analysis at stake, but the current version of your paper doesn't grapple with those issues and this isn't the journal for such a discussion anyway. So, while I think you have a valid complaint, I'm not convinced that a back-and-forth in the pages of this journal would be terribly enlightening about the effect under discussion. So I'm sorry to say that I've decided not to accept your paper.

For your guidance, I append the reviewers' comments below.

Thank you for giving us the opportunity to consider your work.

Yours sincerely,

XXX
Editor-in-Chief
Cognition

Journal Policy Statement: Editorial decisions are final, and unsolicited resubmissions cannot be considered for publication.

Reviewers' comments:

Reviewer 1

This comment focuses on the first five experiments reported in Costa et al.'s 2014 Cognition paper. It argues that the statistical analyses are inadequate and that therefore their conclusions regarding reduced loss aversion in a foreign language are invalid. In my opinion, while the comment is technically correct regarding the results of the first Asian disease experiment, its overall conclusions are incorrect. In a nutshell, while the paper is arguing that Costa et al. are committing a type I error, I believe that the comment's selective analyses are much more likely to lead to a type II error. There is a strong bias in the field these days to pay almost exclusive attention to type I errors, but the harm to science from increasing type II errors is substantial, leading to the dismissal of real and important phenomena. Here is where the comment is technically correct. In the first Asian disease experiment, Costa et al. found a larger numerical difference between losses and gains in the native language condition than in the foreign language condition; in other words, a numerically larger framing effect. While both languages showed a significant framing effect, the authors concluded that the effect is reduced in a FL. It is technically correct that in this particular experiment one should compare the difference in the differences, because there was a significant framing effect in both language conditions.

But the comment goes beyond this technical point, and draws sweeping conclusions that are not warranted for several reasons. The comment has limited applicability for the Costa et al. paper, mainly because the argument does not hold for most of the studies. For example, in the second Asian disease study there was a significant framing effect for NL but not for FL. It is perfectly valid to conclude that the framing effect was reduced in FL (if not that it disappeared). Applying the comment's test in this case is overly conservative and obscures the fact that there was no significant effect in FL.

This is particularly important because the prediction is directional and the reduction is in the predicted direction. This point is really important, and much more general, so let me use a related situation to illustrate it. Think about the NL condition as demonstrating an effect, and the FL condition as attempting to replicate it. Many researchers these days attempt to replicate published studies and sometimes they fail. These failures to replicate are taken as evidence against the existence of the phenomena they attempted to replicate. If you apply the logic of the current comment, you should only doubt the original phenomena if you conduct a statistical test to show that the difference in the differences is significant. It is not enough to show that there is no effect in the study that attempted to replicate it. This does not make sense, but it is a logical conclusion from the current comment.

The comment also misses the theoretical motivation for predicting when one should expect a difference and when not, which makes the Costa et al. paper a lot more nuanced and much more convincing than the current comment would lead one to believe. For example, the study that used the cognitive reflection test found no foreign language effect, which is consistent with the theoretical assumption that the effect is motivated by reduction in emotional reaction (which is irrelevant to the CRT). I also disagree with the approach to scientific inquiry that the comment promotes. It creates the impression that there is only one valid way to analyze data. Statistics is a toolkit that provides a variety of tools. The difference between the approach of the paper by Costa and the approach of the comment is not one between valid and invalid approaches. It is the difference between more or less conservative methods. Each one has its place, but using an overly conservative approach requires justification. It runs the risk of missing out on important discoveries, as is the case here. Costa et al's paper presents a consistent pattern of results across many studies, that even when not always significantly different by the most conservative test, is always there when it should be, always in the predicted direction, and it is always absent when it should be absent. The main question for me is, does the comment warrant the conclusion that Costa et al.'s claims are not supported by the data? In my opinion, just as with any published paper, one could quibble with some technical aspects of the analyses, and the report could have been perhaps more clear about the theoretical motivation, but this does not even come close to supporting the conclusions of the comment. I therefore don't believe it makes a contribution that merits publication in Cognition.

Reviewer 2

This commentary makes two points regarding the claims made in Costa et al.

First, Costa et al. "fail to provide a statistical test of the foreign-language effect for their first five studies". Second, "the conclusion rather is that there is not much evidence in favor of a foreign-language effect in any of the five studies." Let´s see what the evidence says.

