Response to Comment on Jakob et al.: How to Read a Paper

Marissa-Julia Jakob[1] Moritz C. Dechamps[1] Markus A. Maier
Ludwig-Maximilians-Universität, München

Abstract: We thank Peter Bancel for highlighting the potential strengths and weaknesses of our study and for the intensive examination of the applied methods in order to improve our understanding of the reported effect and support future attempts to scientifically study micro-PK effects. Many arguments put forward in the Comment deserve a deeper examination and we appreciate the discussion raised by Bancel, but we do not agree with the main points of his criticism. In the following, we briefly address the central arguments provided against our analysis strategy and our interpretation of the data and present our view on this matter.[1]

1. Evidence claims not adjusted for multiple testing

Bancel first discusses the issue of multiple testing and its impact on the validity of our claims about Bayesian evidence. One argument was that our claim of strong evidence is flawed because adjustments for multiple testing should have been performed due to the dependence of the tests conducted in the study that used three different micro-PK tasks within one sample.

As Bancel acknowledges, within Bayesian testing approaches adjustments for multiple testing are only mandatory when the tests performed are dependent (Sjölander & Vansteelandt, 2019). This is usually the case when some kind of interrelations between the three micro-PK measurements are present in the data. However, in our view, the fact that the same sample is tested across the three measurements does not automatically imply statistical dependence. We agree that evidence for an effect in all three hypotheses should be based on a common mechanism that is an unconscious belief about the self. However, we argue that the hypotheses can be interpreted

[1] Shared first authorship.

https://doi.org/10.31156/jaex.25782

as independent because they are based on three highly specific psychological beliefs about someone's reality that should only produce micro-PK effects when the stimuli used in the respective task match the content of the relevant belief. In addition, it is possible that the "dependent" personality trait is elicited significantly better than the others because of a better match between this belief and its corresponding stimuli, which would lead to strong evidence for only one of the tests, independent of the evidence obtained for the others. Since we did not know beforehand whether the stimulus material used in the three DVs sufficiently matched the beliefs tested in the study, we consequently formulated our main hypothesis less strictly: We predicted that we would find strong Bayesian evidence for H_1 for at least one of the three micro-PK tasks performed. In addition, we counterbalanced the order of the micro-PK tasks used in the study to control for any order effects between tasks.

Bancel goes on to argue that the data sets for the three PT groups are also dependent, because the groupings are highly correlated. Although this is true, we would like to point out that it is not the correlation of personality traits that is relevant here, but a possible systematic relation between the correlated PT scores and the three micro-PK tasks. From a theoretical point of view, statistical dependence in this context would imply that any (group-related) performance on one micro-PK task should have an impact on performance on any of the other micro-PK tasks included in this study. This assumed interrelation of QRNG outcomes would in itself constitute a micro-PK effect. The micro-PK performance in our study is based on QRNG outputs that involve a true random mechanism.

In any textbook of statistics, the core example of statistical independence is true random events. Regardless of whether different random events are produced by different individuals or by the same individual, they are always considered to be statistically independent. Thus, in theory, one should assume a priori that such measurements are statistically independent. Otherwise, the a priori assumption of statistical dependence between truly random events would not only a priori claim the existence of micro-PK, but would also, in the case of using quantum-based RNGs, question the validity of the Bell theorem and lead to violations of the locality assumption in macroscopic domains. This in turn would make it impossible to document objectively (for which locality would be required) the existence of any effects under study. In other words, the empirical documentation of micro-PK effects to accepted scientific standards would become impossible or even unnecessary, since the effect to be demonstrated is already defined a priori as existing.

