CORTEX 74 (2016) 348–352

Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Reply

Challenges in replicating brain-behavior correlations: Rejoinder to Kanai (2015) and Muhlert and Ridgway (2015)

Wouter Boekel[*], *Birte U. Forstmann and Eric-Jan Wagenmakers*

University of Amsterdam, The Netherlands

1. Introduction

In a recent study we attempted to replicate five studies that had previously reported a combined total of 17 significant structural brain behavior (SBB) correlations (Boekel et al., 2015). We preregistered our analysis plan and used confirmatory Bayesian hypothesis tests to quantify the evidence that our data provided for the presence or absence of the SBB correlations. For about half of the 17 SBB correlations that we set out to replicate the data suggested at least moderate evidence for their absence, and for 16 out of the 17 correlations the data produced no evidence for their presence. Subsequent exploratory analyses using Bayesian parameter estimation and a Bayesian replication test sketched a more nuanced perspective of the replication results. Nevertheless, our overall results suggest that confirmatory replication studies in the cognitive neurosciences deserve a more prominent role.

Our confirmatory replication has attracted two commentaries from researchers who are skeptical about our results. In "Failed replications, contributing factors and careful interpretations", Muhlert and Ridgway (2015) critique our replication attempt for having low sample size and incomplete correction for nuisance variables. In addition, they note that there are differences in the VBM processing pipelines used by the original authors and those used in the replication attempt, and suggest that these differences may have contributed towards the discrepant results.

In "Open questions in conducting confirmatory replication studies", Kanai (2015) also critiques our replication approach and points out that our confirmatory ROI approach may underestimate the SBB correlations. In addition, Kanai feels that the process of refereeing a preregistered study demands clearer guidelines.

We wish to thank the discussants for their interesting suggestions and constructive comments. At the moment little guidance exists with respect to the design and interpretation of purely confirmatory replication studies in the cognitive neurosciences, and we hope this discussion can help stimulate the development of common goals and guidelines.

Below, we discuss the key concerns raised in the commentaries. We also suggest ways in which future replication studies can take into account the issues raised by these commentaries.

2. Concern 1: low sample size

Both Muhlert and Ridgway (2015), and Kanai (2015) point out that the sample size of our replication attempt was lower than the sample sizes of the original findings for 16 out of 17 effects. For 9 out of these 17 effects, our data remained evidentially ambiguous (i.e., $1/3 < BF_{01} < 3$). If we had gathered more data, these replication attempts might have provided us with more evidence, in favor of either hypotheses. Larger sample sizes generally provide a more accurate account of the size and location of the effect that is investigated.

We acknowledged the sample size issue explicitly in our original article: "With respect to sample size, it should be noted that while our sample size was lower than most original studies, our results showed that in our data set, 8 out of a total of 17 hypothesized effects were contradicted with moderate or strong levels of evidence. Thus, even though larger samples

* *Corresponding author*. University of Amsterdam, The Netherlands.
http://dx.doi.org/10.1016/j.cortex.2015.06.018

are always better than smaller samples from a pre-experimental perspective, our Bayesian post-experimental perspective shows that even with 36 participants it is possible to obtain informative results. Nevertheless, we encourage additional replication attempts of SBB correlations using larger sample sizes in order to further decrease uncertainty about the replicability of these effects." (p. 130)

Wagenmakers, Verhagen, Ly, Bakker, et al. (in press) provide concrete illustrations of situations in which low-powered experiments yield diagnostic results, and situations in which high-powered experiments yield nondiagnostic results. In addition, Wagenmakers, Verhagen, and Ly (in press) show that some real, high-powered data sets can produce evidence that is only anecdotal. The Bayesian bottom line is that a power analysis is useful for planning a study, as it takes into account all possible outcomes that can be expected for an intended sample size. However, once a specific data set is observed the power analysis is inferentially irrelevant, as all that counts is the evidence for the data that have been observed.