1. Regarding the first point, it is true that a significant effect of a given factor plus a no significant effect of another factor does not mean a significant effect of the interaction. So, we acknowledge that Wickelmaier is right in that we do not show in the paper a direct positive test of such global interaction effect. However, this does not mean that the interaction does not actually exists in our data. We can argue about the advantages and shortcuts of the several alternative methods for analyzing the data. However, the important point here is whether there is evidence of an interaction between the framing effect and the language in which the problem is presented and analyzed. And Wickelmaier is wrong in his conclusion that such interaction does not show up in the data, as we will show below.

   The global conclusion about whether there is actually an interaction in our studies should be reached by meta-analytic methods. As we are dealing with several separate studies the best suited method is the combined estimate of the effects. As in such procedure the studies are weighted by the inverse of their variances, the studies with larger samples have higher weights in the synthesized estimate. Conventional statistical procedures are incorrect for this kind of analyses because the effect sizes of the studies have different variances, and the homocedasticity assumption is not accomplished.

   We have now calculated the LogOR of each language condition from each study, as our effect size index. The variance of each of those values is estimated as (Lipsey & Wilson, 2001),

$$Var_{LogOR} = \frac{1}{N_a} + \frac{1}{N_b} + \frac{1}{N_c} + \frac{1}{N_d}$$

where *a*, *b*, *c* and *d* are the four cells in the contingency table. The studies with the Ticket/Money problem and the Discount problem shared the same sample of participants, so that those two effect size estimates are not independent. In order to manage that situation we have followed the suggestion of Borenstein, Hedges, Higgins and Rothstein (2009) of averaging the two estimates obtained with the same sample but with different outcomes or tasks. The average LogOR in that sample is 1.287 for the native language group and 0.7186 for the foreign language group[1]. We

1 The calculation of the estimated variance of the average effect size needs the value of the correlation between the two measures. As that value is not available, we have assumed a medium value of r=.50; however, with the minimum value of r=0 or the maximum value of r=1 our conclusions do not change.

have adjusted a random-effects model, as it is more conservative and realistic than a fixed-effect model. In our small meta-analysis this is especially important, as the tasks are different although they look for the same effect. Nevertheless, in this particular case the results mimic those of the fixed-effect model, as the estimated specific variance of the effects equals zero. When fitting a model with the language as a dichotomous moderator the effect of such moderator is statistically significant: $Q_b(1) = 4.724$, $p = .030$. The combined effects are in the expected direction: $OR_{native} = 3.155$; $OR_{foreign} = 1.755$. A statistically significant OR means that there is evidence of a framing effect, while the significant effect of the language means that the framing effect is significantly different in each language. That is, the meta-analytic analysis shows that there is a statistically significant interaction.

At a descriptive level, the logged OR is higher for the native than for the foreign language condition in all four studies. However, it could be argued that we do not have 8 separate and completely independent estimates, as each study provide one estimate of each language condition. Although it is not probable any dependency to explain the interaction, we have reanalyzed the data calculating for each study the following effect size index,

$$Dif_{LogOR} = LogOR_{Native} - LogOR_{Foreign}$$

The expected value of such index is zero under the null hypothesis that the framing effect is the same for both languages. As the design on each study involves two-independent groups, the variance of such index equals,

$$Var_{Dif_{LogOR}} = Var(LogOR_{Native}) + Var(LogOR_{Foreign})$$

The combined effect size from the four studies equals 0.6028 and the transformation back to OR equals 1.827 [CI95%: 3.108; 1.074]. So, the set of four studies provides statistically significant evidence that the language moderates the framing effect: the effect is larger with the native language than with the foreign language. Hence, we think that the author's conclusion "the conclusion rather is that there is not much evidence in favor of a foreign-language effect in any of the five studies" is wrong.

2. The second point is even more problematic in our view. The author concludes "the conclusion rather is that there is not much evidence in favor of a foreign-language effect in any of the five studies.", seems to imply that we did not find any significant effect of a foreign language in our article. We think that this conclusion is somewhat misleading. In our article, we present 9 studies (not only 5). Indeed the results of the 4 studies that are not mentioned by the author represent the bulk of the contribution of Costa et al. Namely, the establishment of the foreign language effect in other settings that are

not related to framing, but have to do with other fundamental aspects of decision-making (risk, uncertainty, etc). Furthermore, since these are direct studies of the foreign language effect (in particular, they do not involve the comparison of differences), these are not subject to the critique of the comment.