Despite this theoretical paradox, which we encounter when the statistical dependence of true random events is postulated a priori, let us assume for a moment that Bancel's argument is valid. Our theoretical background is consistent with his view. We assume that micro-PK exists and that it affects our three different micro-PK results. Furthermore, since the three PT groups are empirically correlated, the three micro-PK measures may (or may not) also be related in some way. If they were related, they would have to be considered statistically dependent and multiplicity controls would have to be performed. Thus, Bancel's argument is empirical rather than theoretical. To address this empirical argument, we will next provide empirical tests of the statistical independence of our three micro-PK measures. If the three measures were statistically independent, then an above-chance score (hit) or a chance and below-chance score (miss) on one of the tests should be indicative of a hit or a miss on the other tests. We tested this assumption with three separate Pearson's chi-squared tests of independence. The results indicated that the likelihood of scoring a hit or a miss was not significantly associated with any outcome (hit or miss) on the other tasks (dependent vs. avoidant χ^2 (1, N = 2,403) = 0.43, p = .51; dependent vs. obsessive-compulsive χ^2 (1, N = 2,403) = 1.08, p = .30; avoidant vs. obsessive-compulsive χ^2 (1, N = 2,403) = 0.13, p = .72). These analyses indicate that hits or misses are completely randomly distributed across the three tasks and they empirically document that statistical independence can be assumed between the three micro-PK tasks. The power to detect even small violations (w = .1) of statistical independence was 99% in each χ^2 test.

Since statistical independence exists across the three micro-PK measurements multiplicity control for our Bayesian tests did not need to be performed (Sjölander & Vansteelandt, 2019) and the strong Bayesian evidence (BF_{10} > 10) reported for the micro-PK effect found within the dependent PT-group can still considered to be valid.

Finally, if one still prefers to maintain the statistical dependence assumption, one needs to perform multiplicity controls, and these would affect expectations about the prior model probabilities. One possibility is to adjust the model probabilities in such a way that not each (of the three) null hypotheses has a probability of ½ (leading to the a priori statement that the probability of finding no effect in all three tests is $½^3$= .125), but that the total prior probability of finding no differences in all three tests combined is ½ (null control method; Williams et al., 2016). Following de Jong (2019), this could be achieved by changing the prior model probability for the H_0 from 0.5 to $0.5^{⅔}$ = 0.63 (see Westfall et al., 1997). It is important to note that this does not change the individual Bayes factors for each test, but changes the posterior probability of an effect when considering the entire study by a factor of 0.37/0.63=0.59 as seen in (1).

$$\frac{P(H_1|D}{P(H_0|D} = \frac{P(H_1)}{P(H_0)} \times \frac{P(D|H_1)}{P(D|H_0)} (1)$$

$$Posteriorodds = Priorodds \times BF$$

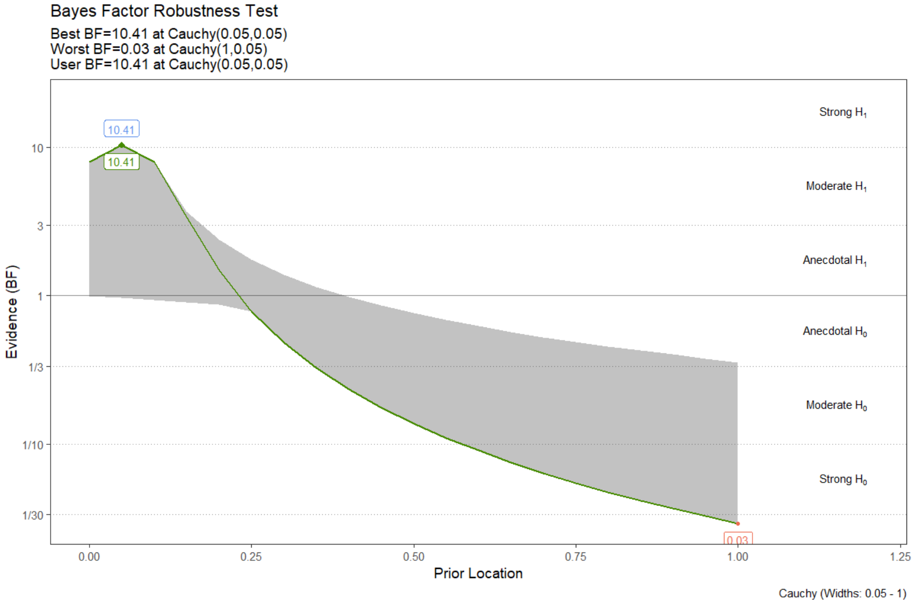
2. Lack of Context for Bayesian *t* Tests