In our data, for 8 out of 17 effects under scrutiny, our confirmatory Bayesian test yielded non-anecdotal evidence in favor of the null-hypothesis, despite our relatively modest sample size. Thus, after the data have been observed, all that matters is the evidence. Low samples sizes mean that one can expect the evidence to be inconclusive, but that need not always be the case, and our data demonstrate that something can be learned even when sample size is low.

To provide a different perspective on what our data reveal despite the relatively low sample size, Fig. 1 plots effect sizes of the original studies against those of our replication attempt. The blue line represents effect size equality. In general the effects cluster in the area to the right of the line, representing an attenuation of effect sizes in the replication studies. Thus, our results suggest that overall, the effect sizes from our studies are lower than those reported in the original studies.

Muhlert and Ridgway (2015) are concerned that readers might interpret the term "failed replications" to mean that there is compelling evidence for the absence of these effects. When we used the term, we meant to convey the fact that there was no evidence in favor of the presence of the SBB correlations. We agree that the term may be easily misunderstood, and that the ultimate assessment of a replication attempt requires a combination of testing and estimation, coupled with good judgment. The absence of evidence is not the same as evidence of absence, and one of the main advantages of a Bayesian analysis is that it can discriminate between the two possibilities. In our data, for some SBB correlations we find evidence of absence, and for others we find that the evidence is absent. We hope and expect that readers will turn to the concrete results of the Bayes factors and credible intervals to form their own opinion about the extent to which our results constitute a failure to replicate the original findings.

In order to provide a more nuanced perspective on the results from our replication attempts, we reported exploratory replication Bayes factor analyses and plotted posterior distributions for the correlations under scrutiny. These exploratory results can be used to identify potential candidates for further investigation. For example, the exploratory replication Bayes

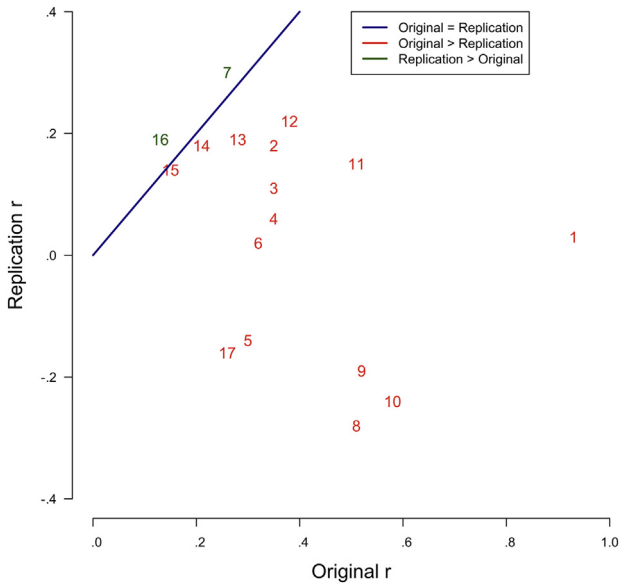

Fig. 1 – Original and replication effect sizes plotted against each other, in order to show the attenuation or amplification of effect sizes. Effects 13 through 17 were flipped in sign for illustration purposes. An effect plotted on the blue line indicates no change in effect size from the original finding to the replication attempt. Effects in green plotted to the left of the blue line indicate amplified effect sizes from the original finding to the replication attempt, whereas effects in red plotted to the right of the blue line indicate attenuation of effect sizes.

factor analysis of the correlation between social network size (SNS) and grey matter volume (GM) shows moderate evidence in favor of an effect similar to the one found in the original study (Table 1; effect #7). This correlation could be further

Table 1 – Summary of the results from 17 replication attempts from Boekel et al. (2015). The data show an overall attenuation of effect size. Both confirmatory (BF_{01}) and exploratory (BF_{0r}) Bayes factors suggest non-anecdotal evidence in favor of the null hypothesis for about half of the replication attempts.