Regarding, the studies that the author comments on (those related to framing effects) We think that the author also fails to put them in context. In these studies, especially in the first three, we aim at replicating three other studies that have shown a foreign language effect in decision making in the context of framing effects. These studies were conducted by Keysar and collaborators. Hence, we did have a clear hypothesis about the directionality of the effect, namely that framing effects would be smaller when problems are presented in a foreign language than when problems are presented in a native language. This is precisely, in numerical terms, what we observed. In four of the five studies, the magnitude of the framing effect was twice larger in the native than in the foreign language. Consider for a minute, that with these results we would have argued that Keysar et al, was wrong and that as the author claims, we do not find evidence of a foreign language effect. That is, we should argue that our consistent pattern of results was due to chance (chance that as it happen leads to a systematic pattern consistent with what we know about the phenomenon). One wonders why the author of this commentary has failed to put in context our study, obviating research presented in our article and previous very relevant research.

3. The tone of the article, including the title, does not seem to be appropriate for a scientific exchange.

Given these considerations we are far from convinced that this commentary will in any way help to advance in our knowledge of the presence of a foreign language effect on decision making. If the author wants to pursue this avenue, he may want to take a more rigorous path and consider all the existence experimental evidence together.

References

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: John Wiley and sons.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Letter to the editor

Date: Tue, 21 Apr 2015 18:48:24 +0200 (CEST)
From: florian.wickelmaier@uni-tuebingen.de
To: Journal Cognition <cognition@elsevier.com>
Subject: Re: Your Submission COGNIT-D-15-00115

Ms. Ref. No.: COGNIT-D-15-00115
Title: Better think thrice: A comment on Costa, Foucart, Arnon, Aparici, and Apesteguia (2014)
Cognition

Dear Prof. XXX,

Thank you for considering my comment. I, of course, accept your decision to not publish my comment but feel like there is an issue that should be addressed nonetheless.

The intention of my comment was to correct a serious statistical mistake: Costa et al. repeatedly claim the existence of an effect but do not provide a statistical test to support their claims. My comment is not about whether or not the effect exists (it may or may not), my comment is about the missing statistical test. I do not agree with the view that being convinced an effect is real relieves a researcher from statistically testing it.

Both reviewers admit that such a test is missing. Reviewer 2 proposes a test that is based on the difference of log odds ratios. My own proposal is based on the ratio of odds ratios, so the idea is similar. A test as suggested by Reviewer 2 or by myself should have been included in the original article, and I find it difficult to understand how the absence of it went unnoticed by the original reviewers.

I do not share your view that there are deep issues of statistical analysis at stake. I do think that the issue is rather simple: effects are claimed without being tested. This constitutes a severe scientific mistake. I am convinced that it is in the best interest of Cognition and of the scientific community to have this mistake corrected.

Your decision is to not publish my attempt at providing such a correction, but I urge you to correct this mistake nevertheless. If it is not done by a comment, it should be done by an erratum to the

original article in which the authors provide an analysis of the interaction effects, possibly along the lines suggested by Reviewer 2.

I will give three reasons why I think an erratum is necessary:

(1) There is no doubt that the authors made a severe mistake, and the original reviewers overlooked it. Since the mistake is obvious, chances are, that others will detect it as well. You might periodically receive complaints like mine. Thus, it seems desirable to settle the issue once and for all by correcting this mistake.

Maybe it is my fault that I did not clearly enough state that the authors did something objectively wrong. But it is a well known, widespread and serious statistical mistake. This is not a matter of opinion. I was requesting Costa and Keysar as reviewers because I believe they are especially in need of learning about this mistake. But had you sent this to a statistician or a methodologically trained psychologist the answer would just have been: "Oh no, did they really publish this? Not again." Or as Gelman and Stern (2006) put it:

"As teachers of statistics, we might think that 'everybody knows' that comparing significance levels is inappropriate, but we have seen this mistake all the time in practice." (p. 328)

One might have hoped that almost ten years later the situation had improved. Sadly, however, Reviewer 1 still makes the same mistake: "For example, in the second Asian disease study there was a significant framing effect for NL but not for FL. It is perfectly valid to conclude that the framing effect was reduced in FL (if not that it disappeared)." This illustrates how difficult it seems to be for statistical knowledge to enter this community. And it underlines how necessary my comment is.