Bancel emphasizes the importance of conducting sensitivity checks in Bayesian analysis, especially when using informed priors and dealing with small effects. He argues that an a priori justification for the prior based on the results of previous studies does not justify overlooking the sensitivity of the results to parameter choices, and links this issue to experimenter psi in psi research. We agree that the choice of the prior involves a degree of freedom for the experimenter. In Bayesian analysis, this is considered a strength, since it is advantageous to specify a test for the expected effect directly. It is therefore good practice to use these features of Bayesian statistics when they can be applied with some confidence. This is the case here, as we had a fairly certain expectation of the expected effect size for this type of micro-PK experiment based on previous studies and the literature. This deliberately translated expectation was pre-registered a priori, eliminating the degree of freedom in the analysis, and was ultimately confirmed by parts of the results of the study that showed an effect size of *d* = .07, which fits well with our a priori estimate.

Not surprisingly, a sensitivity analysis shows that this effect is not very robust to parameter changes (see robustness analysis in fig. 1 below). This is to be expected, since a very small effect can only be detected with a reasonable sample size if the test used is sensitive to that effect size. This is partly why we prefer Bayesian methods in this line of research, and is not a problem as long as the parameters are registered in advance.

Figure 1

Robustness analysis of the effect in the dependent sample. It can be seen that the BF has its highest value for the a priori chosen informed Cauchy parameters. The evidence transitions into pointing towards H0 for a very narrow prior (green line) at around Cauchy(0.25, 0.05). The gray area shows the BFs for different prior widths.



3. An Undetermined Type I Error Rate of the Primary Hypothesis

Bancel criticizes a lack of discussion of the study's false positive error rate (FPE) and its relation to a frequent testing of sequential evidence. He argues that the three tests are interdependent and therefore the FPE in this case is approximately 15%. We appreciate the effort put into the false positive error rate (FPE) analysis. Bancel cites Schönbrodt and Wagenmakers (2018), who provide a simple way to calculate the FPE in Bayesian designs and provide comparable but slightly lower results. This analysis shows an FPE of 3.8% for a sequential design with n-start = 100, n-max = 1,400, and a step size of 10 participants, which are conservative estimates. This shows that the alpha error for each test is well below the generally accepted threshold of .05.

Following the argument above that the data (the micro-PK trials corresponding to the specific personality trait) of the three tests are independent, it is not necessary to control for the experimental error rate (EER), which is typically done with corrections when testing multiple frequentist comparisons simultaneously. The power of each test is not affected by the outcome of the other tests, and the stopping criteria are based on each test individually reaching a specific BF threshold, not on a combination of tests. Therefore, the Type I error rate remains as reported for each hypothesis evaluated individually.

It can also be seen that the FPE changes only slightly with different step sizes (i.e., test intervals), and that the sequential Bayesian designs are usually quite robust to over-testing errors when a minimum N is used for the first test (e.g., changing the step-size to 50 reduces the FPE to 2.8%).

4. Failure to Follow the Pre-Registered Procedure

Bancel criticizes a deviation from the pre-registered stopping rule during data collection, leading to an undefined degree of experimenter freedom in the protocol. Specifically, the data collection was not stopped when the BF > 10 criterion was met for the first time.

We agree that some deviations from the pre-registered protocol were made. All deviations are explicitly mentioned in the original article. However, they do not compromise the evidence of the studies, since Bayesian evidence only becomes more precise as more data are added. According to the protocol, data collection should have stopped when the evidence criterion of BF > 10 was reached for the first time at *n* = 820. In the analysis process, we missed this exact point and observed a decrease in evidence shortly thereafter. Therefore, we decided to collect more data until a conclusive result was reached, as suggested by Schönbrodt et al. (2017). In Bayesian statistics, it is always possible to continue collecting data while updating the data analysis, as Bayes factors are consistent because they converge either to zero (if H_0 is true) or to infinity (if an effect is present) for one-tailed designs (Bayarri & Berger, 2004; Morey & Rouder, 2011). Schönbrodt et al. (2017) showed that sequential Bayesian designs have a lower long-term rate of misleading evidence than frequentist procedures with 5% Type I and 20% Type II error rates, while most errors occur at small sample sizes. Unlike *p*-values, the interpretation of Bayes factors does not depend on stopping rules (Rouder, 2014). Therefore, adding data to a Bayesian analysis is never an issue, and stopping criteria are rather a means to design studies efficiently. This was the case here, and the stopping rule of *N* = 1,000 participants in the absence of evidence was implemented for economic reasons. With the narrowly informed prior we chose, obtaining strong evidence for H_0 would require a very large sample size. At the time we uploaded the pre-registration, our resources were limited to 1,000 participants. As the process progressed, more resources became available to continue data collection beyond *N* = 1,000.