Effect#	r_{orig}	r_{rep}	BF_{01}	BF_{0r}
1	.93	.03	3.90	180.20
2	.35	.18	1.73	1.06
3	.35	.11	2.66	2.06
4	.35	.06	3.51	3.32
5	.30	−.14	7.76	9.56
6	.32	.02	4.35	3.88
7	.26	.30	.57	.27
8	.51	−.28	11.74	249.41
9	.52	−.19	9.40	170.51
10	.58	−.24	10.57	848.06
11	.51	.15	2.04	4.13
12	.38	.22	1.24	.73
13	−.28	−.19	1.51	.67
14	−.21	−.18	1.71	.67
15	−.15	−.14	2.23	.81
16	−.13	−.19	1.60	.65
17	−.26	.16	8.58	7.70

examined in a new data set, using the combined data of previous findings and replications as a prior distribution for an informed Bayesian hypothesis test. In a similar way, the correlation between executive control and cortical thickness in right MTG (Table 1; effect #16) was larger in our replication sample (−.19) than in the original sample (−.13). Nevertheless, our confirmatory Bayes factor analysis suggests that the data are ambiguous (BF_{01} = 1.60) for this SBB correlation. The reason for this is the large difference in sample size between the original study (n = 132) and the replication study (n = 35). While the point estimate of the correlation is larger in our replication sample, the posterior distribution is wider and has more mass near r = 0 (see Fig. 8 and Fig. S17 in Boekel et al., 2015). The replication Bayes factor for this effect tilts more toward the alternative hypothesis (BF_{01} = .65), providing another example for which the addition of data could result in a successful replication. It should be noted that while our exploratory Bayes factor analyses add a layer of nuance to our replication attempt, the general results are similar to our confirmatory analyses. For 9 out of 17 effects, the replication Bayes factor provides non-anecdotal evidence in favor of the null-hypothesis (Table 1).

3. Concern 2: attenuation of effect sizes due to the exploratory nature of discovery

As summarized by Fig. 1, our replication attempts yielded a general attenuation of effect sizes. There are several possible explanations for this attenuation. Muhlert and Ridgway point out that given the exploratory nature of many of the original studies, the attenuation of effect sizes in a replication attempt is not surprising. Specifically, the methods used for detecting an effect in an exploratory study may result in an overestimation of the true effect size (Kriegeskorte, Lindquist, Nichols, Poldrack, & Vul, 2010). This is especially likely in experiments where the sample size is small, such that the effect sizes need to be relatively large in order to pass the classical .05 level of significance. Consequently these effect sizes will likely reduce with subsequent replication attempts. This attenuation, although often a disappointing feature of a confirmatory replication attempt, is a necessary step in identifying the true size of an effect.

4. Concern 3: correction for nuisance variables

Muhlert and Ridgway (2015) suggest that our incomplete correction for nuisance variables is another potential contributor to the attenuation of effect size. Specifically, we corrected the structural brain measures for nuisance variables such as age and sex, but we did not do this for the behavioral measures. This means that if our behavioral measures are correlated with the nuisance variables, our effect sizes may be underestimated. In order to investigate this possibility, we recomputed our results, this time also correcting behavioral data for nuisance variables. Table 2 shows the results. As Muhlert and Ridgway suggested, there seems to be a small overall increase in effect sizes. However, as indicated by the Bayes factors, none of our interpretations were altered in any meaningful way.

Table 2 – Partial versus complete correction for nuisance variables. There is a small general increase in effect sizes after complete correction for nuisance variables. However, Bayes factors were not affected in any meaningful way.