Maybe I failed at clearly stating that comparing p-values instead of testing the interaction is a mistake, not only when p-values are close, but also when they are very different:

"By this, we are not merely making the commonplace observation that any particular threshold is arbitrary--for example, only a small change is required to move an estimate from a 5.1% significance level to 4.9%, thus moving it into statistical significance. Rather, we are pointing out that even large changes in significance levels can correspond to small, nonsignificant changes in the underlying

quantities." (Gelman & Stern, 2006, p. 328)

Gelman and Stern (2006) provide a compelling demonstration:

"Consider two independent studies with effect estimates and standard errors of 25 +/- 10 and 10 +/- 10. The first study is statistically significant at the 1% level, and the second is not at all statistically significant, being only one standard error away from 0. Thus, it would be tempting to conclude that there is a large difference between the two studies. In fact, however, the difference is not even close to being statistically significant: the estimated difference is 15, with a standard error of sqrt(10^2 + 10^2) = 14." (p. 328)

(2) I am afraid that the damage done to the scientific community by not correcting such mistakes is larger than you might admit, especially for the next generation of scientists. Let me illustrate this: We came across this paper in a lab course on experimental psychology. We, the teachers, thought: This looks like a nice study for students to try on their own, so they did using German and English as a new language pair. On a second look, however, I discovered that the proposed analysis was wrong. I told my students that this must have been a mistake, and I instructed them on how to test for an interaction. What am I supposed to tell my students now? My original intention in sending this comment was to correct this mistake so future generations of students would not come across it anymore. Honestly, I am really surprised that this led to a discussion of underlying effects and their existence and not to a simple correction of the error.

Related to this, I see an even broader risk for future Ph.D. students planing to start a career based on the original article. It is tempting to design a series of experiments investigating the foreign-language effect with different language pairs, e.g., German and English, Swedish and Finish, etc. These students might take the original article as a model and repeat the wrong analysis over and over again. One might object that reviewers will make them aware of the error. But then it might be too late: They will have based their sample size calculation on the wrong analysis, so their sample size will be too small for detecting an interaction even if it existed.

(3) An erratum would be a single brief statement, there would not be a back-and-forth of opinions. It would not be embarrassing for the

authors; they might even get credits for presenting a state-of-the-art method of analysis.

Reviewer 2 starts out by asking: "Let's see what the evidence says." This is exactly the right question. It should have been asked long time ago by the authors, by the original reviewers, and by the editor. I believe it is not too late for an answer, and the authors should provide it in the erratum.

Let me conclude by stating that I was approaching you with this issue out of my honest intentions to serve the scientific community. It is not my intention to embarrass anybody. Errors happen of course all the time, and we should learn from them. Whether or not the effect exists may be a matter of opinion. What constitutes a failure to replicate may be another matter of opinion. However, that you cannot compare p-values instead of testing for an interaction, is no matter of opinion, but a statistical truth. Doing so nevertheless is an error.

I hope my arguments convinced you that the error cannot just be ignored and that the community deserves a clarification. In order to maximize its usefulness, a correction should be linked to the original article in which the error occurred. Therefore, it should be published in Cognition and not elsewhere. Let us all work together so that a mistake like this will not happen again in the future. I would be happy to contribute by serving as a reviewer of the erratum.

Yours sincerely,

Florian Wickelmaier

Reference

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. American Statistician, 60, 328-331. doi: 10.1198/000313006X152649

Response from the editor

Date: Mon, 27 Apr 2015 07:36:54 -0400
From: XXX
To: florian.wickelmaier@uni-tuebingen.de
Subject: Fwd: FW: Your Submission COGNIT-D-15-00115

Dear Prof. Wickelmaier,

Your note clarified a misunderstanding that I had. Now that I understand that the analysis you were complaining about concerned data from within a single experiment, I agree with you that it was a serious error. The difficult statistical issue I referred to arises when comparing effects between experiments.

You have convinced me that there's a serious problem with Costa et al.'s analysis. But I also remain convinced by his subsequent analyses that he has a real effect. I think your suggestion to have him write an erratum (or a corrigendum) was an excellent one. I've been in touch with him and, after some back and forth, I've allowed him 10 days to produce a short corrigendum. I've asked him to acknowledge your input in his text. If he does not come through by Fri, May 8, my plan is to give you the opportunity to write a very short note reminding everybody of the statistical issue, perhaps mentioning Costa et al. as well as any other recent perpetrators of the error, without attempting to undermine the findings of their paper.

thanks for your input,
XXX