Furthermore, we never intended to perform a frequentist overall test of the three micro-PK tasks. Instead, we view them as three independent experiments, each with its own independent and dependent variables. This was emphasized in the pre-registration form, which allowed us to stop data collection as soon as any of the three experiments reached our statistical evidence criterion. It is up to the research community to continue data collection if they are interested.

5. Inadequate Treatment of the Control Group Analysis

Finally, Bancel notes the lack of a direct comparison between the target and control groups and criticizes the use of one-tailed tests for the former and two-tailed tests for the latter, which makes the tests inappropriate for comparison. As explained in the pre-registration form, we deliberately chose a one-sample *t*-test design against the expected value under chance rather than a direct group comparison for theoretical reasons. Our main point is not that the two groups are different, but that there is a micro-PK effect due to intentional observation within the PT-high group.

The group split can be seen as a sample pre-selection of individuals with pronounced traits (PT-high groups) that favor the micro-PK effect. A direct comparison of the groups is not meaningful, because the chosen splitting criterion is based on a continuous measure (VDS-30 questionnaire). Therefore, it is possible that the control group also shows (weaker) micro-PK effects in the same or in the opposite direction as the experimental group, since they can also be considered as motivated observers. Consequently, a two-tailed *t*-test was performed for the control group, since a non-random, less-than-chance result does not fit H_0 either. In contrast, the hypotheses for the experimental group were formulated as one-tailed, following the predictions of the Emotional Transgression Model as our theoretical background. Note that the group differences are not explained by the one-tailed vs. two-tailed setting variation, which can be checked with the data set and analysis scripts provided at OSF.

In summary, some of the criticisms in the comment are valid, but we disagree with the conclusion that they lead to nonconfirmatory results and reduce the significance of the strong evidence for the micro-PK effect found in one of the three experiments. Note that the analyses of all PT-high data suggested by the author of the comment yielded p = .06, which just barely exceeds the convention of .05. As noted in the Discussion section of our paper, further analyses of the combined score of the three tasks will be included in another paper we are currently working on. In addition, we disagree that the citation of “Jahn et al. (2000)” was misused in our paper. We mentioned this study as an example of an initial micro-PK effect that could not be replicated. However, our claim was not that the replications failed due to a decline effect. Moreover, we believe that such minor complaints should be part of a review process rather than raised in a comment.

References

Bayarri, M. J., & Berger, J. O. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science, 19(1),* 58-80. https://doi.org/10.1214/088342304000000116

de Jong, T. (2019). A Bayesian approach to the correction for multiplicity. https://doi.org/10.31234/osf.io/s56mk

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods, 16*, 406–419. https://doi.org/10.1037/a0024377

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review, 21*, 301- 308. https://doi.org/10.3758/s13423-014-0595-4

Schönbrodt, F.D., Wagenmakers, E.J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychon Bull Rev 25*, 128–142. https://doi.org/10.3758/s13423-017-1230-y

Schönbrodt, F. D., Wagenmakers, E. J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods, 22(2),* 322. https://doi.org/10.1037/met0000061

Sjölander, A., & Vansteelandt, S. (2019). Frequentist versus Bayesian approaches to multiple testing. *European Journal of Epidemiology, 34(9),* 809–821. https://doi.org/10.1007/s10654-019-00517-2