Effect#	Pearson's r		BF_{01}	
	Partial nuisance correction	Complete nuisance correction	Partial nuisance correction	Complete nuisance correction
1	.0324	.0679	3.8962	3.2978
2	.1774	.1833	1.7300	1.6571
3	.1118	.1153	2.6645	2.6093
4	.0627	.0646	3.5142	3.4787
5	−.1365	−.1401	7.7605	7.8485
6	.0167	.0171	4.3511	4.3424
7	.3076	.3209	.5659	.4905
8	−.2798	−.3030	11.7388	12.3844
9	−.1855	−.1927	9.3958	9.5917
10	−.2374	−.2472	10.5733	10.8397
11	.1528	.1588	2.0415	1.9576
12	.2169	.2212	1.2437	1.1977
13	−.1937	−.1976	1.5055	1.4591
14	−.1782	−.1783	1.7094	1.7076
15	−.1400	−.1401	2.2306	2.2290
16	−.1869	−.1870	1.6015	1.5997
17	.1621	.2017	8.5804	9.6141

5. Concern 4: VBM processing pipeline differences

Muhlert and Ridgway (2015) provide an interesting example of the differences in VBM signal intensity between different preprocessing pipelines, using our replication data (for a recent investigation of pipeline differences in SBB research, see Martinez et al., 2015). Their Fig. 2 shows a collection of disconnected regions containing high between-method correlations (r > .8). This figure unfortunately does not show the entire distribution of correlations across the brain. However, it is safe to say that there are indeed differences between analysis methods, which may have impaired our ability to detect a true effect. It should be noted that the same holds true for the original findings: The difference in analysis methods can potentially also result in an overestimation of an effect size. Because of this, replications are an essential tool to investigate the reliability of empirical findings. Furthermore, by specifying pipelines before data are collected, preregistration reduces analytical freedom and consequently the rate of false-positives.

6. Concern 5: ROI approach potentially leads to underestimation of effect sizes

Fig. 1 of Kanai (2015) illustrates the concern that our confirmatory ROI approach might have resulted in an underestimation of effect size. For the convenience of the reader, with permission we have inserted Kanai's figure here (Fig. 2). The mechanism underlying this underestimation is the spatial

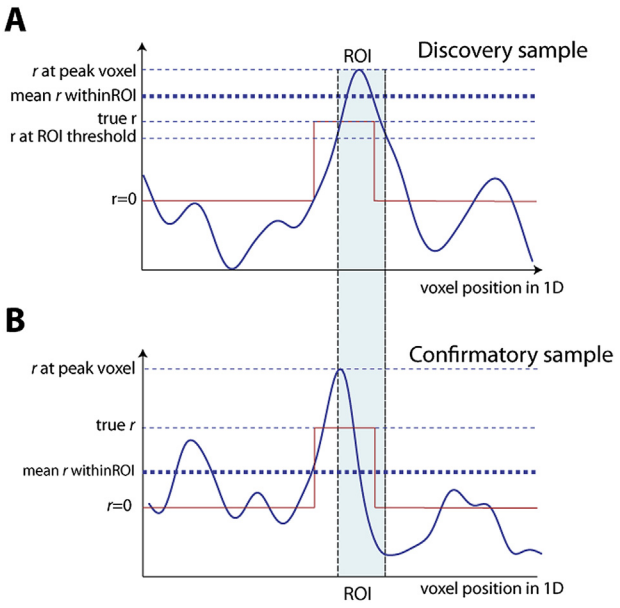

Fig. 2 – Taken from Kanai (2015) with permission, this figure illustrates the issues of over- and underestimation of effect sizes, and uncertainty in effect location. A) An example of a discovery sample. The exploratory nature of the discovery results in an overestimation of the effect size, and uncertainty in terms of the effect location. B) Due to the uncertainty in the effect location and our rigid use of ROIs, the effect size is attenuated.

uncertainty that is introduced in the discovery of the effect (Fig. 2A). When we define our confirmatory ROIs based on the original findings, there is a chance that due to the spatial uncertainty in the discovery, some voxels of the ROI are in fact not part of the true effect location. Because we extracted the average signal from the entire ROI, this could result in an underestimation of the effect size.