Westfall, P. H., Johnson, W. O., & Utts, J. M. (1997). A Bayesian perspective on the Bonferroni adjustment. *Biometrika, 84(2),* 419–427. https://doi.org/10.1093/biomet/84.2.419

Williams, P., Heathcote, A., Nesbitt, K., & Eidels, A. (2016). Post-error recklessness and the hot hand. *Judgment and Decision Making, 11(2),* 174–184. https://doi.org/10.1017/S1930297500007282

Response au Commentaire sur Jakob et al.: How to Read a Paper (Comment lire un article)

Marissa-Julia Jakob Moritz C. Dechamps Markus A. Maier

Résumé: Nous remercions Peter Bancel d'avoir souligné les forces et les faiblesses potentielles de notre étude et d'avoir examiné en profondeur les méthodes appliquées afin d'améliorer notre compréhension de l'effet rapporté et de soutenir les futures tentatives d'étude scientifique des effets des micro-PK. De nombreux arguments avancés dans le commentaire méritent un examen plus approfondi et nous apprécions la discussion soulevée par Bancel, mais nous ne sommes pas d'accord avec les principaux points de sa critique. Dans ce qui suit, nous abordons brièvement les principaux arguments avancés à l'encontre de notre stratégie d'analyse et de notre interprétation des données, et nous présentons notre point de vue sur la question.

Translation into French by Antoine Bioy, Ph. D.

Antwort auf den Kommentar zur Jakob et al.: Wie man eine Arbeit liest

Marissa-Julia Jakob Moritz C. Dechamps Markus A. Maier

Wir danken Peter Bancel für das Aufzeigen möglicher Stärken und Schwächen unserer Studie und für die intensive Überprüfung mit den angewandten Methoden, um unser Verständnis des berichteten Effekts zu verbessern und zukünftige Versuche, Mikro-PK-Effekte wissenschaftlich zu untersuchen, zu unterstützen. Viele der in dem Kommentar vorgebrachten Argumente verdienen eine eingehendere Prüfung, und wir schätzen die von Bancel geführte Diskussion, stimmen aber nicht mit den Hauptpunkten seiner Kritik überein. Im Folgenden gehen wir kurz auf die zentralen Argumente ein, die gegen unsere Analysestrategie und unsere Interpretation der Daten vorgebracht wurden, und legen unsere Meinung dazu dar.

Translation into German by Eberhard Bauer, Ph. D.

Resposta ao Comentário sobre Jakob et al.: Como Ler um Artigo

Marissa-Julia Jakob Moritz C. Dechamps Markus A. Maier

Resumo: Agradecemos a Peter Bancel por destacar os potenciais aspectos positivos e fragilidades de nosso estudo e pela rigorosa análise dos métodos aplicados, visando melhorar nossa compreensão acerca do efeito relatado e apoiar futuras tentativas de estudar cientificamente os efeitos micro-PK. Muitos argumentos apresentados no Comentário merecem uma análise mais profunda, e agradecemos a discussão levantada por Bancel, porém, não concordamos com os principais pontos de sua crítica. A seguir, abordamos brevemente os argumentos centrais sugeridos contra nossa estratégia de análise e nossa interpretação dos dados e apresentamos nossa visão sobre o tema.

Translation into Portuguese by Antônio Lima, Ph. D.

Respuesta al Comentario sobre Jakob et al.: Cómo Leer un Artículo

Marissa-Julia Jakob Moritz C. Dechamps Markus A. Maier

Resumen: Damos las gracias a Peter Bancel por señalar los posibles puntos fuertes y débiles de nuestro estudio y por el examen intensivo de los métodos aplicados con el fin de mejorar nuestra comprensión del efecto mencionado y apoyar futuros intentos de estudiar científicamente los efectos micro-PK. Muchos de los argumentos expuestos en el comentario merecen un examen más profundo y agradecemos el debate planteado por Bancel, pero no estamos de acuerdo con los puntos principales de su crítica. Abordamos brevemente los argumentos centrales aportados contra nuestra estrategia de análisis y nuestra interpretación de los datos y presentamos nuestra opinión al respecto.

Translation into Spanish by Etzel Cardeña, Ph. D.