Note that there are two types of uncertainty in both the discovery sample as well as the confirmatory sample. In the discovery sample, there is often an overestimation of the effect size due to the exploratory nature of the analysis (Kriegeskorte et al., 2010). In addition, there is uncertainty regarding the exact location of the effect, again due to the exploratory nature of the analysis. In the confirmatory sample, the spatial uncertainty introduced by the discovery sample will often result in an attenuation of the effect size. Our ROI method cannot reduce this spatial uncertainty in the confirmatory sample, as we did not allow the effect to be present in any other area. Future replication attempts should find ways of taking into account the spatial uncertainty introduced by the discovery, while still remaining confirmatory in nature.

These issues certainly have the potential to impair the ability of a replication attempt to detect a true effect. However, they also point out problems of first discoveries: overestimation of effect sizes, and spatial uncertainty in effect locations.

We feel that these problems further emphasize the need for more replication research and meta-analyses. If we were to combine the results from both the discovery sample (Fig. 2A) and the confirmatory sample (Fig. 2B) we would get a more accurate overall perspective on the true effect size and spatial location of the effect. The conclusion drawn from this figure should not be that our replication underestimates effect sizes, but that replication plays an important part in the scientific process of updating knowledge and determining true effect size and location.

7. Concern 6: review process/alternative analyses

Finally, Kanai (2015) wonders whether reviewers of a preregistered manuscript should be allowed to suggest alternative analyses when reviewing the final manuscript including the results. We feel that while reviewers are certainly allowed to suggest alternative analyses, authors should also be allowed to reject them. In a manuscript that is the result of a preregistered study, this rejection should not impact the decision to accept the manuscript for publication. Instead, this decision should rely on the authors' adherence to the preregistered protocol and the sensible interpretation of their findings. In order to facilitate further exploration of the data set, however, authors should make their data publicly available. This way, the results of alternative analysis methods can still be published, albeit not in the original paper. Of course, we strongly recommend that a critical re-analysis of a replication data set is also conducted in a purely confirmatory fashion, including a preregistration document posted, for instance, on the Open Science Framework (https://osf.io/). Because the critical re-analysis is already informed by the data, the statistical results are already contaminated to some extent; preregistration prevents the (explicit or implicit) exploration of alternative methods of analysis until the desired result is found.

An intelligent reader will not make up his mind after reading a single paper. Different experiments show different results, and ideally a scientist will conduct research based on the general notions of a field of research, rather than on the outcome of a single study. Because of this, it is not necessary to include all possible alternative analyses in a single paper. Instead, data should be made publicly available, so that other researchers can conduct their own alternative analyses, and potentially publish their results. These alternative analyses should be preregistered or else labeled as exploratory, after which they can be further investigated, possibly by preregistered replication attempts. In this way, we can begin to elucidate the vastly complex patterns of conditions under which particular effects of interest have certain effect sizes and locations.

8. Impact and future directions

Both the results from our confirmatory replication study and the subsequent commentaries suggest that confirmatory replication studies deserve a more prominent role in the cognitive neurosciences. Future replications should optimize their methods in order to increase the accuracy of their replication attempt. Specific to SBB correlations and other neuroimaging findings, spatial uncertainty should be taken

into account when performing a replication attempt. In order to mitigate the intrusion of QRPs, alternative analyses which take into account spatial uncertainty should also be preregistered. In order to prevent us from fooling ourselves and having our desires and wishes guide our statistical reporting we should consistently and clearly indicate the difference between exploratory and confirmatory analyses in our research, and take caution when interpreting exploratory findings, until preregistered replications have confirmed those initial findings.

One way of taking spatial uncertainty into account is presented in Kanai (2015), point 4. In this section, the correlation between CFQ and GM in left SPL is replicated in our data set using a different, less conservative method. This method relies less heavily on the complete ROI identified by the initial finding, as it conducts a voxel-wise test within a restricted ROI based on the peak voxel coordinates of the original finding. By allowing for more freedom in spatial localization of the effect of interest, this method could be used in subsequent replication attempts to take into account the spatial uncertainty that is often introduced when a discovery is made. Kanai points out that this approach has limitations, such as the arbitrary size of the ROI in which the voxelwise tests are conducted, and the inability of this approach to quantify evidence in favor of the null hypothesis. Another way to take into account spatial uncertainty might be to perform a new, explorative voxel-wise test on the combined data from the original study and the replication attempt to identify the region in which both data sets show a significant correlation. This procedure minimizes the potential for the next replication attempt to underestimate the effect size.

Another promising endeavor is the adversarial collaboration (e.g., Matzke et al., 2015). In this approach, proponents and skeptics of a certain discovery decide to work together to design a confirmatory replication attempt of the discovery and agree on a common plan of analysis. Using Bayesian inference, the evidence may be monitored sequentially as the data accumulate, until the evidence is compelling in favor of either hypotheses. Adversarial collaborations can be multi-site endeavors, potentially resulting in much larger sample sizes than what can reasonably be obtained in single-lab studies.

Adversarial collaborations, Bayesian inference, preregistration, and methods for reducing spatial uncertainty together provide a promising starting point for future replication attempts in the cognitive neurosciences in general, and in structural brain-behavior research in particular.

9. Conclusion

The findings from our replication attempts suggest that results in structural brain-behavior research might not be as reliable as previously thought. The subsequent commentaries of both Kanai (2015) and Muhlert and Ridgway (2015) propose factors that may have contributed to our inability to replicate certain effects. Additional research is needed to investigate these factors in order to provide an accurate account of these (and other) effects in terms of their size, location, and the specific conditions under which they apply. Replication is pivotal in the search for scientific truth. Our confirmatory replication and the subsequent commentaries represent an initial step towards a more reliable and replicable field of research. With this work we hope to stimulate other researchers to undertake similar replication attempts.

REFERENCES

Boekel, W., Wagenmakers, E.-J., Belay, L., Verhagen, J., Brown, S., & Forstmann, B. U. (2015). A purely confirmatory replication study of structural brain-behavior correlations. *Cortex, 66*, 115–133.

Kanai, R. (2015). Open questions in conducting confirmatory replication studies: commentary on "A purely confirmatory replication study of structural brain-behaviour correlations" by Boekel et al., 2015. *Cortex*. http://dx.doi.org/10.1016/j.cortex.2015.02.020.

Kriegeskorte, N., Lindquist, M. A., Nichols, T. A., Poldrack, R. A., & Vul, E. (2010). Everything you never wanted to know about circular analysis, but were afraid to ask. *Journal of Cerebral Blood Flow & Metabolism, 30*(9), 1551–1557.

Martinez, K., Madsen, S. K., Joshi, A. A., Joshi, S. H., Román, F. J., Villalon-Reina, J., et al. (2015). Reproducibility of brain-cognition relationships using three cortical surface-based protocols: an exhaustive analysis based on cortical thickness. *Human Brain Mapping, 36*(8), 3227–3245.

Matzke, D., Nieuwenhuis, S., van Rijn, H., Slagter, H. A., van der Molen, M. W., & Wagenmakers, E.-J. (2015). The effect of horizontal eye movements on free recall: a preregistered adversarial collaboration. *Journal of Experimental Psychology: General, 144*, e1–e15.

Muhlert, N., & Ridgway, G. R. (2015). Failed replications, contributing factors and careful interpretations: commentary on "A purely confirmatory replication study of structural brain-behaviour correlations" by Boekel et al., 2015. *Cortex*. http://dx.doi.org/10.1016/j.cortex.2015.02.019.

Wagenmakers, E.-J., Verhagen, A. J., & Ly, A. (2015). How to quantify the evidence for the absence of a correlation. *Behavior Research Methods*. http://dx.doi.org/10.3758/s13428-015-0593-0.

Wagenmakers, E.-J., Verhagen, A. J., Ly, A., Bakker, M., Lee, M. D., Matzke, D., et al. (2015). A power fallacy. *Behavior Research Methods*. http://dx.doi.org/10.3758/s13428-014-0517-4.

Received 20 May 2015
Reviewed 3 June 2015
Revised 18 June 2015
Accepted 19 June 2015